THE IMPORTANCE OF DIVERSITY IN PRECISION MEDICINE RESEARCH

EDITED BY: Dana C. Crawford, Jessica Nicole Cooke Bailey and William Scott Bush

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88966-099-5
DOI 10.3389/978-2-88966-099-5

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

THE IMPORTANCE OF DIVERSITY IN PRECISION MEDICINE RESEARCH

Topic Editors:
Dana C. Crawford, Case Western Reserve University, United States
Jessica Nicole Cooke Bailey, Case Western Reserve University, United States
William Scott Bush, Case Western Reserve University, United States

Citation: Crawford, D. C., Bailey, J. N. C., Bush, W. S., eds. (2020). The Importance of Diversity in Precision Medicine Research. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88966-099-5

Table of Contents

frontiers in Genetics

EDITORIAL
published: 26 August 2020
doi: 10.3389/fgene.2020.00875

Editorial: The Importance of Diversity in Precision Medicine Research

Jessica N. Cooke Bailey[1]*, *William S. Bush*[1,2] and *Dana C. Crawford*[1,2]*

[1] Department of Population and Quantitative Health Sciences, Cleveland Institute for Computational Biology, Case Western Reserve University, Cleveland, OH, United States, [2] Department of Genetics and Genome Sciences, Case Western Reserve University, Cleveland, OH, United States

Keywords: precision medicine, diversity, genomics, personalized medicine, research participant, genetic ancestry, polygenic risk scores, social determinants of health

Editorial on the Research Topic

The Importance of Diversity in Precision Medicine Research

Personalized or precision medicine is meant to distinguish tailored treatment from trial and error. The contemporary concept has evolved to specifically include the 'omic profile of a patient in the prevention, diagnosis, and treatment of disease. Rapid genomic discoveries made possible through genome-wide association studies (GWAS) coupled with decreasing costs of sequencing and genotyping have shifted precision medicine from an academic exercise to clinical reality for some conditions (e.g., Wigle et al., 2017; Claassens et al., 2019; Hamdan et al., 2019; Lim, 2019; Roden, 2019), while others are not far behind. The emergence of electronic health records (EHRs) now makes it possible to both perform population-scale research and effectively deliver personalized medicine to the individual patient through clinical decision support.

While the promise of precision medicine is great, several identifiable gaps exist in current research that limit its reach to all potential patients. One key deficiency is the lack of diversity among biomedical research participants, which limits both the generalizability and availability of genomic-based treatments or prevention strategies. The vast under-representation of diverse populations in genetic/genomic studies (e.g., Sirugo et al., 2019) is highly problematic as genetic information gleaned from one population is not automatically transferrable across populations (Popejoy and Fullerton, 2016). Without sample diversity, signals revealing powerful insights into genetic association and/or drug response can go undetected due to differences in linkage disequilibrium, allele frequencies, and genetic architecture. New initiatives and studies are now in place to ensure the inclusion of traditionally underrepresented groups, defined by race/ethnicity, socioeconomic status and/or position, geography, and age, in genomic research (Bentley et al., 2020). We therefore anticipate a swell of new data and methodologies accelerating the already rapid pace of precision medicine research.

Our goal for this Research Topic was to present original research, commentaries, perspectives, and reviews on the impact and importance of diversity in precision medicine research. Below, we briefly overview the nine accepted manuscripts and the context in which they address this goal.

IMPORTANCE OF RECRUITMENT AND RETENTION OF DIVERSE PARTICIPANTS

A 2009 analysis of GWAS participants revealed only 4% of DNA samples were from non-European participants (Need and Goldstein, 2009). By 2016, 20% of DNA samples were from non-European samples; however, this more than 2000%-fold increase was mainly due to expansion of studies in primarily East Asian ancestry populations (Popejoy and Fullerton, 2016). Taken together, less than 4% of samples analyzed were from individuals of African and Latin American ancestry,

OPEN ACCESS

Edited by:
Daniel Shriner, National Human Genome Research Institute (NHGRI), United States

Reviewed by:
Nora Franceschini, University of North Carolina at Chapel Hill, United States

***Correspondence:**
Jessica N. Cooke Bailey
jnc43@case.edu
Dana C. Crawford
dana.crawford@case.edu

Specialty section:
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics

Received: *22 June 2020*
Accepted: *17 July 2020*
Published: *26 August 2020*

Citation:
Cooke Bailey JN, Bush WS and Crawford DC (2020) Editorial: The Importance of Diversity in Precision Medicine Research. Front. Genet. 11:875. doi: 10.3389/fgene.2020.00875

Hispanic people, and native or indigenous peoples, despite that these are the most vulnerable and traditionally underserved populations worldwide (Popejoy and Fullerton, 2016). Inclusion of diverse groups is key to diversifying the pool from which precision medicine can be developed. Discerning factors that influence participation and incorporating these findings into inclusive ascertainment strategies are crucial; efforts must be made to understand ways in which diverse groups can be accessed and invited to participate, as well as to identify motivators and/or barriers affecting willingness to participate and to remain in studies (Perreira et al., 2020). Four publications in this Research Topic addressed the topic "development of culturally-appropriate consent and recruitment strategies for precision medicine research" and "barriers to participation in research (such as access to technology, genomic literacy, concerns for digital data privacy, and factors that impact time or means to participate in research)."

To identify addressable issues and adjust enrollment protocols to improve participation among Hispanics, Nuytemans et al. sought to identify motivators of patients and caregivers affected by Parkinson's disease (PD) to participate in genomic research via surveys administered to patients in the University of Miami Health System's Movement Disorders Clinic, wherein approximately 35% of patients identify as Hispanic. Of the more than 150 self-identified white PD patients and caregivers, approximately 60% of whom were Hispanic, Hispanics and non-Hispanics were equally motivated to participate in genetic research for PD, but Hispanic patients were less likely to be influenced by the promise of scientific advancements. This lack of scientific interest was found to be likely confounded by lower levels of obtained education. The authors suggest that a potential reason for the underrepresentation of Hispanics in genetic research is due to reduced invitations to studies.

Also focused on motivators for research participation within a patient population impacted by a chronic disease, Cuccaro et al. surveyed individuals with multiple sclerosis (MS) participating in a genetic study of MS. The majority of approached study participants (95/101) were willing to participate in the survey; of these, over 80% were Hispanic and female. Survey respondents were asked to identify the primary reasons or motivations for participation. The most frequently cited reason was finding a cure, equally endorsed by Hispanic and non-Hispanic participants; having MS and helping future generations were also highly endorsed motivators, with Hispanics more frequently citing having MS and non-Hispanics more frequently citing finding new/better treatments. Overall, ethnicity was the only significant factor associated with willingness to participate.

The dearth of genetic data available for populations other than those of European descent extends to pharmacogenomics (PGx), the study of genomic information relevant to drug response to tailor dosing. Scherr et al. review challenges to recruiting African American participants in genomic studies and extrapolate these findings to PGx. Consistent with prior reports, their review highlighted African American distrust of the healthcare system, medical research, organization, and researchers as barriers to study participation. Authentic, intentional collaborations between researchers and communities are suggested as a means by which to begin overcoming distrust. Another overarching barrier was lack of knowledge or awareness regarding genomic studies. To reduce distrust and increase awareness, they suggest transparent and clearly described study protocols, educational messaging, and recruitment efforts that directly address existing attitudes and beliefs of distrust. Importantly, there was no evidence of lack of interest in research study participation; conversely, they found that African Americans are aware that participation in medical research is crucial to medical and scientific advancement. Thus, Scherr et al. suggest a focused approach to recruiting African American research study participants, including messaging that highlights altruism.

Alzheimer disease (AD) is another common, complex disease with later-in-life onset and for which most genetic and genomic studies to date have focused on individuals of European descent (e.g., Beecham et al., 2017). Feliciano-Astacio et al. describe the ascertainment approach applied in the Puerto Rico Alzheimer Disease Initiative (PRADI), a multisource recruitment effort to increase participation by Puerto Ricans in genomic research of AD, which currently has >670 participants. PRADI's successful recruitment was attained by establishing strong community engagement relationships and tailored recruitment of AD patients and families across multiple sites in Puerto Rico. Focused and deliberate recruitment efforts such as these will help ensure the inclusion of Hispanic and Latino populations in future precision medicine research efforts.

POPULATION DIVERSITY AND GENE EXPRESSION

One publication in this Research Topic addressed "statistical methods for genomic data from multiple populations." A popular statistical method, known as PrediXcan (Gamazon et al., 2015), infers gene expression using genetic data. While now widely used on a variety of datasets derived from many different populations, most gene expression datasets are from majority European-descent populations, and thus construction of reference panels used by PrediXcan are based on European-descent data. Mikhaylova and Thornton evaluate the accuracy of PrediXcan in predicting or inferring gene expression in diverse populations. Using a combination of Genetic European Variation in Disease (Geuvadis) RNA sequencing data and 1000 Genomes Project whole genome sequencing data, Mikhaylova and Thornton demonstrate that the performance of PrediXcan varies by population, with lower performance for African-descent populations compared with others available in the 1000 Genomes Project. The data suggest that prediction models developed using European reference panels are not necessarily transferrable to other populations due to differences in allele frequency, linkage disequilibrium, and genetic admixture.

SOCIAL DETERMINANTS OF HEALTH AND GENETIC ASSOCIATION STUDIES

For complex diseases and traits, genetic variation alone does not sufficiently explain the totality of risk or variation. While this

observation is widely accepted, few genetic association studies incorporate important measures of lifestyle, environmental exposures, or social determinants of health associated with disease risk and health disparities. Hollister et al. address this challenge by applying their recently validated algorithm that defines socioeconomic status using electronic health records (Hollister et al., 2017) to a large clinical population of African American patients. All patients were clinically screened for hypertension, a complex condition disproportionately prevalent in African Americans (Fryar et al., 2017) that is independently associated with many common genetic variants and environmental exposures such as diet and socioeconomic status (Aburto et al., 2013; Giri et al., 2019; de las Fuentes et al., 2020; Glover et al., 2020; Hollister et al.). In the work presented herein, Hollister et al. tested for and possibly identified a statistical interaction between education, a recognized social determinant of health, and genetic variants contributing to blood pressure, underscoring the need for additional study of the potentially modifying effects of non-genetic factors for diseases with noted population differences.

CANDIDATE GENE VARIATION AND CHRONIC OBSTRUCTIVE PULMONARY DISEASE

Two publications in this Research Topic addressed "Genomic discovery in non-European populations." Khanna et al. present a meta-analysis of 14 published studies investigating the association between variants rs4588 and rs7041 in the Vitamin-D binding (*GC*) protein locus and chronic obstructive pulmonary disease (COPD). Both *GC* rs4588 and rs7041 are robustly associated with vitamin D levels in GWAS of mostly European-descent populations (Manousaki et al., 2017; O'Brien et al., 2018). The meta-analysis presented by Khanna et al. include both European- and Asian-descent populations. Both single SNP tests of association for COPD and evaluations of linkage disequilibrium and haplotypes using publicly available genomic and *in silico* data are presented for multiple populations to more fully describe the genetic epidemiology of these loci.

Nandy et al. evaluated the association between serum surfactant protein D (SFTPD) concentration and *SFTPD* rs721917 and chronic obstructive pulmonary disease (COPD) and acute exacerbation COPD (AECOPD). Recent large GWAS of mostly European-descent populations have identified *SFTPD* rs721917 as significantly associated with COPD at genome-wide significance (Hobbs et al., 2017; Sakornsakolpat et al., 2019). Nandy and colleagues identified and meta-analyzed results from eight independent published reports, which included six with serum SFTPD concentrations and three with *SFTPD* rs721917 genotype data for Asian populations from China, Lebanon, and Pakistan. As expected, both COPD and AECOPD were associated with serum SFTPD. However, while *SFTPD* rs721917 was significantly associated with both COPD and AECOPD in this meta-analysis, the direction of effect was opposite of that previously reported by recent GWAS of COPD (Hobbs et al., 2017; Sakornsakolpat et al., 2019). While limited in sample size, this small meta-analysis underscores the importance of generalizing GWAS findings in diverse populations.

POLYGENIC RISK SCORES AND DIVERSE POPULATIONS

One publication in this Research Topic addressed "The use of genetic ancestry for genomic discovery (such as admixture mapping)." Genetic and polygenic risk score studies aggregate cumulative effects across genetic loci; effect sizes are typically estimated from GWAS that have traditionally been performed in samples of European descent. Unfortunately, polygenic risk scores do not always replicate in non-European ancestral groups [reviewed in (Sirugo et al., 2019)]. Focusing on Chinese and Japanese samples, Zhou et al. evaluated lumbar disc degeneration (LDD), another complex, age-related phenotype. The focus of this work was to investigate genetic overlap between LDD and four related risk factors. Strong association between a polygenic LDD score, constructed with weights from European-ancestry studies, and related risk factors was detected. However, phenotype variances explained were lower than in prior European studies, thus, reducing power to detect genetic overlaps. This study again emphasizes the importance of genetic studies inclusive of populations other than Europeans.

Taken together, this Research Topic is composed of nine publications that further emphasize the importance of diversity in precision medicine research and offer solutions to better ensure these translational research efforts are realized in the clinic for all to benefit.

AUTHOR CONTRIBUTIONS

JC, WB, and DC conceived the idea for and wrote this editorial. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The authors acknowledged the Cleveland Institute for Computational Biology and NIH grants R13HG009481 and R13HG010286 for supporting scholarly discussion and conferences associated with this research topic.

REFERENCES

Aburto, N. J., Ziolkovska, A., Hooper, L., Elliott, P., Cappuccio, F. P., and Meerpohl, J. J. (2013). Effect of lower sodium intake on health: systematic review and meta-analyses. *BMJ* 346:f1326. doi: 10.1136/bmj.f1326

Beecham, G. W., Bis, J. C., Martin, E. R., Choi, S. H., DeStefano, A. L., Van Duijn, C. M., et al. (2017). Clinical/scientific notes: the Alzheimer's disease sequencing project: study design and sample selection. *Neurol. Genet.* 3:e194. doi: 10.1212/NXG.0000000000000194

Bentley, A. R., Callier, S. L., and Rotimi, C. N. (2020). Evaluating the promise of inclusion of African ancestry populations in genomics. *npj Genomic Med.* 5, 1–9. doi: 10.1038/s41525-019-0111-x

Claassens, D. M. F., Vos, G. J. A., Bergmeijer, T. O., Hermanides, R. S., Van't Hof, A. W. J., Van Der Harst, P., et al. (2019). A genotype-guided strategy for oral P2Y12 inhibitors in primary PCI. *N. Engl. J. Med.* 381, 1621–1631. doi: 10.1056/NEJMoa1907096

de las Fuentes, L., Sung, Y. J., Noordam, R., Winkler, T., Feitosa, M. F., Schwander, K., et al. (2020). Gene-educational attainment interactions in a multi-ancestry genome-wide meta-analysis identify novel blood pressure loci. *Mol. Psychiatry*. doi: 10.1038/s41380-020-0719-3

Fryar, C. D., Ostchega, Y., Hales, C. M., Zhang, G., and Kruszon-Moran, D. (2017). Hypertension prevalence and control among adults: United States, 2015–2016. *NCHS Data Brief*. 1–8.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* 47, 1091–1098. doi: 10.1038/ng.3367

Giri, A., Hellwege, J. N., Keaton, J. M., Park, J., Qiu, C., Warren, H. R., et al. (2019). Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat. Genet.* 51, 51–62. doi: 10.1038/s41588-018-0303-9

Glover, L. M., Cain-Shields, L. R., Wyatt, S. B., Gebreab, S. Y., Diez-Roux, A. V., and Sims, M. (2020). Life course socioeconomic status and hypertension in African American adults: the jackson heart study. *Am. J. Hypertens.* 33, 84–91. doi: 10.1093/ajh/hpz133

Hamdan, D., Nguyen, T. T., Leboeuf, C., Meles, S., Janin, A., and Bousquet, G. (2019). Genomics applied to the treatment of breast cancer. *Oncotarget* 10, 4786–4801. doi: 10.18632/oncotarget.27102

Hobbs, B. D., De Jong, K., Lamontagne, M., Bossé, Y., Shrine, N., Artigas, M. S., et al. (2017). Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nat. Genet.* 49, 426–432. doi: 10.1038/ng.3752

Hollister, B. M., Restrepo, N. A., Farber-Eger, E., Crawford, D. C., Aldrich, M. C., and Non, A. (2017). Development and performance of text-mining algorithms to extract socioeconomic status from de-identified electronic health records. *Pac Symp. Biocomput.* 22, 230–41. doi: 10.1142/9789813207813_0023

Lim, G. B. (2019). Reduced bleeding with genotype-guided antiplatelet therapy. *Nat. Rev. Cardiol.* 16, 646–647. doi: 10.1038/s41569-019-0276-0

Manousaki, D., Dudding, T., Haworth, S., Hsu, Y. H., Liu, C. T., Medina-Gómez, C., et al. (2017). Low-frequency synonymous coding variation in CYP2R1 has large effects on vitamin D levels and risk of multiple sclerosis. *Am. J. Hum. Genet.* 101, 227–238. doi: 10.1016/j.ajhg.2017.06.014

Need, A. C., and Goldstein, D. B. (2009). Next generation disparities in human genomics: concerns and remedies. *Trends Genet.* 25, 489–494. doi: 10.1016/j.tig.2009.09.012

O'Brien, K. M., Sandler, D. P., Shi, M., Harmon, Q. E., Taylor, J. A., and Weinberg, C. R. (2018). Genome-wide association study of serum 25-hydroxyvitamin D in US women. *Front. Genet.* 9:67. doi: 10.3389/fgene.2018.00067

Perreira, K. M., Los Angeles Abreu, M., Zhao, B., Youngblood, M. E., Alvarado, C., Cobo, N., et al. (2020). Retaining hispanics: lessons from the hispanic community health study/study of latinos (HCHS/SOL). *Am. J. Epidemiol.* 189, 518–531. doi: 10.1093/aje/kwaa003

Popejoy, A. B., and Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature* 538, 161–164. doi: 10.1038/538161a

Roden, D. M. (2019). Clopidogrel pharmacogenetics - Why the wait? *N. Engl. J. Med.* 381, 1677–1678. doi: 10.1056/NEJMe1911496

Sakornsakolpat, P., Prokopenko, D., Lamontagne, M., Reeve, N. F., Guyatt, A. L., Jackson, V. E., et al. (2019). Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nat. Genet.* 51, 494–505. doi: 10.1038/s41588-018-0342-2

Sirugo, G., Williams, S. M., and Tishkoff, S. A. (2019). The missing diversity in human genetic studies. *Cell* 177, 26–31. doi: 10.1016/j.cell.2019.02.048

Wigle, T., Jansen, L., Teft, W., and Kim, R. (2017). Pharmacogenomics guided-personalization of warfarin and tamoxifen. *J. Pers. Med.* 7:20. doi: 10.3390/jpm7040020

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 August 2018
doi: 10.3389/fgene.2018.00267

Trans-Ethnic Polygenic Analysis Supports Genetic Overlaps of Lumbar Disc Degeneration With Height, Body Mass Index, and Bone Mineral Density

*Xueya Zhou[1,2], Ching-Lung Cheung[3,4], Tatsuki Karasugi[5], Jaro Karppinen[6], Dino Samartzis[7], Yi-Hsiang Hsu[8,9,10], Timothy Shin-Heng Mak[4], You-Qiang Song[4,11], Kazuhiro Chiba[12], Yoshiharu Kawaguchi[13], Yan Li[1], Danny Chan[11], Kenneth Man-Chee Cheung[7], Shiro Ikegawa[14], Kathryn Song-Eng Cheah[11] and Pak Chung Sham[1,4]**

[1] Department of Psychiatry, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, Hong Kong, [2] Department of Systems Biology, Department of Pediatrics, Columbia University Medical Center, New York, NY, United States, [3] Department of Pharmacology and Pharmacy, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, Hong Kong, [4] Li Ka Shing Faculty of Medicine, Center for Genomic Sciences, The University of Hong Kong, Hong Kong, Hong Kong, [5] Department of Orthopaedic Surgery, Faculty of Life Sciences, Kumamoto University, Kumamoto City, Japan, [6] Medical Research Center Oulu, University of Oulu and Oulu University Hospital, Oulu, Finland, [7] Department of Orthopaedics and Traumatology, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, Hong Kong, [8] Hebrew SeniorLife, Institute for Aging Research, Roslindale, MA, United States, [9] Harvard Medical School, Boston, MA, United States, [10] Molecular and Integrative Physiological Sciences Program, Harvard School of Public Health, Boston, MA, United States, [11] Li Ka Shing Faculty of Medicine, School of Biomedical Science, The University of Hong Kong, Hong Kong, Hong Kong, [12] Department of Orthopedic Surgery, National Defense Medical College, Tokorozawa, Saitama, Japan, [13] Department of Orthopaedic Surgery, Toyama University, Toyama Prefecture, Japan, [14] Laboratory of Bone and Joint Diseases, Center for Integrative Medical Sciences, RIKEN, Tokyo, Japan

OPEN ACCESS

Edited by:
Dana C. Crawford, Case Western Reserve University, United States

Reviewed by:
Kenneth M. Weiss, Pennsylvania State University, United States
Jing Hua Zhao, University of Cambridge, United Kingdom
Anne E. Justice, Geisinger Health System, United States

***Correspondence:**
Pak Chung Sham
pcsham@hku.hk

Specialty section:
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics

Received: *13 April 2018*
Accepted: *02 July 2018*
Published: *03 August 2018*

Citation:
Zhou X, Cheung C-L, Karasugi T, Karppinen J, Samartzis D, Hsu Y-H, Mak TS-H, Song Y-Q, Chiba K, Kawaguchi Y, Li Y, Chan D, Cheung KM-C, Ikegawa S, Cheah KS-E and Sham PC (2018) Trans-Ethnic Polygenic Analysis Supports Genetic Overlaps of Lumbar Disc Degeneration With Height, Body Mass Index, and Bone Mineral Density. Front. Genet. 9:267. doi: 10.3389/fgene.2018.00267

Lumbar disc degeneration (LDD) is age-related break-down in the fibrocartilaginous joints between lumbar vertebrae. It is a major cause of low back pain and is conventionally assessed by magnetic resonance imaging (MRI). Like most other complex traits, LDD is likely polygenic and influenced by both genetic and environmental factors. However, genome-wide association studies (GWASs) of LDD have uncovered few susceptibility loci due to the limited sample size. Previous epidemiology studies of LDD also reported multiple heritable risk factors, including height, body mass index (BMI), bone mineral density (BMD), lipid levels, etc. Genetics can help elucidate causality between traits and suggest loci with pleiotropic effects. One such approach is polygenic score (PGS) which summarizes the effect of multiple variants by the summation of alleles weighted by estimated effects from GWAS. To investigate genetic overlaps of LDD and related heritable risk factors, we calculated the PGS of height, BMI, BMD and lipid levels in a Chinese population-based cohort with spine MRI examination and a Japanese case-control cohort of lumbar disc herniation (LDH) requiring surgery. Because most large-scale GWASs were done in European populations, PGS of corresponding traits were created using weights from European GWASs. We calibrated their prediction performance in independent Chinese samples, then tested associations with MRI-derived LDD scores and LDH affection status. The PGS of height, BMI, BMD and lipid levels were strongly associated with respective phenotypes in Chinese,

but phenotype variances explained were lower than in Europeans which would reduce the power to detect genetic overlaps. Despite of this, the PGS of BMI and lumbar spine BMD were significantly associated with LDD scores; and the PGS of height was associated with the increased the liability of LDH. Furthermore, linkage disequilibrium score regression suggested that, osteoarthritis, another degenerative disorder that shares common features with LDD, also showed genetic correlations with height, BMI and BMD. The findings suggest a common key contribution of biomechanical stress to the pathogenesis of LDD and will direct the future search for pleiotropic genes.

Keywords: polygenic score, genetic correlation, causality, pleiotropy, lumbar disc degeneration, osteoarthritis

INTRODUCTION

Human intervertebral discs (IVDs) are fibrocartilaginous structures that lie between adjacent vertebrae. These IVDs hold the vertebrae together, facilitate some vertebral motion, and act as shock absorbers to accommodate biomechanical loads (Oxland, 2016). IVD is composed of a gel-like nucleus pulposus surrounded by an annulus fibrosis and separated from the vertebral body by a cartilaginous endplate (Humzah and Soames, 1988). During one's lifetime, due to excessive physical loading, occupational injuries, aging, genetics, and other factors, the IVDs may degenerate and display marked biochemical and morphological changes (Buckwalter, 1995; Urban and Roberts, 2003). Currently, magnetic resonance imaging (MRI) is the gold-standard for evaluating disc degeneration. Based on this imaging, numerous methods are available to grade and summarize different features indicative of degeneration, including signal intensity loss, bulging and herniation, as well as disc space narrowing (Battié et al., 2004; Cheung et al., 2009). Lumbar disc degeneration (LDD) is of clinical importance because it is believed to be a major cause of low back pain (Luoma et al., 2000; Livshits et al., 2011; Samartzis et al., 2011; Takatalo et al., 2011). Its severe form lumbar disc herniation (LDH), in which disc material herniates into the epidural space and compresses a lumbar nerve root, can cause neuropathic pains (sciatica) radiating to the lower extremity (Ropper and Zafonte, 2015).

Twin studies have demonstrated a strong genetic contribution to LDD (Sambrook et al., 1999; Battié et al., 2008). However, searching for genetic variants associated with LDD has been a challenge due to discrepancies and non-standardization of phenotype definitions, inconsistencies with imaging technology, and limited sample sizes in genome-wide association studies (GWASs) (Eskola et al., 2012, 2014; Williams et al., 2013). Similar to most other complex traits, LDD is likely to be polygenic with thousands of trait-associated variants each of which has tiny effect size.

In addition to age, sex, and environmental influences, LDD is also associated with several heritable risk factors including body mass index (BMI) (Liuke et al., 2005; Samartzis et al., 2011, 2012; Takatalo et al., 2013), bone mineral density (BMD) (Harada et al., 1998; Pye et al., 2006; Wang et al., 2011), and serum lipid levels (Leino-Arjas et al., 2008; Longo et al., 2011; Zhang et al., 2016). But it is not fully clear if there is a genetic basis for these phenotype associations. Identifying genetic overlaps between LDD and related traits will be useful for elucidating cause and effect because genetic markers are not subject to reverse causation or confounding and can be used as an instrument to infer causality using Mendelian randomization (Davey Smith and Hemani, 2014), and it can also suggest pleiotropic loci that reveal novel insights into biology (Solovieff et al., 2013).

Several methods have been developed to evaluate genetic overlap between traits by exploiting the polygenic architecture (Dudbridge, 2016). A polygenic score (PGS) of a trait is the summation of alleles across loci weighted by their effect sizes estimated from GWAS (Purcell et al., 2009). In its typical application, GWAS of a *base phenotype* is first done in a discovery sample. PGS can be calculated in an independent testing sample using single-nucleotide polymorphisms (SNPs) whose *p*-values are below some threshold in the GWAS of discovery sample. It can then be used as a predictor of *target phenotypes* in the testing sample using regression analysis. PGS has been widely used to predict disease risk (Chatterjee et al., 2016), evaluate genetic overlaps across traits (Krapohl et al., 2016), and infer genetic architectures (Stahl et al., 2012; Palla and Dudbridge, 2015). Because phenotyping of LDD by MRI is expensive and labor intensive, sample sizes are usually limited for well-phenotyped cohorts. PGS can leverage GWAS meta-analysis results from large consortia to maximize the power to detect genetic overlaps and is most suitable for the current study of LDD. Some other methods, such as bivariate linear mixed-effect model (Lee et al., 2012b; Vattikuti et al., 2012) would require genotypes of individuals of base and target phenotypes. Recently developed linkage disequilibrium score (LDSC) regression (Bulik-Sullivan et al., 2015) makes use of only summary-level association statistics and can account for sample overlaps between different studies, but it requires very large sample sizes that has not been available for LDD.

In this study, we applied PGS to investigate the genetic overlap of LDD with four related risk factors using the GWAS data of Hong Kong Disc Degeneration (HKDD) population-based cohort (Cheung et al., 2009; Samartzis et al., 2012; Li et al., 2016) and a Japanese case-control cohort of LDH that required surgery (Song et al., 2013). We selected BMI, BMD and serum lipids levels as base phenotypes, based on their previous reported associations with LDD (Pye et al., 2006; Longo et al., 2011; Samartzis et al.,

2012). Height was also included because its association with chronic low back pain (Hershkovich et al., 2013; Heuch et al., 2015). Two semi-quantitative scores that summarize different aspects of LDD from lumbar spine MRI were used as target phenotypes in the HKDD cohort; LDH affection status was used as the third target phenotype in the Japanese case-control cohort. Because GWASs of base phenotypes were done in European populations whereas our testing samples were of East Asian ancestry, the performance of PGS in predicting base phenotypes was first evaluated in independent Chinese samples. Then we applied the best performing PGS of the base phenotype to test association with target phenotypes in testing samples (**Figure 1**). Results were then interpreted in light of previous epidemiological evidence and statistical power to detect association. To better understand the mechanism implied by the genetic overlaps and motivated by the suggestion that LDD and osteoarthritis (OA) may share common pathophysiological features (Loughlin, 2011; Ikegawa, 2013), we further tested if the base phenotypes that had genetic overlaps with LDD also showed genetic correlations with OA using the GWAS summary data of the arcOGEN study (Zeggini et al., 2012). Finally, we evaluated the predictive power of trans-ethnic PGS to aid the design of future studies.

MATERIALS AND METHODS

Study Samples

HKDD Cohort

The HKDD Study was a population-based cohort of approximately 3,500 Southern Chinese initiated to assess spinal

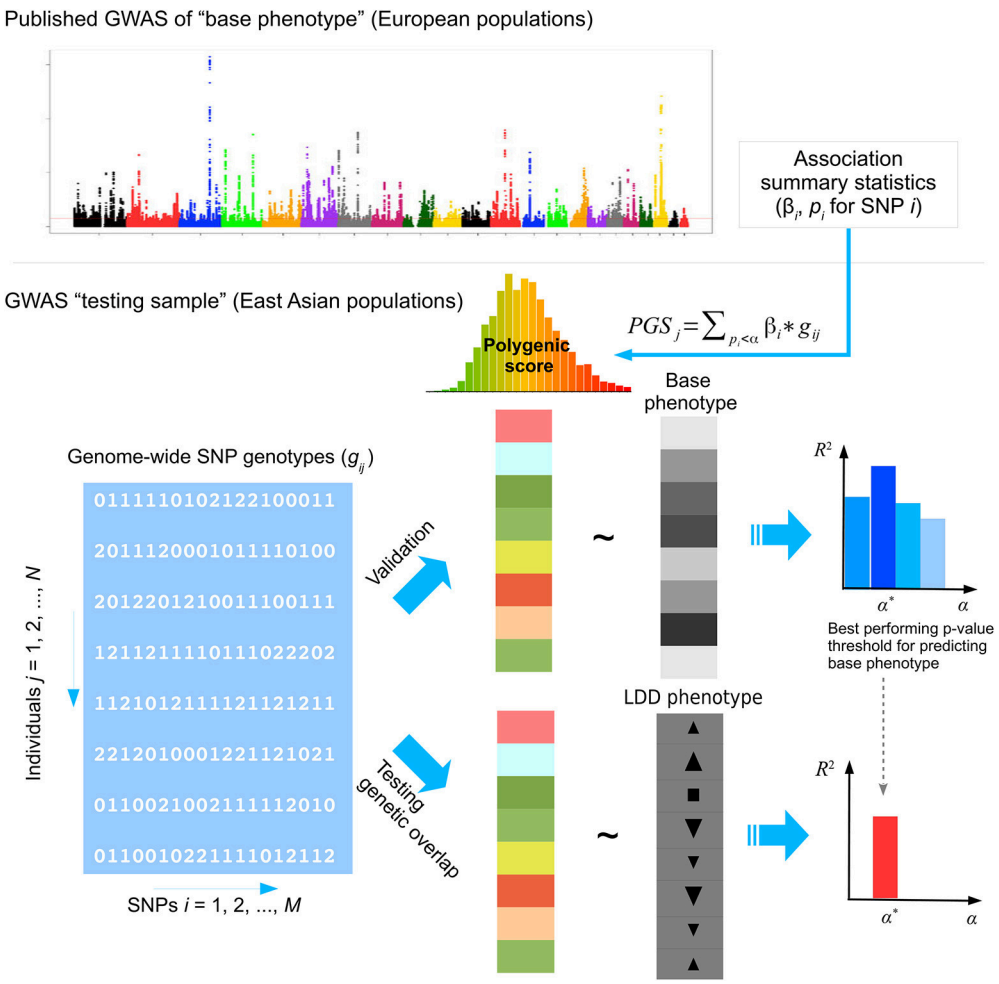

FIGURE 1 | The analysis framework. GWAS summary statistics of base phenotypes were obtained from published studies in European populations. The polygenic score (PGS) in a testing sample of East Asian population was calculated by weighted summation of alleles at approximately independent SNPs whose association *p*-values fall below some threshold in the discovery GWAS done in European populations. The performance of PGS to predict the base phenotype in East Asians was first evaluated in a validation sample. Then the best performing PGS of the base phenotype was used to test genetic overlap with lumbar disc degeneration (LDD) in the testing samples. In this study, we selected height, body mass index (BMI), bone mineral density (BMD) and lipid levels as base phenotypes. The prediction performance of PGS was evaluated in the HKDD cohort for height, BMI, and lipid levels, and in the HKOS cohort for BMD. Three LDD phenotypes were used as target phenotypes, including disc displacement and disc degeneration scores in the HKDD cohort and affection status of lumbar disc herniation (LDH) in the Japanese LDH case-control cohort.

phenotypes and their risk factors. All participants underwent T2-weighted MRI examination of the lumbar spine assessed by expert physicians (JK and KMC). Sample recruitment and MRI procedures have been described in detail previously (Cheung et al., 2009; Samartzis et al., 2012; Li et al., 2016). For the current study, we focused on two major aspects of LDD captured by different MRI features (**Figure 2a**). The first was signal intensity loss within nucleus pulposus, which may represents loss of water content of IVD. Its presence and severity at each lumbar disc was assessed by the Schneiderman's grades (Schneiderman et al., 1987). Based on this grading scheme each disc was given a score of 0–3, whereby 0 indicated normal and higher scores indicated increased severity. A *disc degeneration score* for each individual was calculated by the summation of Schneiderman's grades over all five lumbar discs. We also assessed disc displacement, represented as a bulging/protrusion or extrusion of disc material. An ordinal grade from 0 to 2 was assigned to each lumbar disc to indicate normal, bulge/protrusion or extrusion of disc material; for each individual, a *disc displacement score* was calculated by the summation of the grades over all five lumbar discs (Cheung et al., 2009). Age, sex, physical workload based on occupation, history of smoking, and history of lumbar spine injury were obtained by a questionnaire for all participants. Body height and weight were measured at the time when each subject underwent MRI, and BMI was calculated by dividing weight by height squared (kg/m^2). A subset of the cohort ($N = 815$) also had their blood metabolite profiles measured by quantitative serum nuclear magnetic resonance (NMR) platform (Soininen et al., 2009, 2015). Low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), triglycerides (TG) and total cholesterol (TC) were obtained as part of NMR metabolite measures. Association between LDD scores and other covariates were analyzed using multiple linear regression to account for correlations between predictor variables. The best fitting model was selected using Akaike information criterion.

A total of 2,373 individuals from the HKDD cohort were genotyped by Illumina HumanOmni-ZhongHua-8 Beadchip.

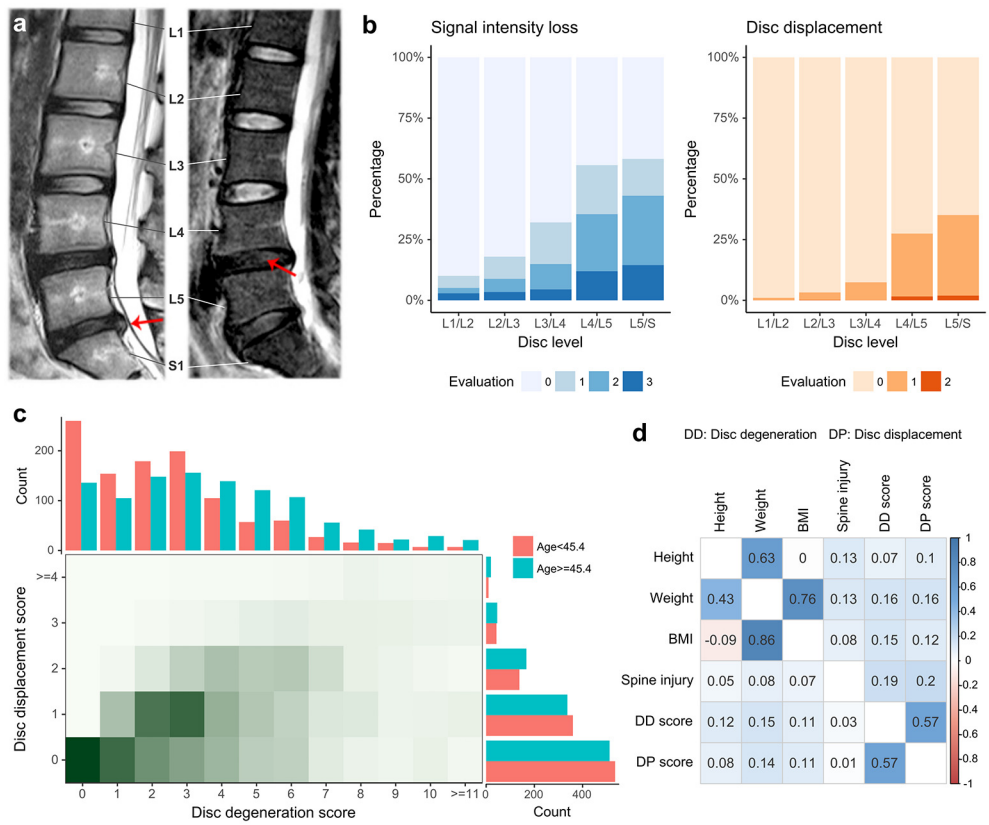

FIGURE 2 | Summary of phenotypes in the HKDD cohort. **(a)** Examples of magnetic resonance imaging show two major aspects of LDD. Disc displacement (left) is shown as bulging of disc material beyond confine of annulus fibrosus. The loss of proteoglycan and water content (right) within nucleus pulposus is reflected by the signal intensity loss. The lumbar spine has 5 intervertebral segments, termed L1 through L5. S1 stands for the first segment of sacral that is intermediately below the lumbar spine. **(b)** The prevalence of signal intensity loss and disc displacement at different levels of lumbar spine discs. Two ordinal grades (0–3 for signal intensity loss, 0–2 for disc displacement) were assigned to each lumbar disc to indicate the presence and severity of LDD, where 0 indicated normal and higher scores indicated increased severity. **(c)** The distribution of two LDD scores. The disc degeneration score and displacement score were defined by the summation of grades over all disc levels for signal intensity loss and disc displacement respectively. The two LDD scores are correlated in the population. The age threshold divides the HKDD cohort in two parts with roughly equal sample sizes. The older subjects tend to have higher disc degeneration scores and disc displacement scores. **(d)** Pairwise Pearson correlations between original phenotypes (upper triangle) and between residual phenotypes after adjusting for age and gender (lower triangle).

Basic genotyping and quality control (QC) procedures have been described in our previous study (Li et al., 2016). In this study, we used more stringent QC criteria that keep only individuals with a call rate >99% and common SNPs with minor allele frequency (MAF) >0.01. The genotypes were imputed to over 8 million common variants in Phase 3 of 1,000 Genomes reference panel using the Michigan Imputation Server (Das et al., 2016) and filtered to keep only common bi-allelic SNPs (MAF>0.01) with imputation quality metrics $r^2 \geq 0.3$.

LDH Case-Control Cohort

The Japanese LDH case-control cohort was part of our previous genetic study of LDD (Song et al., 2013). Hospitalized patients LDH were ascertained on the basis of sciatica or severe low back pain requiring surgical treatment and confirmed by lumbar spine MRI. The controls were unrelated individuals from general Japanese population as part of Japan Biobank Project. All individuals were genotyped by Illumina HumanHap550v3 BeadChip. A total of 366 cases and 3,331 controls passed QC and were used in the association analysis. Genotypes were imputed to 2.5 million SNPs in Phase 2 HapMap Project using IMPUTE2 (Howie et al., 2009), and association analysis at each SNP was performed by logistic regression assuming an additive model using SNPTEST (Marchini et al., 2007).

HKOS GWAS

Hong Kong Osteoporosis Study (HKOS) was a prospective cohort study of over 9,000 Southern Chinese residents in Hong Kong (Cheung et al., 2017). BMD of the lumbar spine (LS-BMD) and femoral neck (FN-BMD) were measured by dual-energy X-ray absorptiometry. The age-corrected and standardized BMD was generated for each gender. A total of 800 unrelated females with extreme BMD were selected in the previous GWAS (Kung et al., 2010). The low BMD subjects were those with BMD Z-score ≤ -1.28 at either the LS or FN; the high BMD subjects were those with BMD Z-score ≥ 1.0 at either of the two skeletal sites. All individuals were genotyped by Illumina HumanHap610Quad Beadchip, whereby 780 individuals passed QC. Association analysis at each SNP was performed by linear regression using PLINK (Purcell et al., 2007). Detailed genotyping, QC, and imputation procedures have been described elsewhere (Kung et al., 2010; Xiao et al., 2012).

arcOGEN Study

The arcOGEN study (http://www.arcogen.org.uk/) was a collection of unrelated, UK-based individuals of European ancestry with knee and/or hip OA from the arcOGEN Consortium (Panoutsopoulou et al., 2011; Zeggini et al., 2012). Cases were ascertained based on clinical evidence with a need of joint replacement or radiographic evidence of disease (Kellgren-Lawrence grade ≥ 2), controls were from ancestry-matched (UK) population. A GWAS meta-analysis that included 7,410 cases and 11,009 controls as the discovery sample has been described previously (Zeggini et al., 2012). The summary statistics of the discovery GWAS were obtained by application to the consortium.

All studies were approved by local ethical committees. Written informed consent was obtained from all participants.

Statistical Analysis

Polygenic Score Regression

We selected height, BMI, BMD, and serum lipid levels as base phenotypes, and obtained GWAS summary data (**Table 1**). PGS was created by the two strategies as described below and used to predict phenotypes through linear regression after accounting for other covariates. The non-genetic covariates were selected based on their association with each phenotype in the baseline multiple linear regression model (listed in footnotes of **Table 3** and **Table S7**). We validated the prediction performance of PGS on base phenotypes in the HKDD cohort (for height, BMI, and lipid levels) and HKOS cohort (for LS- and FN-BMD). Then, the PGS with best prediction performance for each base phenotype was used to test genetic overlap with LDD by predicting two LDD scores in the HKDD cohort and the LDH affection status in the Japanese case-control cohort. When individual-level genotype data in the testing sample were not available (for the HKOS GWAS and LDH case-control cohort), PGS regression was performed using summary statistics based algorithm implemented in gtx R package (Johnson, 2012), and SNP genotypes of HapMap 3 East Asian samples were used as the reference panel for SNP clumping.

TABLE 1 | GWAS summary statistics used in this study.

Phenotype	Study	Publication	Sample size	Population	Availability
Height	GIANT Consortium	Wood et al. (2014)	253,000	European	Public
BMI[§]		Locke et al. (2015)	234,000~322,000[#]	European	Public
Serum lipids	GLGC	Willer et al. (2013)	95,000~189,000[#]	European	Public
BMD	GEFOS Consortium	Estrada et al. (2012)	33,000	European	Application to the consortium
	HKOS[¶]	Kung et al. (2010)	780	Chinese	Contributed by the collaborator
OA	arcOGEN Consortium	Zeggini et al. (2012)	7,400 cases, 11,000 population controls	European	Application to the consortium
Hospitalized LDH	Japan LDH	Song et al. (2013)	366 cases, 3,331 population controls	Japanese	Contributed by the collaborator

BMI, body mass index; BMD, bone mineral density; OA, osteoarthritis; LDH, lumbar disc herniation; GLGC, Global Lipids Genetics Consortium; HKOS, Hong Kong Osteoporosis Study.

[§] *The GIANT consortium's BMI GWAS included samples from multiple ethnicities; only the result of the European samples was used.*

[#] *BMI and lipids summary data were generated from GWAS+Metabochip joint analysis, so sample sizes can vary across different SNPs.*

[¶] *The HKOS GWAS was part of the GEFOS meta-analysis and was the only study of non-European population in that study.*

The first strategy for calculating PGS only used known trait-associated SNPs that reached genome-wide significance in previous studies (GWAS hits). Individual PGS profiles were calculated by summing up the dosage of trait-increasing alleles from imputed genotypes weighted by the reported effect sizes. This strategy has the advantage of including secondary signals within the same locus and increased accuracy of effect size estimates from a larger independent replication sample (for LS- and FN-BMD).

As a second strategy, we performed genome-wide PGS analysis using PRsice (Euesden et al., 2015). Briefly, summary statistics of base GWAS was first aligned with genotyped SNPs of the testing sample. Then SNPs were pruned based on *p*-value informed clumping algorithm [linkage disequilibrium (LD) $r^2 < 0.1$ across 500 kb] that selected SNPs most associated with the base phenotype in a locus to generate sets of independent SNPs. PGS was created using clumped SNPs whose *p*-value in the base GWAS were below pre-specified threshold. We varied *p*-value thresholds (1.0E-7, 1.0E-5, and from 1.0E-4 to 0.5 with a step of 0.0001) to select the one that maximized the variance explained (R^2) for the base phenotype in the validation sample.

Correcting Sample Overlap and Extreme Selection

The HKOS GWAS was part of BMD GWAS meta-analysis conducted by the GEFOS consortium. To make it an independent testing data, we inverted the fixed effect meta-analysis to subtract the contribution of HKOS from the GEFOS summary statistics (**Appendix 1** in the **Supplementary Material**). When calculating PGS using the BMD GWAS hits, we used effect estimates from the stage II replication sample to avoid the issue of overlapping sample.

The HKOS GWAS adopted an extreme phenotype design to increase association power, which also resulted in upward biased estimates of R^2 by PGS using linear regression. To get an estimate of R^2 in the unselected sample ($\hat{R}^2$) for comparing with the previous report, we corrected the R^2 estimate in the selected sample ($\hat{R}^{2'}$) by:

$$\hat{R}^2 \approx \frac{\hat{R}^{2'}}{f - (f-1)\hat{R}^{2'}}$$

where f is the increased phenotype variance due to extreme phenotype selection (=2.739 in the HKOS GWAS sample). The derivation and validation of this approximation formula is given in **Appendix 3** (**Supplementary Material**).

R^2 for Case-Control Data on the Liability Scale

For the case-control data, it is meaningful to estimate the disease liability explained by PGS under the liability threshold model (Falconer and Mackay, 1996), so that the result can be compared to the heritability of LDH (Heikkila et al., 1989). We first converted summary statistics generated by the logistic regression to those of linear regression by first-order approximation. Then we used summary statistics based PGS regression to obtain an estimate of R^2 on the observed scale. Finally, the observed R^2 was converted to the liability scale using the transformation formula by Lee et al. (2012a) assuming the disease prevalence of 0.02 (Jordan et al., 2009). More details are given in **Appendix 4** (**Supplementary Material**).

Inference of Genetic Architecture and Projecting Prediction Performance

We applied AVENGEME (Palla and Dudbridge, 2015) to estimate parameters of genetic architecture of height and BMI from the PGS results. The procedure is described in **Appendix 2** (**Supplementary Material**). Briefly, for a presumed SNP heritability h_1^2, the method estimated the fraction of markers that are null ($\hat{\pi}_0$) and genetic correlation between the discovery and testing samples ($\hat{\sigma}_{12}$). If the genetic architectures of the discovery and testing sample are the same, then the genetic correlation between two samples can be estimated as $\hat{\rho}_G = \hat{\sigma}_{12}/h_1^2$.

The same model was also applied to predict the expected R^2 for height and BMI in Chinese population under different study designs. To project R^2 using PGS created by weights from European GWAS, model parameters were set to the maximum likelihood estimates fitted to the observed PGS results. We also increased the discovery sample size by 500,000 to evaluate the increase of R^2 in the future. To predict R^2 using PGS created by weights from East Asian GWAS, we used the same set of model parameters, but set the discovery GWAS sample size to match the published study of East Asians (Wen et al., 2014; He et al., 2015) and assumed no heterogeneity of effect sizes between the discovery and testing samples ($\sigma_{12} = h_1^2$). To incorporate between-sample heterogeneity within East Asians, we changed between population genetic correlation to 0.9 (so $\sigma_{12} = 0.9h^2$), which is a lower bound for height and BMI in Europeans (de Vlaming et al., 2017) and in different Chinese GWAS samples (data not shown).

SNP Heritabilities

Phenotypes analyzed in the HKDD cohort were adjusted by covariates that are associated with the phenotype (listed in **Table 2**) by linear regression; and residues were inverse normal transformed when necessary. SNP heritabilities of the adjusted phenotypes were estimated using GCTA v1.25 (Yang et al., 2011a) after excluding individuals so that no pair of individuals had estimated coefficient of relatedness >0.05 as recommended by the GCTA developers (Yang et al., 2017).

Estimating Genetic Correlation Between Anthropometric Traits and OA

To test the genetic overlaps of OA with height, BMI and BMD, we applied LDSC regression (Bulik-Sullivan et al., 2015) to the GWAS summary statistics following the recommended procedure. BMD summary statistics were corrected to remove the contribution from the HKOS GWAS (the only non-European study) as in PGS regression.

Due to sample overlaps in different European GWASs, PGS could not have been applied to the genome-wide summary data. But for known BMD associated SNPs whose effect size estimates from an independent replication sample were available (Estrada et al., 2012), we also used PGS regression under summary statistic

TABLE 2 | SNP heritability estimates of phenotypes analyzed in the HKDD cohort.

Phenotype	Adjustment and transformation	$\hat{h}^2$(SE)	*p*-value
Height	Age, sex; inverse normal transformation	0.533 (0.170)	6.75E-05
	Age, sex, first two PCs; inverse normal transform	0.383 (0.182)	1.67E-02
BMI	Age, age^2, sex; inverse normal transformation	0.285 (0.171)	2.97E-02
	Age, age^2, sex, first PC; inverse normal transformation	0.249 (0.174)	6.03E-02
Disc degeneration score	Age, sex, lumbar injury	0.218 (0.163)	6.50E-02
	Age, sex, lumbar injury, height	0.232 (0.169)	6.43E-02
	Age, sex, lumbar injury, BMI	0.226 (0.170)	7.26E-02
	Age, sex, lumbar injury, height, BMI	0.219 (0.171)	8.40E-02
	Age, sex, lumbar injury, weight	0.225 (0.171)	7.80E-02
Disc displacement score	Age, sex, lumbar injury	0.291 (0.176)	4.26E-02
	Age, sex, lumbar injury, height	0.269 (0.180)	6.20E-02
	Age, sex, lumbar injury, BMI	0.238 (0.181)	9.17E-02
	Age, sex, lumbar injury, height, BMI	0.216 (0.182)	1.21E-01
	Age, sex, lumbar injury, weight	0.213 (0.182)	1.22E-01

mode to assess the genetic correlation between osteoarthritis and BMD.

Power Analysis

Power calculation was done assuming the test statistics follows non-central chi-squared distribution under the alternative hypothesis. The non-centrality parameter for quantitative trait is $\frac{NR^2}{1-R^2}$, where N is the sample size and R^2 is the phenotype variance explained by PGS. For binary trait, R^2 in the above formula is on the observed scale and can be converted from liability scale using Lee et al. (2012a)'s formula as described in **Appendix 4** (**Supplementary Material**).

RESULTS

Phenotype Summary of the HKDD Cohort

A total of 2,054 unrelated Chinese subjects in the HKDD cohort (60% were females) were included in polygenic analysis. The basic demographic and phenotype summary are shown in **Table S1**. Both signal intensity loss and disc displacement showed a higher prevalence and severity at lower lumbar levels (**Figure 2b**). The disc degeneration and disc displacement scores for each individual were calculated by the summation of grades over all levels. Consistent with a major effect of aging, older individuals tend to have higher disc degeneration and displacement scores (**Figure 2c**). The two LDD scores were correlated with each other ($r = 0.57$; **Figure 2c**). Both of them were also positively correlated with height, body weight, BMI, and lumbar spine injury ($P < 0.001$; **Figure 2d**, **Table S2A**); the correlation remained significant for all except injury after correcting for the effect of age and gender (**Figure 2d**, **Table S2B**). Multiple linear regression analysis showed that the best fitting models for both disc degeneration and displacement scores included age, sex, lumbar injury, height and BMI as covariates (**Table S3**), which together explained 21.5 and 9.6% phenotype variances respectively. The SNP heritability estimates for height and BMI in the HKDD cohort were 0.38 (±0.18) and 0.25 (±0.17), similar to the previous reports in Europeans (Yang et al., 2010, 2011b). For disc degeneration and displacement scores, after adjusting for known covariates, SNP heritability estimates were about 0.2∼0.3 (**Table 2**).

Evaluating the Prediction Performance of PGS of Anthropometric Traits

We first evaluated prediction performance of PGS in Chinese samples and compared them with Europeans. PGS profiles of height and BMI were created in the HKDD cohort using known trait-associated SNPs identified by the GIANT consortium GWAS meta-analyses. They explained 5.7 and 1.2% of height and BMI variances respectively ($P < 1.0E\text{-}10$ for height, $P = 1.6E\text{-}07$ for BMI), after adjusting for age, sex and principle components. It is 2∼3-fold lower than previous reports in independent European samples, which were 16% for height (Wood et al., 2014) and 2.7% for BMI (Locke et al., 2015). The prediction performance of BMD associated SNPs reported by GEFOS consortium were tested in the HKOS GWAS sample (Kung et al., 2010). After correcting for extreme phenotype selection (**Appendix 3** in the **Supplementary Material**), the known BMD-associated SNPs explained 3.4% and 3.0% variance of LS-BMD and FN-BMD in Chinese population ($P < 1.0E\text{-}10$), also lower than previous reported ∼5% in Europeans (Estrada et al., 2012).

Since GWAS hits may only explain a small proportion of phenotype variance, we extended PGS analysis to make use of whole-genome summary statistics (**Figure 3**). As the *p*-value threshold of the discovery GWAS increases, both true and false positive SNPs will be included in the PGS. The *p*-value threshold that optimized phenotype prediction depends on the discovery sample size and unknown genetic architecture (Chatterjee et al., 2013; Dudbridge, 2013), and should be determined empirically. At the optimal *p*-value threshold, we found the phenotype variance explained is similar to that using GWAS hits for FN-BMD, marginally improved for height, slightly worse for LS-BMD, and more than doubled for BMI ($R^2 = 2.6\%$, $P < 1.0E\text{-}10$). Theoretical model fitting under a range of plausible parameters (**Table S4**) suggested that BMD had smaller fraction of trait associated SNPs (0.5∼0.6%) with larger effect sizes compared with BMI and height (estimated fraction of non-null markers: 14∼17%), which explained why sparse PGS models showed better prediction performance for BMD. The trait variances explained by PGS predicted by the models generally captures the trend of empirical observations at different *p*-value thresholds (**Figure 3**). The estimate of between-population genetic covariance for

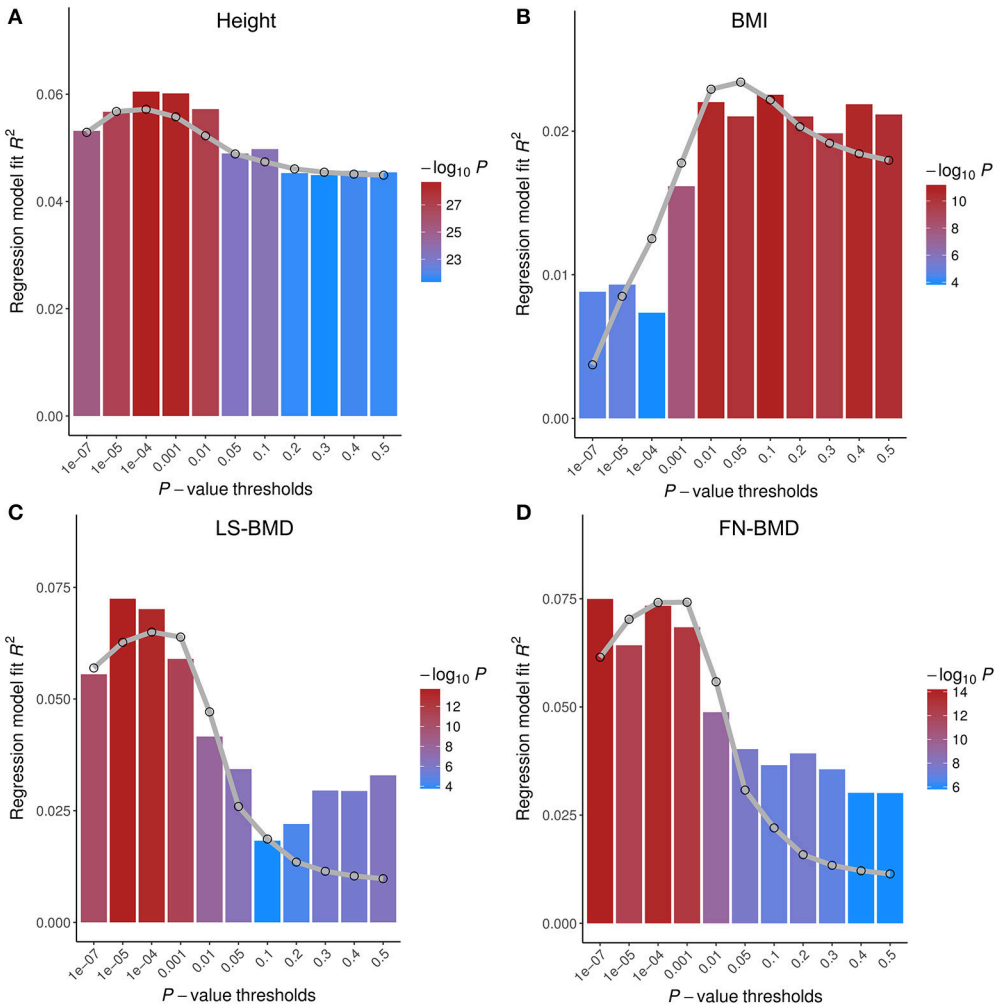

FIGURE 3 | Prediction performance of polygenic scores (PGS) on four base phenotypes in Chinese population. Phenotype variances explained (R^2) are shown at different *p*-value thresholds. The gray lines are the predicted R^2 based on the theoretical model of Dudbridge (2013) with parameters given in **Table S4**. Different parameter sets of each model give similar results. PGS of height **(A)** and BMI **(B)** were tested on HKDD cohort; PGS of bone mineral density at the lumbar spine (LS-BMD, **C**), and femoral neck (FN-BMD, **D**), were evaluated on HKOS sample.

each phenotype was consistently lower than the presumed heritability (**Table S4**), reflecting trans-ethnic heterogeneity in effect sizes.

Testing Genetic Overlap Between Anthropometric Traits and LDD

We then applied PGS to test genetic overlaps between anthropometric traits and LDD (**Table 3**). In the HKDD cohort, the BMI PGS at its optimal threshold was positively associated with both disc displacement score ($R^2 = 0.29\%$, $P = 0.015$) and disc degeneration score ($R^2 = 0.31\%$, $P = 0.011$) after adjusting for sex, age and lumbar injury. The results are consistent with obesity as a major risk factor for LDD development and progression (Hassett et al., 2003; Hangai et al., 2008). The associations remained significant ($P < 0.05$) after further adjusting for height but disappeared after adjusting for BMI or body weight (**Table S5**). The PGS of LS-BMD were positively associated with disc displacement score ($R^2 \approx 0.2\%$; $P<0.05$) and remained significant ($P < 0.05$) after further adjusting for height, BMI or weight (**Table S6**). The same trend was also observed for FN-BMD but did not reach significance. The finding supports the previous reported genetic correlation between BMD and disc bulge in a twin study (Livshits et al., 2010). The lack of association with disc degeneration score is also consistent with the previous study that showed a smaller effect size between BMD and disc signal intensity on MRI (Livshits et al., 2010).

In addition to LDD scores in the general population, we also applied PGS to predict case-control status of symptomatic LDH (Song et al., 2013). The height PGS was positively associated with LDH ($P < 0.01$) and explained 0.35% of disease liability. The association cannot be explained by body weight, because the BMI PGS is better associated with weight but does not show association with LDH ($P > 0.5$). The result provides

TABLE 3 | Genetic overlap of lumbar disc degeneration with anthropometric traits.

Base phenotype	SNP Predictor	*N*(GWAS Sample)[a]	*N*(SNPs)[b]	R^{2c}	Association with disc displacement score[d]			Association with disc degeneration score[d]			Association with lumbar disc herniation requiring surgery		
					sgn(β)	R^2	*p*-value	sgn(β)	R^2	*p*-value	sgn(β)	R^{2e}	*p*-value
Height (Wood et al., 2014)	Known loci	253,000	622	5.76%	+	0.003%	8.04E-01	+	0.135%	9.40E-02	+	0.351%	7.73E-03
	PGS $P \leq 2.0E-04$	3,933	6.69%	+	0.023%	4.88E-01	+	0.162%	6.95E-02	+	**0.341%**	**8.79E-03**
BMI (Locke et al., 2015)	Known loci	234,000~322,000	92	1.24%	+	0.016%	5.46E-01	+	0.053%	2.94E-01	–	0.000%	9.80E-01
	PGS $P \leq 1.23E-02$		4,238	2.64%	+	**0.287%**	**1.45E-02**	+	**0.311%**	**1.08E-02**	+	0.011%	6.44E-01
Lumbar Spine BMD (Estrada et al., 2012)	Known loci	46,000	60	9.33%[f]	+	**0.222%**	**3.16E-02**	+	0.000%	9.92E-01	+	0.093%	1.71E-01
	PGS $P \leq 1.0E-05$	32,000	109	7.53%[f]	+	0.200%	4.10E-02	+	0.009%	6.66E-01	+	0.136%	9.80E-02
Femoral Neck BMD (Estrada et al., 2012)	Known loci	51,000	60	7.78%[f]	+	0.062%	2.55E-01	–	0.017%	5.51E-01	+	0.073%	2.24E-01
	PGS $P \leq 7.0E-04$	33,000	563	8.88%	+	0.153%	7.43E-02	+	0.044%	3.36E-01	+	0.187%	5.27E-02

For the four base phenotypes, we adopted two strategies to create polygenic profiles to predict target phenotypes: using known trait-associated SNPs and their effect size estimates or using independent SNPs of GWAS summary statistics selected based on the optimal p-value threshold for predicting the base phenotype. Nominally significant associations with LDD phenotypes are shown in bold for PGS with best prediction performance for the base phenotype.

[a]Sample size of discovery sample GWAS. For known BMD-associated loci, shown are the sample size of second stage replication cohort.

[b]Number of SNPs in the HKDD cohort that passed QC and have minor allele frequency ≥0.01.

[c]Variance of base phenotype in Chinese population explained by the PGS. For height and BMI, R^2 was estimated in the HKDD cohort ($N = 2,054$). Height was adjusted for age, sex, and the first two principle components (PCs); BMI was adjusted for sex, age, age^2, and the first PC. The residuals were then inverse normal transformed. For BMD, it was estimated in the HKOS GWAS sample (Kung et al., 2010; $N = 780$ females). LS-BMD and FN-BMD were adjusted for age and standardized into Z-scores.

[d]Association with disc displacement and degeneration scores were evaluated by inclusion of polygenic profile score as a covariate to the multiple linear regression model of target phenotype that adjusted for age, sex and lumbar spine injury.

*[e]For LDH, R^2 is the variance of disease liability explained by the PGS (**Appendix 4** in the **Supplementary Material**).*

*[f]The HKOS GWAS sample were selected from extreme ends of BMD distribution. After correcting for extreme-selection (**Appendix 3** in the **Supplementary Material**), we estimate that R^2 by the PGS of GWAS hits is 3.52% for LS-BMD and 2.99% for FN-BMD.*

a genetic basis to the previous epidemiological observation that being tall is a risk factor for hospitalization due to LDH (Wahlstrom et al., 2012) and back surgery (Coeuret-Pellicer et al., 2010).

Testing Genetic Overlap Between Lipid Levels and LDD

Previous studies also reported that increased level of LDL-C, TC, and TG were associated with increased risk of LDH (Leino-Arjas et al., 2008; Longo et al., 2011; Zhang et al., 2016). To test if serum lipid levels have genetic correlation with LDD, we did similar PGS analysis using known lipid associated SNPs and GWAS summary data from the Global Lipids Genetic Consortium (Willer et al., 2013). Prediction performance of PGS was first evaluated in a subset of the HKDD cohort ($N = 620$ with genotypes) whose lipid levels were measured by the high-throughput NMR approach. All PGS were significantly associated with the corresponding lipid levels. Except for LDL-C, the PGS of known lipid loci showed the best prediction performance (**Figure S1**). However, none of them was significantly associated with LDD scores with the expected direction in the HKDD cohort or LDH in the Japanese case-control cohort (**Table S7**). Directly testing the phenotype association in the HKDD cohort by multiple linear regression also showed no association (**Table S8**). Therefore, our data does not support the previously suggested role of atherosclerotic lipids in LDD.

Power Consideration

Given sample sizes and study designs, the two testing samples used in this study show similar profiles of statistical power (**Figure 4**). We have >50% power (at significance level $\alpha = 0.05$) to detect genetic correlation if PGS explains >0.2% variance of adjusted LDD scores (or LDH disease liability). To achieve the same power at $\alpha = 0.01$, it would require PGS to explain >0.33% phenotype (liability) variance. But the current study does not have enough power to detect genetic overlap if the PGS explain less than 0.2% variance of LDD scores (or LDH liability). Therefore, we designed this study to only test phenotypes with previous epidemiological evidence for association with LDD. To further reduce the multiple testing burden, we had only used the PGS which was optimal in predicting the corresponding base phenotype to test the genetic overlap with LDD. The results that were nominally significant and consistent with the expected phenotype correlations can be interpreted as *supportive evidence of genetic overlaps*.

The Association of a Height Associated SNP rs6651255 With LDD Scores

The current study has no power to search for individual SNPs showing pleiotropic associations with LDD and related traits. But we noted that a recent GWAS of LDH with lumbar spine surgery in Iceland population (Bjornsdottir et al., 2017) identified a genome-wide significant SNP rs6651255, which was also a known height associated SNP (Wood et al., 2014). The risk allele T showed an odds ratio of 1.23 and associated with increased height. The same study also reported that increase in the genetically determined height increased the risk of LDH with

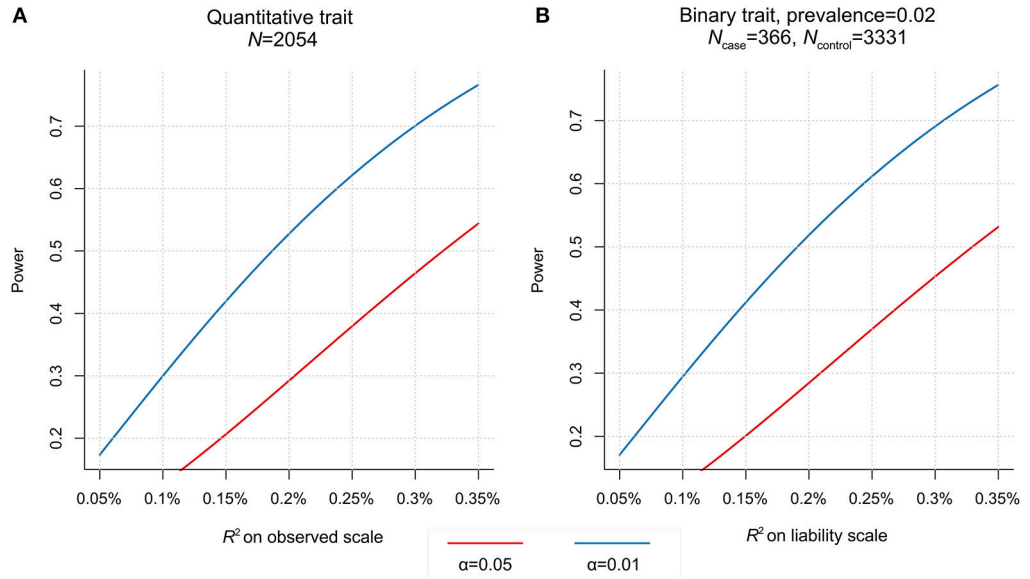

FIGURE 4 | Power to detect association with polygenic score (PGS) in two testing samples. **(A)** For the HKDD cohort ($N = 2{,}054$), given significance level (at $\alpha = 0.01$ or 0.05), the power is determined by the phenotype variance explained by PGS. **(B)** For the Japanese case-control cohort (366 cases, 3,331 controls), assuming the disease prevalence of 0.02, the power is a function of the disease liability explained by PGS.

surgery, but the effect of rs6651255 on LDH was not mediated by height. To replicate this finding in our cohorts, we found an LD proxy rs4733724 (LD $r^2 = 1$ with rs6651255 in 1000 Genomes CEU population) was directly genotyped in the HKDD cohort and reliably imputed in the Japanese case-control cohort. The allele A was coupled to the LDH risk allele and significantly increased both disc displacement and disc degeneration scores ($P < 0.05$; **Table 4**). The effects remained significant after further adjusting for height, BMI or body weight. The same allele was also weakly associated with increased height and increased risk of LDH requiring surgery (odds ratio = 1.11), but the results were not significant as the sample sizes limited the power to detect associations with small effect sizes.

Genetic Correlation Between Anthropometric Traits and OA

Finally, LDD has been suggested to share common features with OA which is also known as degenerative joint disease (Loughlin, 2011; Ikegawa, 2013). To test if osteoarthritis also showed genetic overlaps with the same set of traits as LDD, we assessed the genetic correlations of BMI, BMD and height with osteoarthritis using LDSC regression (**Table 5**). Significant positive genetic correlation was found between BMI and osteoarthritis ($\hat{r}_G = 0.255$, $P = 4.0\text{E-}07$), which is expected given the strong evidence for a causal role of BMI (Panoutsopoulou et al., 2014). Suggestive positive genetic correlations with osteoarthritis were also observed for height ($P = 9.5\text{E-}03$) and LS-BMD ($P = 0.012$) but not for FN-BMD ($P > 0.1$).

The genetic correlation between osteoarthritis and LS-BMD was less significant though its effect was stronger than between osteoarthritis and height, which was possibly due to smaller sample size of the BMD GWAS. Since the genetic architecture of BMD was dominated by fewer number of causal SNPs with larger effect sizes (**Table S4**), it is also possible that LDSC which assumed an infinitesimal model may be less optimal to detect genetic correlations for BMD and other traits. To support this, we calculated PGS of BMD GWAS hits using weights from the second stage replication sample of the GEFOS consortium (Estrada et al., 2012) to predict OA. The PGS of both LS-BMD and FN-BMD were strongly associated with OA case-control status in the acrOGEN sample ($R^2 = 0.13\%$, $P = 7.8\text{E-}07$ for LS-BMD and $R^2 = 0.12\%$, $P = 2.5\text{E-}06$ for FN-BMD). Taken together, the results suggest that like LDD, OA also shares genetic overlaps with height, BMI and BMD.

DISCUSSION

Between-Population Heterogeneity and Its Impact on Prediction Performance of PGS

In this study, we adopted a trans-ethnic PGS strategy to evaluate the genetic overlaps between different traits where GWAS of base phenotypes were done in Europeans and validation and testing samples were East Asians. Although most GWAS findings were generally replicated in populations different from the initial discovery, heterogeneity commonly existed in the estimated effect sizes (e.g., Carlson et al., 2013; Marigorta and Navarro, 2013), which would reduce the power of PGS to predict phenotypes in populations from a different ethnicity (e.g., Johnson et al., 2015). Consistent with this, we found in Chinese validation samples that variance of height, BMI and BMD explained by the PGS of corresponding GWAS hits were all lower than in Europeans. For height and BMI,

TABLE 4 | Association of rs4733724-A allele with lumbar disc degeneration and height in East Asian samples.

Phenotype	Covariates	$\hat{\beta}$ (SE)	*p*-value
Disc displacement score ($N = 2{,}054$)	Age, sex, lumbar injury	0.078 (0.032)	1.61E-02
	Age, sex, lumbar injury, BMI	0.074 (0.033)	2.28E-02
	Age, sex, lumbar injury, height	0.074 (0.033)	2.43E-02
	Age, sex, lumbar injury, weight	0.072 (0.032)	2.68E-02
	Age, sex, lumbar injury, height, BMI	0.071 (0.032)	2.91E-02
Disc degeneration score ($N = 2{,}054$)	Age, sex, lumbar injury	0.182 (0.089)	4.16E-02
	Age, sex, lumbar injury, BMI	0.187 (0.090)	3.73E-02
	Age, sex, lumbar injury, height	0.183 (0.090)	4.16E-02
	Age, sex, lumbar injury, weight	0.180 (0.089)	4.33E-02
	Age, sex, lumbar injury, height, BMI	0.176 (0.089)	4.83E-02
Hospitalized LDH (366 cases, 3,331 controls)	NA	0.105[a] (0.081)	1.96E-01
Height ($N = 2{,}050$)	Age, sex, PC1, PC2	0.032 (0.035)	3.59E-01

The SNP rs4733724 was genotyped in the HKDD cohort and reliably imputed in the Japanese LDH case-control cohort. The A allele was previously reported to be associated with increased height in Europeans (Wood et al., 2014). The rs4733724-A allele is coupled to rs6651255-T, the latter of which was recently found to increase the risk (odds ratio = 1.23) of LDH requiring surgery in Icelanders (Bjornsdottir et al., 2017). The frequency of rs4733724-A allele is 0.72 in East Asians and 0.23 in Europeans.
[a]Odds ratio = 1.11.

TABLE 5 | Genetic correlations estimated by LD-score regression.

Trait 1	Trait 2	$\hat{r}_G$ (SE)	*p*-value
GENETIC CORRELATIONS BETWEEN ANTHROPOMETRIC TRAITS			
Height	BMI	−0.055 (0.022)	1.36E-02
Height	LS-BMD	0.071 (0.032)	2.72E-02
Height	FN-BMD	0.036 (0.033)	2.83E-01
BMI	LS-BMD	0.067 (0.028)	1.75E-02
BMI	FN-BMD	0.071 (0.028)	9.90E-03
LS-BMD	FN-BMD	0.669 (0.032)	4.64E-96
GENETIC CORRELATIONS BETWEEN OA AND ANTHROPOMETRIC TRAITS			
OA	BMI	0.255 (0.050)	4.02E-07
OA	Height	0.117 (0.045)	9.50E-03
OA	LS-BMD	0.192 (0.076)	1.18E-02
OA	FN-BMD	0.094 (0.068)	1.63E-01

BMI, body mass index; BMD, bone mineral density; LS, lumbar spine; FN, femoral neck; OA, osteoarthritis.

assuming the genetic architecture is the same between European and Chinese, the observed PGS results suggest that between-population genetic correlations are about 0.4∼0.6 (**Table S4**, Materials and Methods). The rough estimations are within the range of previous estimate for type 2 diabetes and rheumatoid arthritis between European and East Asian using a different methodology (Brown et al., 2016).

The use of European GWAS in the current study is mainly due to large sample sizes and publicly available summary statistics. GWAS meta-analyses of height and BMI were also conducted in East Asians (Wen et al., 2014; He et al., 2015). Although sample sizes are much smaller (N≈36,000 for height, 87,000 for BMI), they are expected to have more similar effect sizes to the Chinese sample. To evaluate the tradeoff between sample size and effects heterogeneity, we projected expected prediction performance (R^2) of height and BMI using a theoretical model with parameters of genetic architecture compatible with the observed PGS results (Materials and Methods). Despite smaller sample sizes, using East Asian GWAS as the discovery sample is expected have comparable maximum R^2 as European GWAS to predict height in the Chinese population (**Figure 5**). For BMI, depending on the presumed SNP heritability, using East Asian GWAS shows comparable or better maximum R^2 (**Figure 5**, **Figure S2**). Further increase the European GWAS sample sizes by half a million, a scale similar to the on-going UK biobank study, the increase in R^2 for height is capped at 8% but roughly doubles for BMI. Notably, when using East Asian GWAS as the discovery sample, the best prediction performance can only be achieved at p-value thresholds >0.01. However, whole-genome summary statistics of East Asian GWASs were not publicly available for us before the start of this study. Also consistent with the theoretical predictions, incorporating East Asian GWAS top hits to the PGS of GWAS hits only marginally increased R^2 in predicting height and BMI, and their associations with the LDD scores remained insignificant (**Table S9**).

Given genetic architecture and sample sizes, the power of PGS in detecting genetic overlaps is mainly determined by the performance PGS in predicting the corresponding base phenotype. Therefore, the theoretical results suggest that the use of European GWAS as discovery sample in PGS analysis can still be a favorable approach in cross-trait analysis in the East Asian population. But we caution that the trans-ethnic PGS strategy may not be suitable for other populations like African. Nevertheless, whenever possible ancestry-matched GWAS of base phenotype with large sample sizes should be used to improve the power. Since summary data from large scale GWAS in non-European populations have started to become available recently (e.g., Akiyama et al., 2017), new method will be needed to integrate GWAS data from multiple ethnicities to further improve the PGS prediction performance.

The Influence of Phenotype Definition

In this study, we analyzed three LDD phenotypes, including two semi-quantitative scores derived from MRI assessment and one clinically defined symptom. The PGS of height, BMI and BMD were associated with at least one LDD phenotype. It highlights the complexity in operationally defining LDD, as the current diagnostic approach only captures certain aspects of the degenerative process. Therefore, comparison between different studies should clarify how phenotypes are defined. And it will be

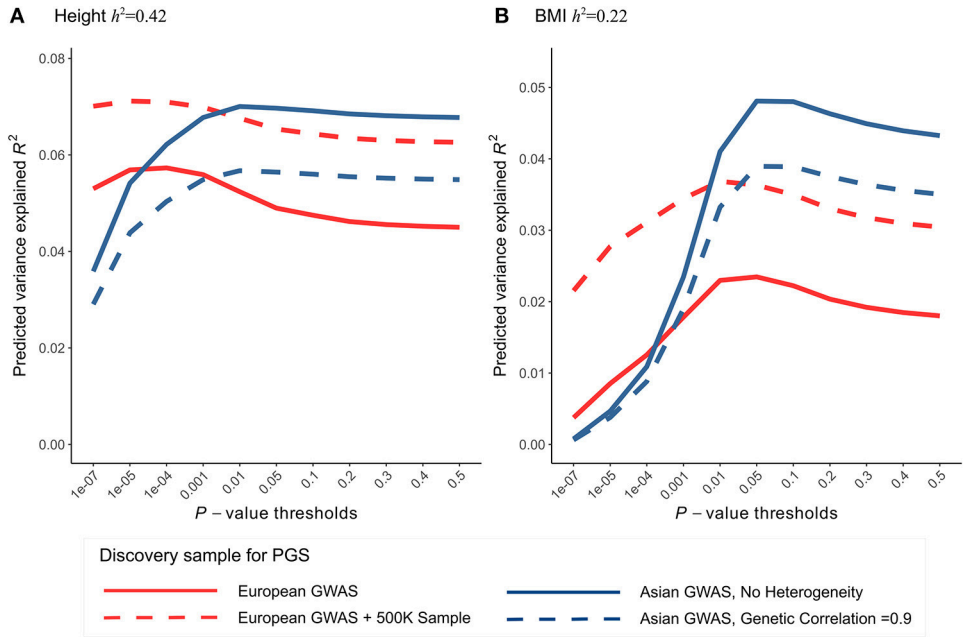

FIGURE 5 | Comparing the prediction performance of polygenic score (PGS) in Chinese sample using GWAS of European and East Asian. Expected phenotype variances explained by PGS (R^2) were calculated using the theoretical model of Dudbridge (2013) with parameters compatible with the observed PGS results of height and BMI. **(A)** For height, the latest European GWAS has sample size 252K. Assuming SNP heritability of 0.42 in both populations, expected R^2 in Chinese population (red line) was calculated using the parameters best fit to **Figure 3**. In comparison, the latest East Asian GWAS has sample size only 36K, expected R^2 was calculated using the same set of parameters except that we assumed no heterogeneity in effect sizes (i.e., genetic correlation = 1) between discovery and testing sample (blue line). To predict the gain in R^2 when using even larger European GWAS in the future, we further increased the discovery GWAS sample size by 500K (red dashed line). We also relaxed the assumption of no heterogeneity within East Asian and calculate expected R^2 assuming genetic correlation of 0.9 (blue dashed line). **(B)** For BMI, European GWAS has sample size 234K; East Asian GWAS has sample size 87K. Expected R^2 were calculated similarly as height, assuming SNP heritability of 0.22.

fruitful to jointly evaluate multiple MRI features in future genetic studies. However, although MRI is the current gold standard that gives best resolution in defining LDD, it is too expensive to be carried out in large samples.

An alternative strategy is to use the a "proxy phenotype" such as patient-based LDH in which large number of cases can be identified based on electronic medical records. Use of proxy phenotype has been demonstrated to improve the power in GWAS (e.g., Okbay et al., 2016). Increase in sample sizes can outweigh the dilution of genetic effects, but it may also capture certain aspects of the trait that is irrelevant to the phenotype of interest (e.g., Kong et al., 2017). In the current and our previous study (Song et al., 2013), LDH requiring surgery was presumed to represent an extreme end of disc displacement in the population. In this regard, it is surprising that the PGS of height strongly associated with LDH but not LDD scores, and PGS of BMI and BMD were associated with LDD scores but not LDH. Although the lack of expected associations can be false negatives due to insufficient power, we cannot rule out the possibility that ascertainment of LDH patients based on severe low back pain or sciatica may enrich polygenic factors other than LDD.

Biological Interpretations

The observed genetic overlaps can be explained by either causality or genetic pleiotropy or both. Interestingly, BMI, BMD and height also showed suggestive evidence of positive genetic correlation with OA. It is possible that they can be explained by some common mechanisms. Although formal assessment of causality could utilize the Mendelian randomization paradigm in larger sample sizes, PGS can be used to nominate candidate phenotypes (Evans et al., 2013). Overweight or obesity has been established as one of the major risk factors for the development and progression of both LDD (Hassett et al., 2003; Hangai et al., 2008) and OA (Bierma-Zeinstra and Koes, 2007). It is commonly believed that increased body weight or BMI exerts more physical loading to the IVD and vertebral endplate (Videman et al., 2007) or joint cartilage (Guilak, 2011), and leads to increased wear and tear of the structures. For BMD, in addition to its correlation with LDD, previous studies also found the increase in BMD in OA patients and an inverse association between OA and osteoporosis (Hannan et al., 1993; Arden et al., 1996). It was postulated that increased BMD is associated with a loss of resilience of subchondral bone which may results in increased mechanical stress on joint cartilage (Foss and Byers, 1972; Radin and Rose, 1986) and similarly on IVD (Harada et al., 1998). The causal mechanism of tall stature on LDH that leads to hospitalization or surgery remains unclear. One possibility may be related to increased disc height, because a previous study using finite element modeling demonstrated that discs with taller height and smaller area were prone to larger motion, higher annular fiber stress and larger degree of disc displacement (Natarajan and Andersson, 1999). Another possibility may be

altered spinal alignment in taller individuals that predispose them to lumbar spine injury. Notably, the postulated mechanisms all point to the pathophysiological role of biomechanical stress. Some other mechanisms have also been proposed (Katz et al., 2010; Samartzis et al., 2013). For example, obesity is also believed to lead to local inflammatory response of secondary mediators secreted by adipocytes known as adipokines. The causal role of adipokines and inflammatory markers can also be tested using their genetic predictors as instrumental variables in future studies.

Alternatively, the observed genetic correlations are also consistent with the genetic pleiotropy and shared pathways among skeletal phenotypes. In supporting this notion, several individual OA associated SNPs were associated with height or BMD (Reynard and Loughlin, 2013; Hackinger et al., 2017), and OA and LDD were found to share some common genetic risk factors (Song et al., 2008; Williams et al., 2011). At single SNP level, we also replicated the recent finding of Bjornsdottir et al. (2017) and showed that the height-increasing allele SNP rs6651255 was associated with the increase of two LDD scores in the HKDD cohort. The previous study did not find association of the same SNP with other related skeletal phenotypes like OA of the spine or osteoporotic vertebral fractures and suggested that the association was driven by the neuropathic pain rather than herniated lumbar discs. However, they did not examine the association of the SNP with radiologically defined LDD phenotypes. Our results in the large population-based cohort with MRI assessment suggest that the same SNP also influences the changes in composition and morphology of lumbar discs. Future genetic studies on LDD with larger sample sizes should search for additional pleiotropic SNPs to better understand bone-cartilage relationships.

In summary, the current study is the first attempt to evaluate genetic overlap between LDD and related traits using GWAS data. Our trans-ethnic polygenic analysis supports the genetic correlations of height, BMI and BMD with LDD, and sheds new light on understanding the pathological mechanism of degenerative skeletal disorders.

DATA AVAILABILITY

The genome-wide association summary statistics of the HKDD cohort is available at https://goo.gl/6gpt9g.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the ethical principles and guidelines for the protection of human participants of research, Human Research Ethics Committee, The University of Hong Kong. The protocol was approved by the Human Research Ethics Committee, The University of Hong Kong. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

AUTHOR CONTRIBUTIONS

XZ and PCS conceived the study and coordinated the research. KS-EC obtained the funding. XZ developed the methods and performed analysis. JK and KM-CC evaluated MRI images of the HKDD cohort. TK, Y-QS, KC, YK, and SI contributed the Japanese LDH GWAS summary data. C-LC contributed the HKOS GWAS summary data. Y-HH contributed the GEFOS consortium GWAS summary data. DS contributed the NMR data. YL and DC applied for the arcOGEN consortium GWAS summary data. XZ drafted the manuscript with inputs from PCS, KS-EC. DS, TS-HM, C-LC, and JK reviewed and revised manuscript. PCS and KS-EC participated discussion.

ACKNOWLEDGMENTS

This work was supported by Research Grant Council of Hong Kong Theme-based Research Scheme Functional Analyses of How Genomic Variation Affect Personal Risk for Degenerative Skeletal Disorders (T12-708/12N), and General Research Fund 776513M, 17128515 and 17124027. GEFOS study was funded by the European Commission (HEALTH-F2-2008-201865-GEFOS). arcOGEN study was funded by a special purpose grant from Arthritis Research UK (grant 18030). We thank Ms. Pei Yu for curating the HKDD phenotype database, and Dr. Eleftheria Zeggini for providing the arcOGEN GWAS summary data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00267/full#supplementary-material

REFERENCES

Akiyama, M., Okada, Y., Kanai, M., Takahashi, A., Momozawa, Y., Ikeda, M., et al. (2017). Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nat. Genet.* 49, 1458–1467. doi: 10.1038/ng.3951

Arden, N. K., Griffiths, G. O., Hart, D. J., Doyle, D. V., and Spector, T. D. (1996). The association between osteoarthritis and osteoporotic fracture: the Chingford Study. *Br. J. Rheumatol.* 35, 1299–1304. doi: 10.1093/rheumatology/35.12.1299

Battié, M. C., Videman, T., Levälahti, E., Gill, K., and Kaprio, J. (2008). Genetic and environmental effects on disc degeneration by phenotype and spinal level: a multivariate twin study. *Spine* 33, 2801–2808. doi: 10.1097/BRS.0b013e31818043b7

Battié, M. C., Videman, T., and Parent, E. (2004). Lumbar disc degeneration: epidemiology and genetic influences. *Spine* 29, 2679–2690. doi: 10.1097/01.brs.0000146457.83240.eb

Bierma-Zeinstra, S. M., and Koes, B. W. (2007). Risk factors and prognostic factors of hip and knee osteoarthritis. *Nat. Clin. Pract. Rheumatol.* 3, 78–85. doi: 10.1038/ncprheum0423

Bjornsdottir, G., Benonisdottir, S., Sveinbjornsson, G., Styrkarsdottir, U., Thorleifsson, G., Walters, G. B., et al. (2017). Sequence variant at 8q24.21 associates with sciatica caused by lumbar disc herniation. *Nat. Commun.* 8:14265. doi: 10.1038/ncomms14265

Brown, B. C., Consortium Asian Genetic Epidemiology Network Type 2 Diabetes, Ye, C. J., Price, A. L., and Zaitlen, N. (2016). Transethnic

genetic-correlation estimates from summary statistics. *Am. J. Hum. Genet.* 99, 76–88. doi: 10.1016/j.ajhg.2016.05.001

Buckwalter, J. A. (1995). Aging and degeneration of the human intervertebral disc. *Spine* 20, 1307–1314. doi: 10.1097/00007632-199506000-00022

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P. R., et al. (2015). An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* 47, 1236–1241. doi: 10.1038/ng.3406

Carlson, C. S., Matise, T. C., North, K. E., Haiman, C. A., Fesinmeyer, M. D., Buyske, S., et al. (2013). Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. *PLoS Biol.* 11:e1001661. doi: 10.1371/journal.pbio.1001661

Chatterjee, N., Shi, J., and García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* 17, 392–406. doi: 10.1038/nrg.2016.27

Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S. J., and Park, J. H. (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* 45, 400-5–405e1-3. doi: 10.1038/ng.2579

Cheung, C. L., Tan, K. C. B., and Kung, A. W. C. (2017). Cohort profile: the Hong Kong Osteoporosis Study and the follow-up study. *Int. J. Epidemiol.* 47, 397–398f. doi: 10.1093/ije/dyx172

Cheung, K. M., Karppinen, J., Chan, D., Ho, D. W., Song, Y. Q., Sham, P., et al. (2009). Prevalence and pattern of lumbar magnetic resonance imaging changes in a population study of one thousand forty-three individuals. *Spine* 34, 934–940. doi: 10.1097/BRS.0b013e3181a01b3f

Coeuret-Pellicer, M., Descatha, A., Leclerc, A., and Zins, M. (2010). Are tall people at higher risk of low back pain surgery? A discussion on the results of a multipurpose cohort. *Arthritis Care Res.* 62, 125–127. doi: 10.1002/acr.20023

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., et al. (2016). Next-generation genotype imputation service and methods. *Nat. Genet.* 48, 1284–1287. doi: 10.1038/ng.3656

Davey Smith, G., and Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* 23, R89–R98. doi: 10.1093/hmg/ddu328

de Vlaming, R., Okbay, A., Rietveld, C. A., Johannesson, M., Magnusson, P. K., Uitterlinden, A. G., et al. (2017). Meta-GWAS accuracy and power (MetaGAP) calculator shows that hiding heritability is partially due to imperfect genetic correlations across studies. *PLoS Genet.* 13:e1006495. doi: 10.1371/journal.pgen.1006495

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* 9:e1003348. doi: 10.1371/annotation/b91ba224-10be-409d-93f4-7423d502cba0

Dudbridge, F. (2016). Polygenic epidemiology. *Genet. Epidemiol.* 40, 268–272. doi: 10.1002/gepi.21966

Eskola, P. J., Lemmela, S., Kjaer, P., Solovieva, S., Mannikko, M., Tommerup, N., et al. (2012). Genetic association studies in lumbar disc degeneration: a systematic review. *PLoS ONE* 7:e49995. doi: 10.1371/journal.pone.0049995

Eskola, P. J., Männikkö, M., Samartzis, D., and Karppinen, J. (2014). Genome-wide association studies of lumbar disc degeneration–are we there yet? *Spine J.* 14, 479–482. doi: 10.1016/j.spinee.2013.07.437

Estrada, K., Styrkarsdottir, U., Evangelou, E., Hsu, Y. H., Duncan, E. L., Ntzani, E. E., et al. (2012). Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nat. Genet.* 44, 491–501. doi: 10.1038/ng.2249

Euesden, J., Lewis, C. M., and O'Reilly, P. F. (2015). PRSice: polygenic risk score software. *Bioinformatics* 31, 1466–1468. doi: 10.1093/bioinformatics/btu848

Evans, D. M., Brion, M. J., Paternoster, L., Kemp, J. P., McMahon, G., Munafò, M., et al. (2013). Mining the human phenome using allelic scores that index biological intermediates. *PLoS Genet.* 9:e1003919. doi: 10.1371/journal.pgen.1003919

Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*. Pearson Education.

Foss, M. V., and Byers, P. D. (1972). Bone density, osteoarthrosis of the hip, and fracture of the upper end of the femur. *Ann. Rheum. Dis.* 31, 259–264. doi: 10.1136/ard.31.4.259

Guilak, F. (2011). Biomechanical factors in osteoarthritis. *Best Pract. Res. Clin. Rheumatol.* 25, 815–823. doi: 10.1016/j.berh.2011.11.013

Hackinger, S., Trajanoska, K., Styrkarsdottir, U., Zengini, E., Steinberg, J., Ritchie, G. R. S., et al. (2017). Evaluation of shared genetic aetiology between osteoarthritis and bone mineral density identifies SMAD3 as a novel osteoarthritis risk locus. *Hum Mol Genet.* 26, 3850–3858. doi: 10.1093/hmg/ddx285

Hangai, M., Kaneoka, K., Kuno, S., Hinotsu, S., Sakane, M., Mamizuka, N., et al. (2008). Factors associated with lumbar intervertebral disc degeneration in the elderly. *Spine J.* 8, 732–740. doi: 10.1016/j.spinee.2007.07.392

Hannan, M. T., Anderson, J. J., Zhang, Y., Levy, D., and Felson, D. T. (1993). Bone mineral density and knee osteoarthritis in elderly men and women. The Framingham study. *Arthritis Rheum* 36, 1671–1680. doi: 10.1002/art.1780361205

Harada, A., Okuizumi, H., Miyagi, N., and Genda, E. (1998). Correlation between bone mineral density and intervertebral disc degeneration. *Spine* 23, 857–861. discussion: 862. doi: 10.1097/00007632-199804150-00003

Hassett, G., Hart, D. J., Manek, N. J., Doyle, D. V., and Spector, T. D. (2003). Risk factors for progression of lumbar spine disc degeneration: the Chingford Study. *Arthritis Rheum.* 48, 3112–3117. doi: 10.1002/art.11321

He, M., Xu, M., Zhang, B., Liang, J., Chen, P., Lee, J. Y., et al. (2015). Meta-analysis of genome-wide association studies of adult height in East Asians identifies 17 novel loci. *Hum. Mol. Genet.* 24, 1791–1800. doi: 10.1093/hmg/ddu583

Heikkilä, J. K., Koskenvuo, M., Heliövaara, M., Kurppa, K., Riihimäki, H., Heikkila, K., et al. (1989). Genetic and environmental factors in sciatica. Evidence from a nationwide panel of 9365 adult twin pairs. *Ann. Med.* 21, 393–398. doi: 10.3109/07853898909149227

Hershkovich, O., Friedlander, A., Gordon, B., Arzi, H., Derazne, E., Tzur, D., et al. (2013). Associations of body mass index and body height with low back pain in 829,791 adolescents. *Am. J. Epidemiol.* 178, 603–609. doi: 10.1093/aje/kwt019

Heuch, I., Heuch, I., Hagen, K., and Zwart, J. A. (2015). Association between body height and chronic low back pain: a follow-up in the Nord-Trondelag Health Study. *BMJ Open* 5:e006983. doi: 10.1136/bmjopen-2014-006983

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 5:e1000529. doi: 10.1371/journal.pgen.1000529

Humzah, M. D., and Soames, R. W. (1988). Human intervertebral disc: structure and function. *Anat. Rec.* 220, 337–356. doi: 10.1002/ar.1092200402

Ikegawa, S. (2013). The genetics of common degenerative skeletal disorders: osteoarthritis and degenerative disc disease. *Annu. Rev. Genomics Hum. Genet.* 14, 245–256. doi: 10.1146/annurev-genom-091212-153427

Johnson, L., Zhu, J., Scott, E. R., and Wineinger, N. E. (2015). An examination of the relationship between lipid levels and associated genetic markers across racial/ethnic populations in the multi-ethnic study of atherosclerosis. *PLoS ONE* 10:e0126361. doi: 10.1371/journal.pone.0126361

Johnson, T. (2012). "Efficient calculation for multi-SNP genetic risk scores," in *American Society of Human Genetics Annual Meeting* (San Francisco, CA).

Jordan, J., Konstantinou, K., and O'Dowd, J. (2009). Herniated lumbar disc. *BMJ Clin. Evid.* 2009:1118.

Katz, J. D., Agrawal, S., and Velasquez, M. (2010). Getting to the heart of the matter: osteoarthritis takes its place as part of the metabolic syndrome. *Curr. Opin. Rheumatol.* 22, 512–519. doi: 10.1097/BOR.0b013e32833bfb4b

Kong, A., Frigge, M. L., Thorleifsson, G., Stefansson, H., Young, A. I., Zink, F., et al. (2017). Selection against variants in the genome associated with educational attainment. *Proc. Natl. Acad. Sci. U.S.A.* 114, E727–E732. doi: 10.1073/pnas.1612113114

Krapohl, E., Euesden, J., Zabaneh, D., Pingault, J. B., Rimfeld, K., von Stumm, S., et al. (2016). Phenome-wide analysis of genome-wide polygenic scores. *Mol. Psychiatry* 21, 1188–1193. doi: 10.1038/mp.2015.126

Kung, A. W., Xiao, S. M., Cherny, S., Li, G. H., Gao, Y., Tso, G., et al. (2010). Association of JAG1 with bone mineral density and osteoporotic fractures: a genome-wide association study and follow-up replication studies. *Am. J. Hum. Genet.* 86, 229–239. doi: 10.1016/j.ajhg.2009.12.014

Lee, S. H., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2012a). A better coefficient of determination for genetic profile analysis. *Genet. Epidemiol.* 36, 214–224. doi: 10.1002/gepi.21614

Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M., and Wray, N. R. (2012b). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* 28, 2540–2542. doi: 10.1093/bioinformatics/bts474

Leino-Arjas, P., Kauppila, L., Kaila-Kangas, L., Shiri, R., Heistaro, S., and Heliövaara, M. (2008). Serum lipids in relation to sciatica among Finns. *Atherosclerosis* 197, 43–49. doi: 10.1016/j.atherosclerosis.2007.07.035

Li, Y., Samartzis, D., Campbell, D. D., Cherny, S. S., Cheung, K. M., Luk, K. D., et al. (2016). Two subtypes of intervertebral disc degeneration distinguished by large-scale population-based study. *Spine J.* 16, 1079–1089. doi: 10.1016/j.spinee.2016.04.020

Liuke, M., Solovieva, S., Lamminen, A., Luoma, K., Leino-Arjas, P., Luukkonen, R., et al. (2005). Disc degeneration of the lumbar spine in relation to overweight. *Int. J. Obes.* 29, 903–908. doi: 10.1038/sj.ijo.0802974

Livshits, G., Ermakov, S., Popham, M., Macgregor, A. J., Sambrook, P. N., Spector, T. D., et al. (2010). Evidence that bone mineral density plays a role in degenerative disc disease: the UK Twin Spine study. *Ann. Rheum. Dis.* 69, 2102–2106. doi: 10.1136/ard.2010.131441

Livshits, G., Popham, M., Malkin, I., Sambrook, P. N., Macgregor, A. J., Spector, T., et al. (2011). Lumbar disc degeneration and genetic factors are the main risk factors for low back pain in women: the UK Twin Spine Study. *Ann. Rheum. Dis.* 70, 1740–1745. doi: 10.1136/ard.2010.137836

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature* 518, 197–206. doi: 10.1038/nature14177

Longo, U. G., Denaro, L., Spiezia, F., Forriol, F., Maffulli, N., and Denaro, V. (2011). Symptomatic disc herniation and serum lipid levels. *Eur. Spine J.* 20, 1658–1662. doi: 10.1007/s00586-011-1737-2

Loughlin, J. (2011). Knee osteoarthritis, lumbar-disc degeneration and developmental dysplasia of the hip–an emerging genetic overlap. *Arthritis Res. Ther.* 13:108. doi: 10.1186/ar3291

Luoma, K., Riihimäki, H., Luukkonen, R., Raininko, R., Viikari-Juntura, E., and Lamminen, A. (2000). Low back pain in relation to lumbar disc degeneration. *Spine* 25, 487–492. doi: 10.1097/00007632-200002150-00016

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* 39, 906–913. doi: 10.1038/ng2088

Marigorta, U. M., and Navarro, A. (2013). High trans-ethnic replicability of GWAS results implies common causal variants. *PLoS Genet.* 9:e1003566. doi: 10.1371/journal.pgen.1003566

Natarajan, R. N., and Andersson, G. B. (1999). The influence of lumbar disc height and cross-sectional area on the mechanical response of the disc to physiologic loading. *Spine* 24, 1873–1881. doi: 10.1097/00007632-199909150-00003

Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., et al. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* 533, 539–542. doi: 10.1038/nature17671

Oxland, T. R. (2016). Fundamental biomechanics of the spine–what we have learned in the past 25 years and future directions. *J. Biomech.* 49, 817–832. doi: 10.1016/j.jbiomech.2015.10.035

Palla, L., and Dudbridge, F. (2015). A fast method that uses polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait. *Am. J. Hum. Genet.* 97, 250–259. doi: 10.1016/j.ajhg.2015.06.005

Panoutsopoulou, K., Metrustry, S., Doherty, S. A., Laslett, L. L., Maciewicz, R. A., Hart, D. J., et al. (2014). The effect of FTO variation on increased osteoarthritis risk is mediated through body mass index: a Mendelian randomisation study. *Ann. Rheum. Dis.* 73, 2082–2086. doi: 10.1136/annrheumdis-2013-203772

Panoutsopoulou, K., Southam, L., Elliott, K. S., Wrayner, N., Zhai, G., Beazley, C., et al. (2011). Insights into the genetic architecture of osteoarthritis from stage 1 of the arcOGEN study. *Ann. Rheum. Dis.* 70, 864–867. doi: 10.1136/ard.2010.141473

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460, 748–752. doi: 10.1038/nature08185

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Pye, S. R., Reid, D. M., Adams, J. E., Silman, A. J., and O'Neill, T. W. (2006). Radiographic features of lumbar disc degeneration and bone mineral density in men and women. *Ann. Rheum. Dis.* 65, 234–238. doi: 10.1136/ard.2005.038224

Radin, E. L., and Rose, R. M. (1986). Role of subchondral bone in the initiation and progression of cartilage damage. *Clin. Orthop. Relat. Res.* 213, 34–40. doi: 10.1097/00003086-198612000-00005

Reynard, L. N., and Loughlin, J. (2013). Insights from human genetic studies into the pathways involved in osteoarthritis. *Nat. Rev. Rheumatol.* 9, 573–583. doi: 10.1038/nrrheum.2013.121

Ropper, A. H., and Zafonte, R. D. (2015). Sciatica. *N. Engl. J. Med.* 372, 1240–1248. doi: 10.1056/NEJMra1410151

Samartzis, D., Karppinen, J., Chan, D., Luk, K. D., and Cheung, K. M. (2012). The association of lumbar intervertebral disc degeneration on magnetic resonance imaging with body mass index in overweight and obese adults: a population-based study. *Arthritis Rheum.* 64, 1488–1496. doi: 10.1002/art.33462

Samartzis, D., Karppinen, J., Cheung, J. P., and Lotz, J. (2013). Disk degeneration and low back pain: are they fat-related conditions? *Global Spine J.* 3, 133–144. doi: 10.1055/s-0033-1350054

Samartzis, D., Karppinen, J., Mok, F., Fong, D. Y., Luk, K. D., and Cheung, K. M. (2011). A population-based study of juvenile disc degeneration and its association with overweight and obesity, low back pain, and diminished functional status. *J. Bone Joint Surg. Am.* 93, 662–670. doi: 10.2106/JBJS.I.01568

Sambrook, P. N., MacGregor, A. J., and Spector, T. D. (1999). Genetic influences on cervical and lumbar disc degeneration: a magnetic resonance imaging study in twins. *Arthritis Rheum.* 42, 366–372.

Schneiderman, G., Flannigan, B., Kingston, S., Thomas, J., Dillin, W. H., and Watkins, R. G. (1987). Magnetic resonance imaging in the diagnosis of disc degeneration: correlation with discography. *Spine* 12, 276–281. doi: 10.1097/00007632-198704000-00016

Soininen, P., Kangas, A. J., Würtz, P., Suna, T., and Ala-Korpela, M. (2015). Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circ. Cardiovasc. Genet.* 8, 192–206. doi: 10.1161/CIRCGENETICS.114.000216

Soininen, P., Kangas, A. J., Würtz, P., Tukiainen, T., Tynkkynen, T., Laatikainen, R., et al. (2009). High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. *Analyst* 134, 1781–1785. doi: 10.1039/b910205a

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., and Smoller, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* 14, 483–495. doi: 10.1038/nrg3461

Song, Y. Q., Cheung, K. M., Ho, D. W., Poon, S. C., Chiba, K., Kawaguchi, Y., et al. (2008). Association of the asporin D14 allele with lumbar-disc degeneration in Asians. *Am. J. Hum. Genet.* 82, 744–747. doi: 10.1016/j.ajhg.2007.12.017

Song, Y. Q., Karasugi, T., Cheung, K. M., Chiba, K., Ho, D. W., Miyake, A., et al. (2013). Lumbar disc degeneration is linked to a carbohydrate sulfotransferase 3 variant. *J. Clin. Invest.* 123, 4909–4917. doi: 10.1172/JCI69277

Stahl, E. A., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B. F., et al. (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.* 44, 483–489. doi: 10.1038/ng.2232

Takatalo, J., Karppinen, J., Niinimäki, J., Taimela, S., Nayha, S., Mutanen, P., et al. (2011). Does lumbar disc degeneration on magnetic resonance imaging associate with low back symptom severity in young Finnish adults? *Spine* 36, 2180–2189. doi: 10.1097/BRS.0b013e3182077122

Takatalo, J., Karppinen, J., Taimela, S., Niinimäki, J., Laitinen, J., Blanco Sequeiros, R., et al. (2013). Body mass index is associated with lumbar disc degeneration in young Finnish males: subsample of Northern Finland birth cohort study 1986. *BMC Musculoskelet. Disord.* 14:87. doi: 10.1186/1471-2474-14-87

Urban, J. P., and Roberts, S. (2003). Degeneration of the intervertebral disc. *Arthritis Res. Ther.* 5, 120–130. doi: 10.1186/ar629

Vattikuti, S., Guo, J., and Chow, C. C. (2012). Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLoS Genet.* 8:e1002637. doi: 10.1371/journal.pgen.1002637

Videman, T., Levälahti, E., and Battié, M. C. (2007). The effects of anthropometrics, lifting strength, and physical activities in disc degeneration. *Spine* 32, 1406–1413. doi: 10.1097/BRS.0b013e31806011fa

Wahlström, J., Burström, L., Nilsson, T., and Järvholm, B. (2012). Risk factors for hospitalization due to lumbar disc disease. *Spine* 37, 1334–1339. doi: 10.1097/BRS.0b013e31824b5464

Wang, Y., Boyd, S. K., Battié, M. C., Yasui, Y., and Videman, T. (2011). Is greater lumbar vertebral BMD associated with more disk degeneration? A

study using microCT and discography. *J. Bone Miner Res.* 26, 2785–2791. doi: 10.1002/jbmr.476

Wen, W., Zheng, W., Okada, Y., Takeuchi, F., Tabara, Y., Hwang, J. Y., et al. (2014). Meta-analysis of genome-wide association studies in East Asian-ancestry populations identifies four new loci for body mass index. *Hum. Mol. Genet.* 23, 5492–5504. doi: 10.1093/hmg/ddu248

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., et al. (2013). Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* 45, 1274–1283. doi: 10.1038/ng.2797

Williams, F. M., Bansal, A. T., van Meurs, J. B., Bell, J. T., Meulenbelt, I., Suri, P., et al. (2013). Novel genetic variants associated with lumbar disc degeneration in northern Europeans: a meta-analysis of 4600 subjects. *Ann. Rheum. Dis.* 72, 1141–1148. doi: 10.1136/annrheumdis-2012-201551

Williams, F. M., Popham, M., Hart, D. J., de Schepper, E., Bierma-Zeinstra, S., Hofman, A., et al. (2011). GDF5 single-nucleotide polymorphism rs143383 is associated with lumbar disc degeneration in Northern European women. *Arthritis Rheum.* 63, 708–712. doi: 10.1002/art.30169

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* 46, 1173–1186. doi: 10.1038/ng.3097

Xiao, S. M., Kung, A. W., Gao, Y., Lau, K. S., Ma, A., Zhang, Z. L., et al. (2012). Post-genome wide association studies and functional analyses identify association of MPP7 gene variants with site-specific bone mineral density. *Hum. Mol. Genet.* 21, 1648–1657. doi: 10.1093/hmg/ddr586

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42, 565–569. doi: 10.1038/ng.608

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011a). GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88, 76–82. doi: 10.1016/j.ajhg.2010.11.011

Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., et al. (2011b). Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* 43, 519–525. doi: 10.1038/ng.823

Yang, J., Zeng, J., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2017). Concepts, estimation and interpretation of SNP-based heritability. *Nat. Genet.* 49, 1304–1310. doi: 10.1038/ng.3941

Zeggini, E., Panoutsopoulou, K., Southam, L., Rayner, N. W., Day-Williams, A. G., Lopes, M. C., et al. (2012). Identification of new susceptibility loci for osteoarthritis (arcOGEN): a genome-wide association study. *Lancet* 380, 815–823. doi: 10.1016/S0140-6736(12)60681-3

Zhang, Y., Zhao, Y., Wang, M., Si, M., Li, J., Hou, Y., et al. (2016). Serum lipid levels are positively correlated with lumbar disc herniation–a retrospective study of 790 Chinese patients. *Lipids Health Dis.* 15:80. doi: 10.1186/s12944-016-0248-x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 April 2019
doi: 10.3389/fgene.2019.00261

Accuracy of Gene Expression Prediction From Genotype Data With PrediXcan Varies Across and Within Continental Populations

Anna V. Mikhaylova and Timothy A. Thornton**

Department of Biostatistics, University of Washington, Seattle, WA, United States

Using genetic data to predict gene expression has garnered significant attention in recent years. PrediXcan has become one of the most widely used gene-based methods for testing associations between predicted gene expression values and a phenotype, which has facilitated novel insights into the relationship between complex traits and the component of gene expression that can be attributed to genetic variation. The gene expression prediction models for PrediXcan were developed using supervised machine learning methods and training data from the Depression Genes and Networks (DGN) study and the Genotype-Tissue Expression (GTEx) project, where the majority of subjects are of European descent. Many genetic studies, however, include samples from multi-ethnic populations, and in this paper we evaluate the accuracy of PrediXcan for predicting gene expression in diverse populations. Using transcriptomic data from the GEUVADIS (Genetic European Variation in Disease) RNA sequencing project and whole genome sequencing data from the 1000 Genomes project, we evaluate and compare the predictive performance of PrediXcan in an African population (Yoruban) and four European ancestry populations for thousands of genes. We evaluate a range of models from the PrediXcan weight databases and use Pearson's correlation coefficient to assess gene expression prediction accuracy with PrediXcan. From our evaluation, we find that the predictive performance of PrediXcan varies substantially among populations from different continents (F-test p-value $< 2.2 \times 10^{-16}$), where prediction accuracy is lower in the Yoruban population from West Africa compared to the European-ancestry populations. Moreover, not only do we find differences in predictive performance between populations from different continents, we also find highly significant differences in prediction accuracy among the four European ancestry populations considered (F-test p-value $< 2.2 \times 10^{-16}$). Finally, while there is variability in prediction accuracy across different PrediXcan weight databases, we also find consistency in the qualitative performance of PrediXcan for the five populations considered, with the African ancestry population having the lowest accuracy across databases.

Keywords: transcriptome, expression quantitative trait loci (eQTL), genetic diversity, genetic mapping, complex traits

OPEN ACCESS

Edited by:
Dana C. Crawford,
Case Western Reserve University,
United States

Reviewed by:
Georgios Athanasiadis,
University of Copenhagen, Denmark
Binglan Li,
University of Pennsylvania,
United States

****Correspondence:***
Anna V. Mikhaylova
avmikh@uw.edu
Timothy A. Thornton
tathornt@uw.edu

Specialty section:
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics

Received: *30 November 2018*
Accepted: *08 March 2019*
Published: *03 April 2019*

Citation:
Mikhaylova AV and Thornton TA (2019) Accuracy of Gene Expression Prediction From Genotype Data With PrediXcan Varies Across and Within Continental Populations. Front. Genet. 10:261. doi: 10.3389/fgene.2019.00261

1. INTRODUCTION

In the past decade, genome-wide association studies (GWAS) have identified thousands of genetic variants significantly associated with a wide range of human phenotypes (Sudlow et al., 2015; NHLBI, 2016; MacArthur et al., 2017; Visscher et al., 2017). The vast majority of these studies, however, were conducted in samples from European ancestry populations (Need and Goldstein, 2009; Bustamante et al., 2011; Petrovski and Goldstein, 2016; Popejoy and Fullerton, 2016; Bentley et al., 2017; Hindorff et al., 2018). Differences in allele frequencies, genetic architecture, and linkage disequilibrium (LD) patterns across ancestries suggest that GWAS discoveries can fail to generalize across populations, and recent publications have provided compelling evidence that GWAS findings often do not transfer from European populations to other ethnic groups (Adeyemo and Rotimi, 2009; Li and Keating, 2014). For example, Carlson et al. analyzed multi-ethnic data from the PAGE Consortium and concluded that some variants identified in GWAS in European ancestry populations had different magnitude and direction of allelic effects in non-European populations and the differential effects were more persistent in African Americans (Carlson et al., 2013). Moreover, genetic risk prediction models derived from European GWAS were found to be unreliable when applied to other ethnic groups (Carlson et al., 2013). Martin et al. examined the impact of population history on polygenic risk scores and demonstrated that they can be biased and confounded by population structure (Martin et al., 2017). Since genetic risk prediction accuracy depends on genetic similarity between the target and discovery cohorts, Martin et al. advised against interpreting the scores across populations and recommended computing them in genetically similar cohorts.

Associations between genetic variation and molecular traits, such as gene expression, have advanced our understanding of the mechanisms underlying trait-variant associations (Nica et al., 2010; Torres et al., 2014; Albert and Kruglyak, 2015). Prior studies have shown that a large proportion of GWAS variants identified for complex traits are expression quantitative trait loci (eQTLs), i.e., they play a role in regulating gene expression (Nicolae et al., 2010). Thus, eQTLs can aid in prioritizing likely causal variants among the ones identified by GWAS, especially if they are found in non-coding regions, and can help uncover the mechanisms by which genotypes influence phenotypes (Albert and Kruglyak, 2015). As a result, having three types of data—genotype, phenotype and gene expression—on the same set of subjects can be advantageous for improved understanding of the relationships between complex traits, the genetic backgrounds of study subjects, and the underlying biological processes. However, collecting all of these different types of data on the same study subjects is often not feasible due to cost and tissue availability. Additionally, eQTL studies have the same pitfalls as GWASs—the majority of the detected eQTLs are not causal, but may be in LD with causal variants. Similar to variants identified through GWAS, eQTL findings might fail to replicate in diverse populations due to differential LD patterns across populations (Kelly et al., 2017).

Recently, there has been increased interest in integrating eQTL studies and GWASs for improved complex trait mapping. PrediXcan (Gamazon et al., 2015) is one of the most widely used integrative methods for testing associations between a phenotype and gene expression values predicted from SNP genotyping or sequencing data. PrediXcan can have increased power over traditional GWAS methods, particularly when differential changes in gene expression is an intermediary stage of the causal pathway from genetic variation to the outcome of interest. A useful feature of PrediXcan (and other similar methods) is the ability to obtain predicted gene expression values on study subjects when tissue types relevant to phenotypes are not available. We now give a very brief overview of the PrediXcan method. PrediXcan uses machine learning methods and large reference datasets consisting of both genotype and trascriptome data for supervised training to construct prediction models for expression of each gene. With PrediXcan, genetic training data is restricted to common *cis*-variants that are within 1 Mb upstream and downstream from the transcription region (Gamazon et al., 2015). Gene-specific derived SNP weights from the prediction models are then stored in databases, with separate sets of weights for different tissue types. Using these weights, PrediXcan allows for the prediction of gene expression values for study subjects with available genotype data, where predicted expression values are computed as a weighted linear combination of SNP dosages. Finally, the predicted expression values can then be used to test for associations with a phenotype of interest. By conducting tests on gene expression obtained from an aggregation of variants, PrediXcan dramatically reduces multiple testing burden as compared to single variant association testing.

Previous studies have reported differences in gene expression levels across diverse populations from the HapMap3 project, noting that 77% of eQTLs are population specific and only 23% are shared between two or more populations (The International HapMap 3 Consortium et al., 2010; Stranger et al., 2012). More distantly related populations have more differentially expressed genes than closely related populations, although this can often be explained by the expression of different gene transcripts across populations (Lappalainen et al., 2013). One potential limitation of PrediXcan, however, is that the method may not perform well in diverse populations, as the supervised learning for PrediXcan was conducted using data from the Depression Genes and Networks (DGN) and the Genotype-Tissue Expression (GTEx) Project—both of which consist primarily of European-ancestry subjects (Lonsdale et al., 2013; Battle et al., 2014). Many genetic studies include samples from multi-ethnic populations, and understanding the accuracy of gene expression prediction with PrediXcan across populations is of interest to many genetic researchers.

Recent works have evaluated the performance of PrediXcan in diverse populations (Gottlieb et al., 2017; Li et al., 2018). Li et al. evaluated PrediXcan whole-blood prediction models and investigated the factors that influence prediction accuracy using the Yoruban (YRI) and European (CEU) samples from the Genetic European Variation in Health and Disease (GEUVADIS) (Lappalainen et al., 2013) cohort. In this paper, the PrediXcan performance was reported to be unsatisfactory for most genes

due to predicted gene expression values not correlating well with the observed values (Li et al., 2018). Differences in prediction accuracy with PrediXcan between the YRI and CEU, however, were not directly compared. Gottlieb et al. investigated the performance of PrediXcan for a small subset of 116 genes that are in the warfarin-response pathway in European and African American samples where they concluded that PrediXcan performed poorly in African Americans (Gottlieb et al., 2017).

Here, we evaluate the predictive performance of PrediXcan both across and within continental populations using thousands of genes across the genome. Using the GEUVADIS transcriptome data and whole genome sequencing data from the 1000 Genomes Project (Lappalainen et al., 2013; Auton et al., 2015), we consider four closely related European ancestry populations and one African population. In our analysis, we test the null hypotheses of (1) no difference in prediction accuracy with PrediXcan across European and African continental populations; and (2) no difference in predictive performance among the four European derived populations. We obtain predicted gene expression levels using seven PrediXcan weight databases derived from whole blood and lymphoblastoid cell lines (LCL) transcriptome data for each individual. To evaluate differences in prediction accuracy among the populations, we use a linear mixed effects model framework where Pearson's correlation coefficients for observed and predicted gene expression levels are included as the outcome and the populations are included as categorical predictors. In addition, we evaluate the utility of whole-blood-based models when making predictions for LCL expression data. We find from our analyses that accuracy of PrediXcan for gene expression prediction not only differs between European and African continental populations, but also among closely related populations of European ancestry. Furthermore, prediction accuracy with PrediXcan is the lowest in Africans across all seven weight databases considered, which further illustrates the need to develop new predictive models using training data composed of individuals who have similar ancestry to the target sample for which gene expression is to be predicted (Mogil et al., 2018).

2. MATERIALS AND METHODS

2.1. Datasets

We obtained gene expression data from the GEUVADIS Consortium and whole genome sequencing data from the 1000 Genomes Project. The gene expression data consisted of RNA sequencing on lymphoblastoid cell line (LCL) samples for 464 individuals from five populations. Of these, 445 subjects were in the 1000 Genomes Phase 3 dataset, including 358 subjects of European descent, and 87 subjects of African descent. European samples included: Utah residents with Northern and Western European ancestry (CEU, $n = 89$), British individuals in England and Scotland (GBR, $n = 86$), Finnish in Finland (FIN, $n = 92$), and Toscani in Italy (TSI, $n = 91$). African samples included individuals of African descent from Yoruba in Ibadan, Nigeria (YRI, $n = 87$). Gene expression measurements were available for 23,722 genes.

We used seven PrediXcan weight databases: DGN whole-blood (further referred to as DGN), GTEx v6 1KG whole blood, GTEx v6 1KG LCL, GTEx v6 HapMap whole blood, GTEx v6 HapMap LCL, GTEx v7 HapMap whole blood (GTEx WB), and GTEx v7 HapMap LCL (GTEx LCL). The databases were downloaded from http://predictdb.org/.

2.2. Filtering Procedure for Poorly Predicted Genes

Linear regression models were used to identify genes whose predicted values were not associated with the observed values at significance level of 0.05 in order to filter out genes that have poor prediction accuracy across all subjects. For each gene, we fit a linear regression model with observed gene expression as the outcome and predicted gene expression as the predictor of interest. A Wald test was used to assess significance of the coefficient for each gene in the linear model. Genes with corresponding p-values that were higher than a nominal significance level of 0.05 were identified and labeled as "poorly predicted."

We then calculated Pearson's correlation coefficient, r, between observed and predicted expression values for every gene, in each population separately. A few genes had the same predicted gene expression levels across all subjects. Since we could not calculate the correlation if one of the variables was constant, we excluded those genes. Thus, for every gene considered there were five Pearson's correlation coefficients, one for each population. Note that we used r instead of the square of Pearson correlation, r^2, in order to take directionality of correlation into account when assessing predictive performance. We found that using r^2 as a measure of predictive accuracy can be misleading as there were genes for which predicted and observed expression values had a significant negative correlation.

It should be noted that we also performed an evaluation of the performance of PrediXcan without doing any filtering of genes in order to assess the impact on the analysis when poorly predicted genes are excluded, as discussed below.

2.3. Assessing Prediction Accuracy Differences Across Populations and Across Tissues

In the analyses described below to assess differences in prediction accuracy with PrediXcan across populations, two sets of genes were considered—all genes without any filtering and a subset of genes using the filtering process previously described.

We first compared prediction performance between the two continental groups—European and African. For each gene, we calculated two Pearson's correlation coefficients between observed and predicted gene expression levels—one based on all European samples and the other one based on the African samples. We then used a paired t-test to assess differences in mean prediction accuracy between the correlation coefficients for European samples vs correlation coefficients for African samples.

To assess differences in prediction accuracy across the five populations, we used a linear mixed effects model approach where we fit the following model:

$$r_{ij} = \beta_0 + \gamma_i + \beta_1 \mathbb{I}_{FIN,i} + \beta_2 \mathbb{I}_{GBR,i} + \beta_3 \mathbb{I}_{TSI,i} + \beta_4 \mathbb{I}_{YRI,i} + \epsilon_{ij}, \quad (1)$$

where r_{ij} is the correlation coefficient for gene i in population j; and $\mathbb{I}_{FIN,i}$, $\mathbb{I}_{GBR,i}$, $\mathbb{I}_{TSI,i}$, and $\mathbb{I}_{YRI,i}$ are indicator variables that are equal to 1 if the gene correlation was calculated on the population indicated in the subscript, and otherwise are equal to 0. Thus, we modeled population as a categorical predictor, with the CEU population as a reference. To account for variation between genes, we included a random intercept γ_i for each gene and we assumed that $\gamma_i \sim \mathcal{N}(0, \sigma_\gamma^2)$. We also included an error term ϵ_{ij}, such that $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. We used repeated measures ANOVA to test the null hypothesis of $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ for no difference in mean Pearson's correlation coefficients among the populations. A Wald test was used to assess significance of differences in mean Pearson's correlation coefficients between CEU, the reference population, and each of the other four populations.

We also ran a similar analysis where we excluded the CEU population due to potentially lower quality of the CEU cell lines, as reported in the literature (Çaliskan et al., 2014; Yuan et al., 2015). We fit a model identical to (1), excluding the CEU and using the FIN population as a reference:

$$r_{ij} = \beta_0 + \gamma_i + \beta_1 \mathbb{I}_{GBR,i} + \beta_2 \mathbb{I}_{TSI,i} + \beta_3 \mathbb{I}_{YRI,i} + \epsilon_{ij}, \quad (2)$$

where the notation is the same as above.

Additionally, we tested for differences in prediction accuracy across the four European populations. For this analysis, we included only individuals of European ancestry and fit the following linear mixed effects model:

$$r_{ij} = \beta_0 + \gamma_i + \beta_1 \mathbb{I}_{FIN,i} + \beta_2 \mathbb{I}_{GBR,i} + \beta_3 \mathbb{I}_{TSI,i} + \epsilon_{ij}, \quad (3)$$

where CEU is included as the reference population in the model. As in the previously described analyses, a repeated measures ANOVA was used to test for differences in prediction accuracy across the four European populations.

To evaluate how the PrediXcan performance with whole-blood (WB) databases differed from LCL databases, we restricted the set of genes to only those that were present in both the WB and LCL databases. First, we presented scatter plots of correlation coefficients comparing WB and LCL databases in the five populations separately. Then we recalculated Pearson's correlation coefficients between observed and predicted expression values with all the five populations combined but separately for every database, i.e., as a result, we had two correlation coefficients per gene, one that corresponded to a GTEx WB database and one to a GTEx LCL database. We compared each pair of GTEx WB and GTEx LCL databases using a paired t-test between LCL-based correlation coefficients and WB-based correlation coefficients. All the statistical analyses described above were performed in `R` version 3.3.3 (R Core Team, 2014). All plots were generated with `ggplot2` (Wickham, 2016).

3. RESULTS

3.1. Overview of PrediXcan Weight Databases

In **Table 1**, we summarize the main features of the PrediXcan weight databases that we used in the analyses. Compared to DGN database, GTEx databases have fewer gene models and smaller training sample sizes. HapMap and 1KG-based models differ in the number of variants used for training: GTEx Hapmap models were trained on the HapMap genotyping data while GTEx 1KG were trained on the 1000 Genomes sequencing data, so the latter utilize more variants when predicting expression. While GTEx LCL databases are based on relatively small training sets, they are derived from the same tissue as the GEUVADIS RNA-seq data we analyzed. Lastly, DGN and GTEx v7 sets of weights were trained only on Europeans samples, while GTEx v6 databases had a small fraction of non-Europeans.

To avoid repetition, results using the DGN, GTEx v7 WB, and GTEx v7 LCL databases are included in the main text, while the results for the other four databases are provided in the **Supplementary Material**.

TABLE 1 | Summary of PrediXcan databases used in analyses.

PrediXcan database	Training set size	Number of models	Number of SNPs used
DGN whole blood	922	13,171	249,696
GTEx v6 1KG whole blood	338	6,759	185,786
GTEx v6 1KG LCL	114	3,759	125,045
GTEx v6 HapMap whole blood	338	6,588	136,941
GTEx v6 HapMap LCL	114	3,441	91,237
GTEx v7 HapMap whole blood	315	6,297	140,931
GTEx v7 HapMap LCL	96	3,045	88,143

TABLE 2 | Number of genes for which Pearson correlation coefficients are available by population and by PrediXcan weight database.

PrediXcan database	DGN	GTEx v7 WB	GTEx v7 LCL
Genes with observed and predicted expression values	10,387	5,432	2,777
By population:			
CEU	10,385	5,432	2,777
FIN	10,385	5,432	2,777
GBR	10,385	5,432	2,777
TSI	10,385	5,432	2,776
YRI	10,354	5,419	2,767
Genes before filtering	10,354	5,419	2,767
Genes after filtering	3,493	2,288	1,699

3.2. PrediXcan Prediction Accuracy Differs Across Diverse Populations

Using DGN, GTEx WB, and GTEx LCL models and sequence data, gene expression was predicted for 10,387, 5,432, and 2,777 genes, respectively (see **Table 2**). The number of genes with available predictions varied by population, where the four European populations had a similar number of gene predictions while the counts for YRI were slightly lower. We excluded 33 genes, 13 genes, and 10 genes from DGN, GTEx WB, and GTEx LCL, respectively, due to there being no variation in predicted gene expression values for at least one of the populations. For

the remaining genes, we identified those that had poor prediction accuracy based on associations between observed and predicted values, as described in section Materials and Methods on filtering poorly predicted genes. From the genes predicted with the DGN database, two-thirds were labeled by this criterion as "poorly predicted," while slightly less than a half were labeled as such from gene sets predicted using the GTEx databases. As previously mentioned, we also considered the performance of PrediXcan without doing any filtering of the genes. For every weight database, we had two sets of genes—before and after filtering—where the latter set is a much smaller subset of the former. Both versions were used and evaluated in downstream analyses.

We first evaluated performance of PrediXcan for the two continental populations, European and African. We compared Pearson's correlation of predicted and observed gene expression values for the combined sample consisting of all individuals from the four European-ancestry populations to Pearson's correlation calculated for the YRI African population sample. As only two groups were being compared in this analysis, a paired *t*-test was used to assess differences in prediction accuracy, where the pairing was based on the gene. With or without the filtering of genes, we find the mean difference in gene correlation coefficients between the European and African samples to be highly significantly different from zero, regardless of the weight database used (all *p*-values $< 2.2 \times 10^{-16}$), with the African population having lower prediction accuracy than the European samples.

Next, we computed gene correlation coefficients, separately in each of the five populations. Violin plots display the correlation coefficients by population across genes before and after filtering (see **Figures 1A,B**, respectively). **Figure 1A** shows correlation coefficients for the genes before any filtering was done and we observe that LCL-derived models perform better than WB-derived: i.e., DGN and GTEx v7 WB correlation distributions are centered at values close to 0, whereas GTEx LCL correlation distributions are centered at higher values, especially for the four European populations. We also note that prediction accuracy is slightly lower for the African populations than for any of the European populations across the three weight databases. This trend is even more obvious after the filtering process. As we can see in **Figure 1B**, the overall performance accuracy improved after filtering in all the populations, as expected. However, the difference in prediction performance in Europeans vs. Africans is even more apparent. The four European populations have similar prediction accuracy, whereas it is lower for the African population. Similarly to panel A, LCL-derived prediction models perform better than WB-derived in filtered genes in **Figure 1B**.

Afterwards, we binned the genes into six categories based on the gene correlation coefficients (see **Table 3**). The majority of genes have very poor prediction accuracy—of the genes predicted with whole-blood databases, a third have negative correlations and a half have correlations between 0 and 0.2. Of the genes predicted with LCL, a fifth have negative correlations and over a third have correlations between 0 and 0.2. The distribution of gene correlation coefficients is fairly similar across the four European populations, although predictive accuracy seems worse in CEU compared to FIN, GBR, and TSI. The predictive accuracy is the lowest in the African sample. Across all populations, only a small number of genes were predicted with high accuracy (with $r > 0.6$). Furthermore, all European populations have a greater number of well-predicted genes than the African population, regardless of the weight database used.

Next, we assessed the association between the prediction accuracy (as gene correlation coefficients) and population category via repeated measures ANOVA and linear mixed models using both sets of genes, all and filtered. The results for unfiltered and filtered genes were comparable and led to equivalent conclusions. Based on the repeated measures ANOVA, we find that prediction accuracy differs across populations for filtered and unfiltered sets of genes, regardless of the weight database used (*p*-values for all databases were $< 2.2 \times 10^{-16}$). Below, we focus our attention on filtered genes and present the parameter estimates and their 95% confidence intervals calculated using model-based standard errors for the model 1 in **Table 4**. From the linear mixed model 1, we find that the prediction accuracy is significantly higher in FIN, GBR, and TSI and significantly lower in YRI, compared to CEU. This suggests that predictive performance varies not only among distant populations, but also among closely related populations. When we performed the analysis on a full set of genes, without any filtering, regression coefficients were slightly attenuated toward zero; however, the conclusions from hypothesis testing remained the same.

We repeated the analysis described above, this time excluding the CEU population. We present the parameter estimates and the corresponding 95% confidence intervals in **Table 5**. From the repeated measures ANOVA, we find that prediction accuracy differs across the four populations (*p*-values for all databases were $< 2.2 \times 10^{-16}$). Moreover, based on the coefficients and the corresponding *p*-values from the linear mixed model 2, we estimate the prediction accuracy to be significantly higher in GBR and significantly lower in TSI and YRI, compared to the FIN population (see corresponding *p*-values in **Table 5**). This difference in prediction accuracy is the greatest between YRI and FIN when GTEx v7 LCL weight database was used. Like in the analysis above, we notice that predictive performance differs across populations, including European populations.

Finally, we evaluated PrediXcan prediction accuracy on a subset of subjects with European ancestry. Based on the repeated measures ANOVA test, prediction performance differs across the four European populations in genes before and after filtering, regardless of the weight database used (*p*-values for all databases were $< 2.2 \times 10^{-16}$). Because of potentially biased expression patterns of the CEU due to the previously mentioned age of these cell lines, we conducted an analysis where we omitted the CEU population and compared prediction accuracy among the other three European populations. The results were comparable to the analysis of the European populations that included CEU. With a repeated measures ANOVA, we find highly significant differences in prediction accuracy among the FIN, GBR, and TSI populations, with *p*-values less than 10^{-6} across all weight databases with or without filtering of poorly predicted genes.

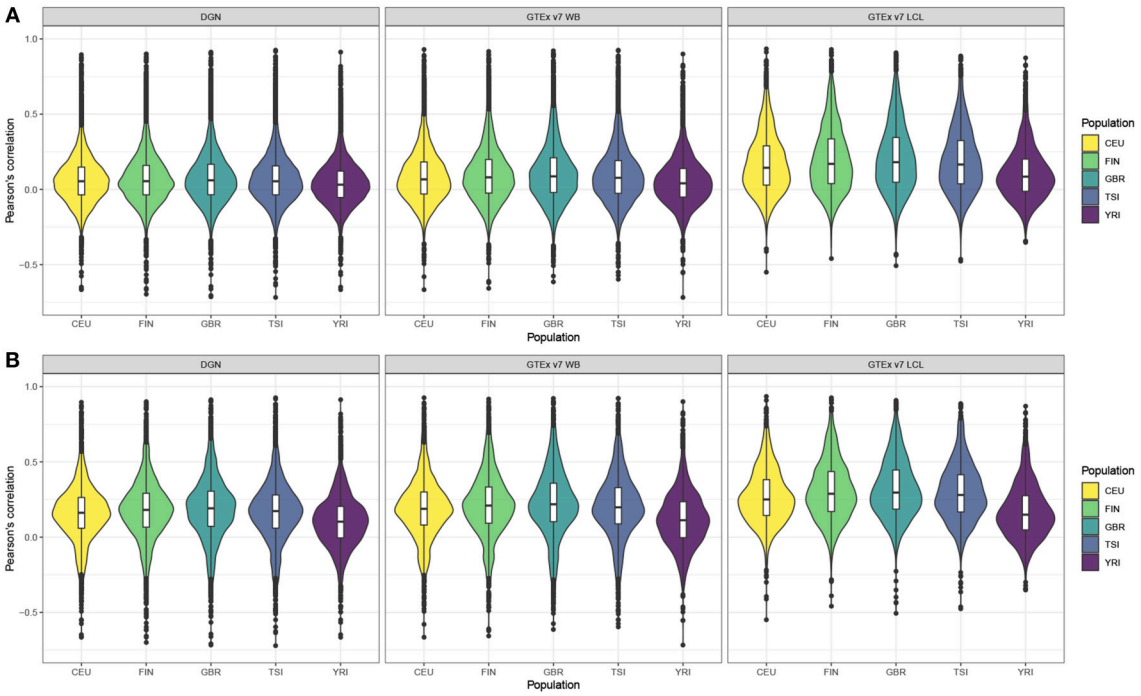

FIGURE 1 | Violin plots of gene expression correlation coefficients by five populations using DGN, GTEx v7 WB, and GTEx v7 LCL weight databases; **(A)** before and **(B)** after filtering out poorly predicted genes.

TABLE 3 | Gene counts per population, per database, per correlation category for the five populations using DGN, GTEx WB, and GTEx LCL weight databases.

	Unfiltered					Filtered				
	CEU	**FIN**	**GBR**	**TSI**	**YRI**	**CEU**	**FIN**	**GBR**	**TSI**	**YRI**
DGN DATABASE										
$r < 0$	3,583	3,491	3,480	3,587	4,156	561	547	554	585	911
$0 < r < 0.2$	5,107	4,976	4,812	4,954	5,001	1,533	1,379	1,258	1,409	1,674
$0.2 < r < 0.4$	1,359	1,480	1,589	1,434	1,016	1,097	1,162	1,209	1,121	728
$0.4 < r < 0.6$	239	302	354	290	147	236	300	353	289	146
$0.6 < r < 0.8$	56	93	105	75	31	56	93	105	75	31
$0.8 < r < 1$	10	12	14	14	3	10	12	14	14	3
GTEx v7 WB DATABASE										
$r < 0$	1,756	1,621	1,622	1,684	2,101	336	309	314	335	590
$0 < r < 0.2$	2,471	2,450	2,366	2,456	2,491	877	786	732	820	993
$0.2 < r < 0.4$	902	958	981	901	668	788	804	793	758	546
$0.4 < r < 0.6$	210	282	329	278	117	207	281	328	275	117
$0.6 < r < 0.8$	69	93	100	85	38	69	93	100	85	38
$0.8 < r < 1$	11	15	21	15	4	11	15	21	15	4
GTEx v7 LCL DATABASE										
$r < 0$	546	488	484	509	774	80	69	55	69	274
$0 < r < 0.2$	1,119	1,031	996	1,050	1,296	560	443	426	477	777
$0.2 < r < 0.4$	718	742	761	736	510	675	681	692	681	461
$0.4 < r < 0.6$	293	361	369	360	145	293	361	369	360	145
$0.6 < r < 0.8$	80	126	137	96	38	80	126	137	96	38
$0.8 < r < 1$	11	19	20	16	4	11	19	20	16	4

TABLE 4 | Results from linear mixed models for population category (with CEU as a reference) and change in gene correlation coefficient among filtered genes.

	DGN			GTEx v7 WB			GTEx v7 LCL		
	Estimate	**95% CI**	***p*-value**	**Estimate**	**95% CI**	***p*-value**	**Estimate**	**95% CI**	***p*-value**
FIN	0.019	(0.014, 0.025)	1.3×10^{-11}	0.021	(0.015, 0.028)	1.3×10^{-9}	0.038	(0.030, 0.046)	$< 10^{-16}$
GBR	0.029	(0.023, 0.034)	$< 10^{-16}$	0.032	(0.025, 0.039)	$< 10^{-16}$	0.051	(0.043, 0.059)	$< 10^{-16}$
TSI	0.010	(0.004, 0.016)	3.9×10^{-4}	0.013	(0.007, 0.020)	4.6×10^{-5}	0.027	(0.019, 0.035)	2.9×10^{-11}
YRI	−0.054	(−0.059, −0.048)	$< 10^{-16}$	−0.070	(−0.077, −0.063)	$< 10^{-16}$	−0.097	(−0.105 −0.089)	$< 10^{-16}$

TABLE 5 | Results from linear mixed models for population category (excluding CEU, with FIN as a reference) and change in gene correlation coefficient among filtered genes.

	DGN			GTEx v7 WB			GTEx v7 LCL		
	Estimate	**95% CI**	***p*-value**	**Estimate**	**95% CI**	***p*-value**	**Estimate**	**95% CI**	***p*-value**
GBR	0.010	(0.004, 0.015)	9.2×10^{-4}	0.011	(0.004, 0.018)	3.1×10^{-3}	0.013	(0.005, 0.021)	$2.0 \times \times 10^{-3}$
TSI	−0.009	(−0.015, −0.003)	1.8×10^{-3}	−0.008	(−0.015, −0.001)	2.8×10^{-2}	−0.011	(−0.019, −0.003)	8.9×10^{-3}
YRI	−0.073	(−0.079, −0.067)	$< 10^{-16}$	−0.091	(−0.098, −0.084)	$< 10^{-16}$	−0.134	(−0.143, −0.126)	$< 10^{-16}$

3.3. PrediXcan Prediction Accuracy Differs Between Tissues

As can be seen in the violin plots in **Figure 1**, both databases based on whole blood perform similarly, and LCL-based database displays improved prediction accuracy. In order to compare pairwise gene correlations, we restricted our analyses to the 1,595 genes common for both GTEx v7 WB and GTEx v7 LCL.

Scatter plots presented in **Figure 2** suggest that the majority of genes have very similar correlation coefficients when using WB and LCL databases across all populations. However, we see more genes in the upper left corner, above the dotted line, indicating that using the LCL database results in more genes with better prediction accuracy. This result is not surprising since the expression data we used were derived from LCL. The results of the paired *t*-test are consistent with the visual examination of the data: the mean difference between gene correlations based on the GTEx v7 LCL models and based on the GTEx v7 WB models is 0.03 (p-value $< 2.2 \times 10^{-16}$), with predictions based on the LCL model having higher performance.

4. DISCUSSION

In this work, we evaluated the performance of PrediXcan and compared the prediction accuracy of the method across five geographically diverse populations from two continents for seven weight databases. Models from all weight databases considered were trained on subjects primarily of European ancestry; three of the databases were derived from LCL and the remaining four from whole blood. As a measure of prediction accuracy, we computed correlation coefficients for each gene in all populations and used both paired *t*-tests and linear mixed effects models to assess evidence of significant differences in prediction performance across populations. We also investigated whether whole blood models are appropriate for predicting gene expression levels in LCL.

We find highly significant differences in prediction accuracy with PrediXcan in the European ancestry populations as compared to the YRI African population, with the prediction accuracy being lower in YRI. The lower accuracy with PrediXcan in the African population is expected since the PrediXcan models were largely trained using European ancestry samples, and this result is consistent with recent works showing that prediction accuracy is expected to be higher when the training and testing cohorts are of similar ancestry (Gottlieb et al., 2017; Li et al., 2018; Mogil et al., 2018). Surprisingly, we also find highly significant differences in prediction accuracy with PrediXcan among the closely related European ancestry populations, with the Finnish, British, and Italian populations having significantly higher prediction accuracy than the CEU. These results are consistent across all seven PrediXcan weight databases we considered. Lastly, we also find that LCL-trained models outperformed whole-blood-trained models across populations, although the prediction accuracy was similar for many of the genes.

Among the European populations, we find that prediction accuracy for the CEU population was the lowest. LCLs are derived from B cells found in whole blood, and they provide a continuous supply of genetic material for GWAS and gene expression studies. However, they do undergo a transformation to become immortal that can change their biology and they do not have the same properties as native tissue (Kelly et al., 2017). Storage conditions, freeze-thaw cycles, and maturity of cell lines can also affect gene expression patterns (Çaliskan et al., 2014; Yuan et al., 2015). The CEU cell lines were collected much earlier than the other cell lines and LCL age can have a confounding effect and bias downstream analyses (Yuan et al., 2015). This factor could have contributed to the differences in prediction accuracy among European populations. We did, however, perform a sensitivity analysis that excluded the CEU population, and there were highly significant differences in prediction accuracy

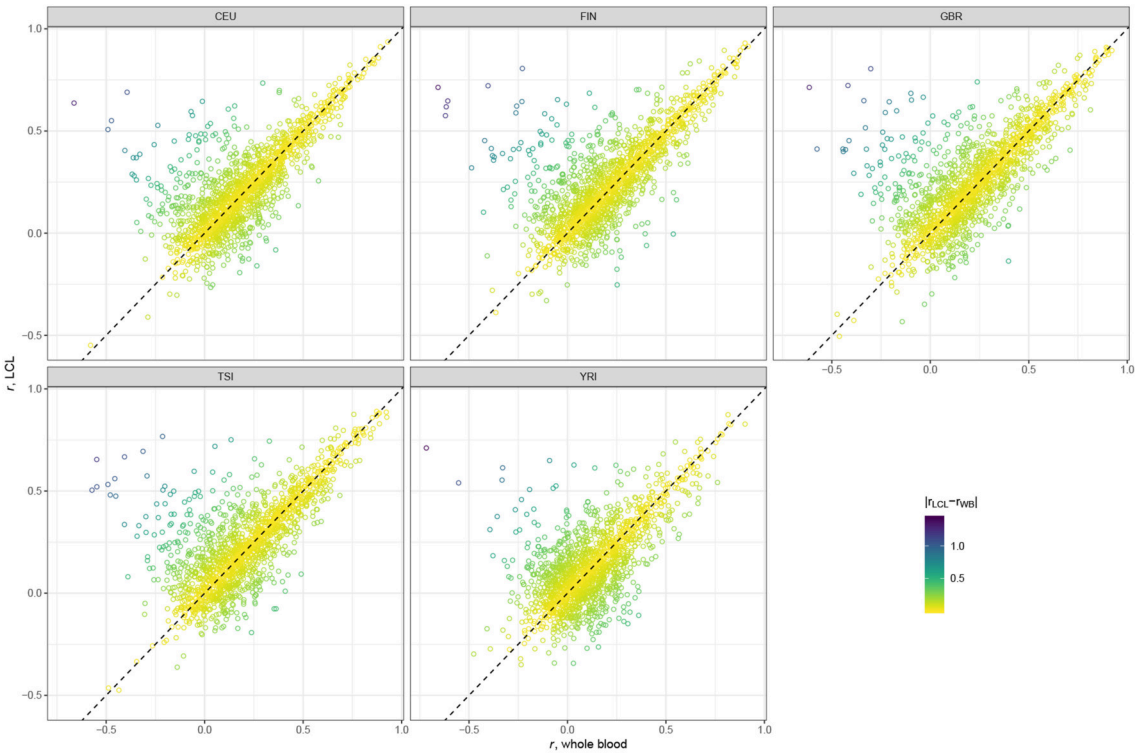

FIGURE 2 | Scatter plots comparing gene correlation coefficients by population using GTEx v7 LCL vs. GTEx v7 WB databases.

with PrediXcan among the FIN, GBR, and TSI populations, as well as between these three combined European populations and the YRI African population, with the YRI having the lowest accuracy.

Overall, PrediXcan accurately predicted gene expression for some genes; however, the majority of genes had very poor correlation between measured and predicted expression levels. For almost half the genes, for example, the correlation was negative. There are some important caveats and limitations to point out with the PrediXcan method. First, the prediction models of PrediXcan are based on common *cis*-variants and they do not take rare *cis*- and *trans*-regulatory elements into account. Common *cis*-eQTLs only account for 9–12% of genetic variance in gene expression, according to a large twin study (Grundberg et al., 2012). Another recent study demonstrates that *trans*-acting variants largely contribute to gene expression variation, with estimates of genetic variance in expression due to *trans*-acting variation ranging from 60 to 90% (Liu et al., 2018). However, individual effects of each *trans*-variant are very weak and difficult to map because they require well-powered studies.

We conclude this paper by highlighting that the lack of genomic data from diverse populations limits the ability to effectively interpret and translate genomic results into clinical applications for individuals from diverse populations, and particularly non-European ancestry populations. The results presented in this paper illustrate that gene expression prediction models are, in general, not transferable across diverse populations from different continents, and further corroborate the importance of including more ancestrally diverse individuals in medical genomics to ensure that everyone gets the benefits of precision medicine and to avoid further exacerbating healthcare inequality (Oh et al., 2015, 2016; Manrai et al., 2016). We also demonstrate that there can be differences in prediction accuracy among closely related European populations, suggesting that prediction models that take into account fine-scale ancestry differences among individuals may be important for improved prediction of gene expression from genetic data. Lastly, our study had only modest sample sizes and evaluated gene expression prediction accuracy with PrediXcan in European and African populations. Future transcriptomic studies with much larger samples sizes are needed for the development of improved gene expression prediction models for multi-ethnic populations, including admixed populations such as African Americans and Hispanic/Latino populations, who have recent ancestry derived from multiple continents.

DATA AVAILABILITY

GEUVADIS expression data is available at Array Express (E-MTAB-264 and E-GEUV-1) at https://www.ebi.ac.uk/arrayexpress/experiments/ and 1000 Genomes project genotype data is available at http://www.internationalgenome.org/.

AUTHOR CONTRIBUTIONS

AM and TT conceived the idea, designed the analysis, interpreted the results, and wrote the paper. AM ran the analysis.

FUNDING

This work was supported by National Institute of Health grant AG054074. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

We thank two reviewers for helpful comments and suggestions that improved the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00261/full#supplementary-material

REFERENCES

Adeyemo, A., and Rotimi, C. (2009). Genetic variants associated with complex human diseases show wide variation across multiple populations. *Public Health Genom.* 13, 72–79. doi: 10.1159/000218711

Albert, F. W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* 16, 197–212. doi: 10.1038/nrg3891

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., et al. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* 24, 14–24. doi: 10.1101/gr.155192.113

Bentley, A. R., Callier, S., and Rotimi, C. N. (2017). Diversity and inclusion in genomic research: why the uneven progress? *J. Commun. Genet.* 8, 255–266. doi: 10.1007/s12687-017-0316-6

Bustamante, C. D., Burchard, E. G., and De la Vega, F. M. (2011). Genomics for the world. *Nature* 475, 163–165. doi: 10.1038/475163a

Çaliskan, M., Pritchard, J. K., Ober, C., and Gilad, Y. (2014). The effect of freeze-thaw cycles on gene expression levels in lymphoblastoid cell lines. *PLoS ONE* 9:e107166. doi: 10.1371/journal.pone.0107166

Carlson, C. S., Matise, T. C., North, K. E., Haiman, C. A., Fesinmeyer, M. D., Buyske, S., et al. (2013). Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. *PLoS Biology* 11:e1001661. doi: 10.1371/journal.pbio.1001661

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* 47, 1091–1098. doi: 10.1038/ng.3367

Gottlieb, A., Daneshjou, R., DeGorter, M., Bourgeois, S., Svensson, P. J., Wadelius, M., et al. (2017). Cohort-specific imputation of gene expression improves prediction of warfarin dose for African Americans. *Gen. Med.* 9, 1–9. doi: 10.1186/s13073-017-0495-0

Grundberg, E., Small, K. S., Hedman, Å. K., Nica, A. C., Buil, A., Keildson, S., et al. (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet.* 44, 1084–1089. doi: 10.1038/ng.2394

Hindorff, L. A., Bonham, V. L., Brody, L. C., Ginoza, M. E., Hutter, C. M., Manolio, T. A., et al. (2018). Prioritizing diversity in human genomics research. *Nat. Rev. Genet.* 19, 175–185. doi: 10.1038/nrg.2017.89

Kelly, D. E., Hansen, M. E., and Tishkoff, S. A. (2017). Global variation in gene expression and the value of diverse sampling. *Curr. Opin. Syst. Biol.* 1, 102–108. doi: 10.1016/j.coisb.2016.12.018

Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A., Monlong, J., Rivas, M. A., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–511. doi: 10.1038/nature12531

Li, B., Verma, S. S., Veturi, Y. C., Verma, A., Bradford, Y., Haas, D. W., et al. (2018). Evaluation of PrediXcan for prioritizing GWAS associations and predicting gene expression. *Pac. Symp. Biocomput.* 23, 448–459. doi: 10.1142/9789813235533_0041

Li, Y. R., and Keating, B. J. (2014). Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Gen. Med.* 6:91. doi: 10.1186/s13073-014-0091-5

Liu, X., Li, Y. I., and Pritchard, J. K. (2018). Trans effects on gene expression can drive omnigenic inheritance. *bioRxiv [preprint].* bioRxiv:425108. doi: 10.1101/425108

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., and Lo, E. (2013). The genotype-tissue Expression (GTEx) project. *Nature* 45, 580–585. doi: 10.1038/ng.2653

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* 45, D896–D901. doi: 10.1093/nar/gkw1133

Manrai, A. K., Funke, B. H., Rehm, H. L., Olesen, M. S., Maron, B. A., Szolovits, P., et al. (2016). Genetic misdiagnoses and the potential for health disparities. *New Engl. J. Med.* 375, 655–665. doi: 10.1056/NEJMsa1507092

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., et al. (2017). Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* 100, 635–649. doi: 10.1016/j.ajhg.2017.03.004

Mogil, L. S., Andaleon, A., Badalamenti, A., Dickinson, S. P., Guo, X., Rotter, J. I., et al. (2018). Genetic architecture of gene expression traits across diverse populations. *PLoS Genet.* 14:e1007586. doi: 10.1371/journal.pgen.1007586

Need, A. C., and Goldstein, D. B. (2009). Next generation disparities in human genomics: concerns and remedies. *Trends Genet.* 25, 489–494. doi: 10.1016/j.tig.2009.09.012

NHLBI (2016). *NHLBI Trans-Omics for Precision Medicine. TOPMed Projects and their Parent Studies*. Available online at: https://www.nhlbiwgs.org/group/project-studies

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., et al. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* 6:e1000895. doi: 10.1371/journal.pgen.1000895

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Eileen Dolan, M., and Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* 6:e1000888. doi: 10.1371/journal.pgen.1000888

Oh, S. S., Galanter, J., Thakur, N., Pino-Yanes, M., Barcelo, N. E., White, M. J., et al. (2015). Diversity in clinical and biomedical research: a promise yet to be fulfilled. *PLoS Med.* 12:e1001918. doi: 10.1371/journal.pmed.1001918

Oh, S. S., White, M. J., Gignoux, C. R., and Burchard, E. G. (2016). Making precision medicine socially precise: take a deep breath. *Am. J. Respir. Crit. Care Med.* 193, 348–350. doi: 10.1164/rccm.201510-2045ED

Petrovski, S., and Goldstein, D. B. (2016). Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biol.* 17:157. doi: 10.1186/s13059-016-1016-y

Popejoy, A. B., and Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature* 538, 161–164. doi: 10.1038/538161a

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Stranger, B. E., Montgomery, S. B., Dimas, A. S., Parts, L., Stegle, O., Ingle, C. E., et al. (2012). Patterns of Cis regulatory variation in diverse human populations. *PLoS Genet.* 8e1002639. doi: 10.1371/journal.pgen.1002639

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., et al. (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 12:e1001779. doi: 10.1371/journal.pmed.1001779

The International HapMap 3 Consortium, Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., et al. (2010). The Integrating common and rare genetic variation in diverse human populations. *Nature* 467, 52–58. doi: 10.1038/nature09298

Torres, J. M., Gamazon, E. R., Parra, E. J., Below, J. E., Valladares-Salgado, A., Wacher, N., et al. (2014). Cross-tissue and tissue-specific eQTLs: Partitioning the heritability of a complex trait. *Am. J. Hum. Genet.* 95, 521–534. doi: 10.1016/j.ajhg.2014.10.001

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., et al. (2017). 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101, 5–22. doi: 10.1016/j.ajhg.2017.06.005

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag New York.

Yuan, Y., Tian, L., Lu, D., and Xu, S. (2015). Analysis of genome-wide RNA-sequencing data suggests age of the CEPH/Utah (CEU) lymphoblastoid cell lines systematically biases gene expression profiles. *Sci. Rep.* 5, 1–5. doi: 10.1038/srep07960

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

SYSTEMATIC REVIEW
published: 17 April 2019
doi: 10.3389/fgene.2019.00339

Systematic Review and Meta-Analysis Confirms Significant Contribution of Surfactant Protein D in Chronic Obstructive Pulmonary Disease

*Debparna Nandy, Nidhi Sharma and Sabyasachi Senapati**

Department of Human Genetics and Molecular Medicine, Central University of Punjab, Bathinda, India

OPEN ACCESS

Edited by:
William Scott Bush, Case Western Reserve University, United States

Reviewed by:
Lifeng Tian, University of Pennsylvania, United States
Lili Ding, Cincinnati Children's Hospital Medical Center, United States

****Correspondence:***
Sabyasachi Senapati
sabyasachi1012@gmail.com;
s.senapati@cup.edu.in

Specialty section:
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics

Received: *13 July 2018*
Accepted: *29 March 2019*
Published: *17 April 2019*

Citation:
Nandy D, Sharma N and Senapati S (2019) Systematic Review and Meta-Analysis Confirms Significant Contribution of Surfactant Protein D in Chronic Obstructive Pulmonary Disease. Front. Genet. 10:339. doi: 10.3389/fgene.2019.00339

Background: Surfactant protein D (SFTPD) is a lung specific protein which performs several key regulatory processes to maintain overall lung function. Several infectious and immune mediated diseases have been shown to be associated with SFTPD. Recent findings have suggested the serum concentration of SFTPD can be used as a diagnostic or prognostic marker for chronic obstructive pulmonary disease (COPD) and acute exacerbation COPD (AECOPD). But these findings lack replication studies from different ethnic populations and meta-analysis, to establish SFTPD as reliable diagnostic or prognostic biomarker for COPD and associated conditions.

Methods: We performed systematic literature search based on stringent inclusion and exclusion criteria to identify eligible studies to perform a meta-analysis. Our objective was to assess the predictability of serum SFTPD concentration and SFTPD allelic conformation at rs721917 (C > T) with COPD and AECOPD outcome. These variables were compared between COPD and healthy controls, where mean difference (MD), and odds ratio (OR) were calculated to predict the overall effect size. Review manager (RevMan-v5.3) software was used to analyse the data.

Results: A total of eight published reports were included in this study. Comparative serum SFTPD concentration data were extracted from six studies and three studies were evaluated for assessment of genetic marker from SFTPD. Our study identified strong association of elevated serum SFTPD with COPD and AECOPD. Significant association of risk was also observed for "T" allele or "TT" genotype of rs721917 from SFTPD with COPD and AECOPD.

Conclusion: Serum concentration and alleleic conformation of SFTPD has a significantly high predictive value for COPD and AECOPD. Thus, these can be tested further and could be applied as a predictive or prognostic marker.

Keywords: SFTPD, COPD, AECOPD, rs721917, meta-analysis

INTRODUCTION

Chronic Obstructive Pulmonary Disease (COPD) affects lungs and exhibits irreversible airflow conditions that leads to improper respiratory function (Carolan et al., 2014). COPD is a global disease burden which accounts for ∼3 million deaths annually (Zemans et al., 2017) and is responsible for the increase in worldwide mortality and morbidity (Dickens et al., 2011). Chronic Obstructive Pulmonary Disease is projected to be the third leading cause of death by 2020 (Dickens et al., 2011). Chronic Obstructive Pulmonary Disease has multiple sub-phenotypic conditions like emphysema, lean body mass, mucus hypersecretion, and acute exacerbation (Dickens et al., 2011; Shakoori et al., 2012; Carolan et al., 2014). Each sub-phenotype is considered to be the outcome of different immune related pathways which are involved in COPD pathogenesis (Ishii et al., 2012).

Surfactant protein D (SF-D or SFTPD) is a highly lung specific glycoprotein secreted by type II alveolar cells and non-ciliated clara cells and functionally involved in maintaining the lung functions (Shakoori et al., 2012; Akiki et al., 2016). This multimeric glycoprotein belongs to lectin super family (Moreno et al., 2014) and takes part in immune regulation and maintenance of lung function (Ju et al., 2012). SFTPD is found to have three domains, namely: the collagen like domain, the neck domain, and the carbohydrate domain (Moreno et al., 2014). The carbohydrate binding domain is responsible for the maintenance of innate immune function in the lungs. Upon calcium binding, this calcium dependent protein cross-talks with defensin and other immunoregulatory molecules (Crouch and Wright, 2001; Jakel et al., 2013; Moreno et al., 2014). Due to considerably high molecular stability i.e., over 6 months in circulation, SFTPD has been investigated to establish it as a biomarker for pulmonary function (Holmskov et al., 2003; Hoegh et al., 2010).

Most COPD patients belong to the stable COPD category (SCOPD) followed by acute exacerbation COPD (AECOPD). AECOPD is characterized by sudden worsening of respiratory conditions including secretion of greenish phlegm (Shakoori et al., 2009). Trends of elevated serum SFTPD concentration among AECOPD patients compared to SCOPD or healthy control group have been reported. Elevated serum SFTPD among AECOPD patient group ($n = 13$; 227 ± 120 ng/ml), compared to SCOPD ($n = 14$; 151 ± 83 ng/ml), and control group ($n = 54$; 127 ± 65 ng/ml) was reported among Pakistanis (Shakoori et al., 2009). In another case-control study on a Chinese population, similar trend was observed, where serum SFTPD level was found to be significantly ($p < 0.001$) higher among AECOPD ($n = 40$; 235.22 ± 48.27 ng/ml) than SCOPD ($n = 71$; 153.54 ± 45.21 ng/ml) and control subjects ($n = 60$; 103.05 ± 24.97 ng/ml) (Ju et al., 2012). Serum SFTPD is often found to show association with different lung function parameters (Liu et al., 2014). Besides COPD, SFTPD is found to be associated with multiple pulmonary and other multifactorial diseases including lung cancer, interstitial pneumonia, asthma, viral infection, and other acute respiratory syndromes (Ishii et al., 2012; Carolan et al., 2014; Zemans et al., 2017). Serum SFTPD level can be pivotal in the diagnosis and monitoring of prognosis of various pulmonary conditions. Among COPD patients, serum concentration of SFTPD was found associated with BODE (body mass index, airflow obstruction, dyspnea, exercise capacity) index of severity (Ju et al., 2012) and mortality (Celli et al., 2012). However, its association with COPD severity was not observed in several other studies (Lomas et al., 2009; Liu et al., 2014; Akiki et al., 2016).

Genetic variations in *SFTPD* have also been established as informative genetic markers for COPD in different populations (Shakoori et al., 2012; Fakih et al., 2018). A non-synonymous variation rs721917:c.92T>C (p.Met31Thr) is associated with altered serum concentration of SFTPD and its multimerization (Sorensen et al., 2009). Presence of "T" or "C" alleles of rs721917 codes for methionine or threonine amino acids, respectively, at 31st position of SFTPD protein. Degradation of SFTPD from its multimerized (high molecular weight) to non-multimerized (low molecular weight) form is associated with respiratory diseases, including COPD (Fakih et al., 2018). This variation was also found associated with COPD among Mexicans and Europeans and with emphysema among Japanese populations (Guo et al., 2001; Foreman et al., 2011; Ishii et al., 2012; Horimasu et al., 2014). Other intronic or synonymous variations (rs2245121, rs911887, rs6413520, and rs7078012) were also identified as associated with altered serum concentration of SFTPD among Europeans (NETT-NAS and ECLIPSE cohorts) (Foreman et al., 2011).

In this study, we performed a systematic review and meta-analysis, with the objective to establish the potential of a single biomarker, SFTPD in identifying, and stratifying the different COPD sub-phenotype(s). We attempt to establish an association between variation in serum SFTPD concentration and the SFTPD genetic variation rs721917 with COPD and its sub- phenotype(s). The rationale of this study is to increase the power by considering multiple studies in a meta-analysis with similar environmental conditions, thereby evaluating the significance of SFTPD as an important diagnostic biomarker.

METHODOLOGY

Identification and Eligibility of Relevant Studies

A search for eligible literature was done till May 12th, 2018. Databases used for the retrieval of eligible articles were PubMed (along with MESH database) and Google Scholar. The following keywords were used to retrieve all the publications: "COPD and SFTPD"; "Chronic obstructive pulmonary disease and surfactant protein D"; "COPD and serum SFTPD"; "COPD and SFTPD genotypes." Publications with the desirable keywords were selected. Further publications were added from the cross-references of the retrieved articles. Details are given in the **Figure 1**.

Study Inclusion/Exclusion Criteria

Scanning of publications with relevant titles and abstracts were done only for case-control studies encompassing COPD, and

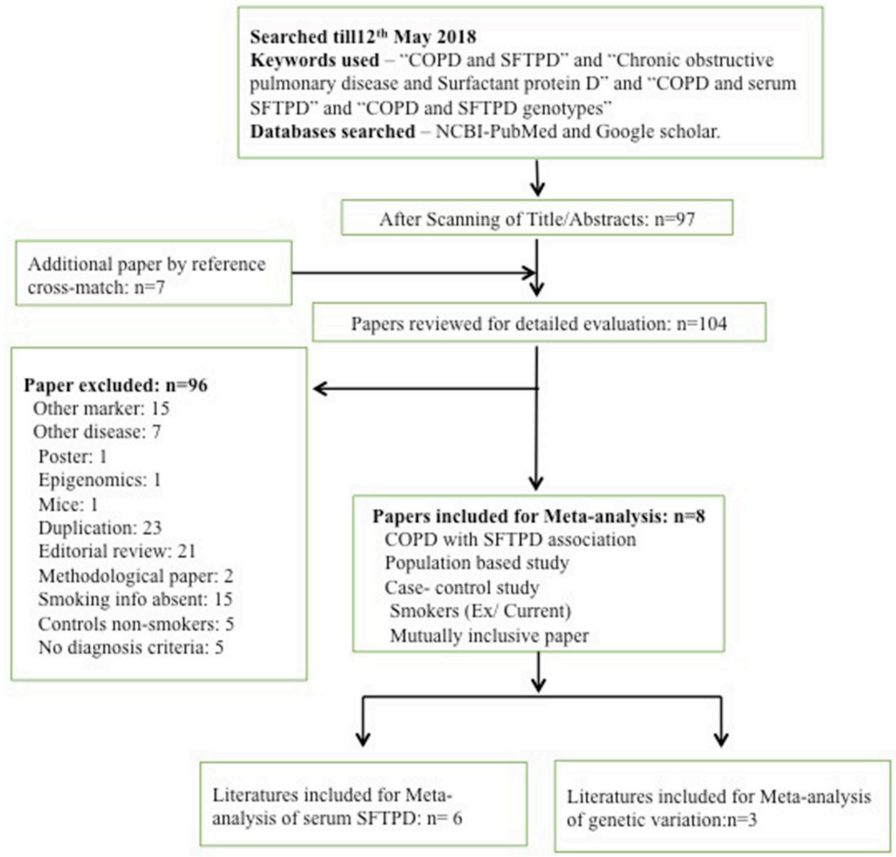

FIGURE 1 | Flow–diagram for the literature search procedure performed for meta-analysis.

TABLE 1 | Summarized results for association of serum SFTPD concentration with COPD and AECOPD.

Sr.no	Study no.	Stratification	COPD/Control	MD (95 % CI)	Effect size (Z)	P (Z test)	I^2 (%)	Pheterogeneity
1.	6	Overall COPD	2109/464	39.26 [36.97, 41.54]	33.65	<0.0001	11	0.34
2.	2	Acute exacerbation COPD (AECOPD)	53/114	130.41[114.62, 46.20]	16.19	<0.00001	0	0.36

AECOPD as one of the major sub-phenotypes. No other sub-phenotypes of COPD were included in the study owing to maintain the focus area of the present meta-analysis. Only those studies were included where study participants were aged more than 35 years and all were smokers. We did not keep sex as a selection criteria. While for genetic polymorphisms, studies having information about rs721917 were only selected.

Data Extraction

Data extraction from the eligible publications was done by two investigators independently and conflicts were resolved through group discussions. Following data was extracted from the finally selected publications:

a. Protein biomarker: author names, number of participants, SFTPD serum/plasma level mean value (cases and controls), diagnostic criteria.
b. Genetic biomarker (rs721917): author names, number of participants, allele distribution among cases and controls, population, and diagnostic criteria.

Statistical Analysis

For statistical analysis, Review Manager (RevMan-v5.3) Copenhagen: The Nordic Cochrane Center, The Cochrane Collaboration, 2014, software was used. As serum biomarker level is a continuous variable mean difference (MD) was calculated. For genetic marker odds ratio (OR) for pooled data was calculated. Different genetic models such as, allelic model, dominant model, recessive model, and additive model were used to analyze the association. Heterogeneity among studies was calculated using I^2 and chi^2 tests, where I^2 more than 50% and chi^2 p-value <0.05 was considered significant heterogeneity. Both the analyses were done using fixed effect model. Meta-OR or Meta-MD were calculated using Z-test with 5% level of

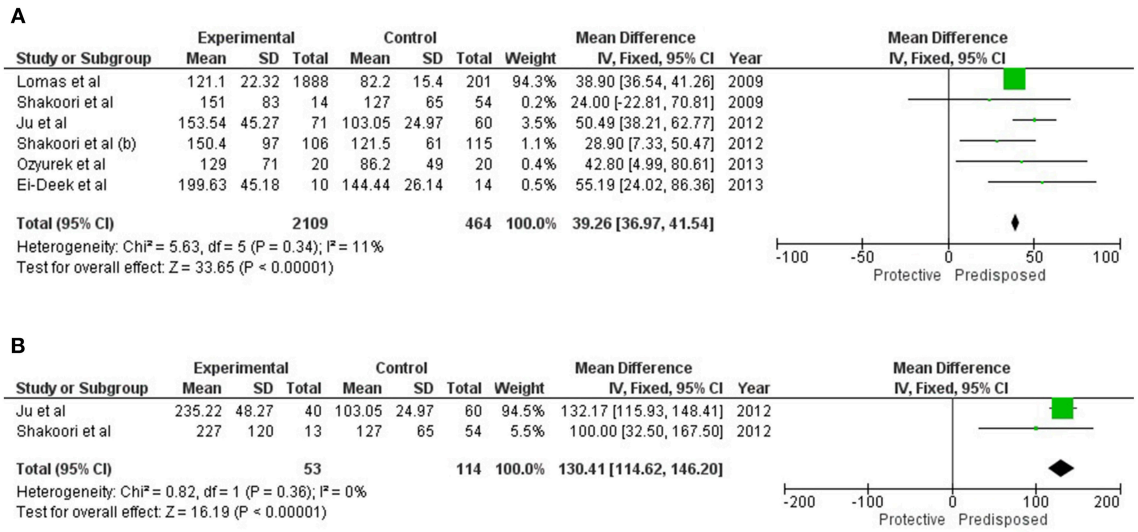

FIGURE 2 | Forest plots showing results of meta-analysis to assess the association of serum SFTPD concentration with **(A)** Overall COPD, and **(B)** AECOPD. The Mean Difference between the COPD subjects and controls were used for estimating the variation among the same. Fixed model was used for meta-analysis as there were negligible study heterogeneity (I^2 < 50%), in both the analyses.

significance and 95% confidence interval. Possible publication bias was evaluated through visual inspection of funnel plots generated using the same software.

RESULTS

Characteristics of Eligible Studies

Following online literature search, a total of 97 publications were obtained. Additionally seven publications were found through cross-references. However, based on our study inclusion-exclusion criteria, a total of 96 publications were excluded. Therefore, only eight publications were found eligible and taken forward for the meta-analysis (**Figure 1**). Eligible studies were reported between 2009 and 2017 (**Supplementary Tables 1**, **2**). Out of these 96 publications, six were assessed to evaluate the risk of SFTPD serum concentration and three were assessed to evaluated for risk of genetic variation in *SFTPD* (rs721917) with overall COPD and acute exacerbation with COPD (AECOPD). Detail characteristics of these studies are presented in the **Supplementary Tables 1**, **2**. No significant publication bias was observed among the studies included in this meta-analysis (**Supplementary Figure 1**).

Association of Serum SFTPD With COPD

Serum concentration of SFTPD (mean and SD) was available for a total of 2,109 cases and 464 healthy controls reported in eligible studies. Elevated serum SFTPD values were found to be significantly associated [M.D = 39.26 (36.97, 41.54; p_Z < 0.00001] with overall COPD (**Table 1** and **Figure 2A**). Two of these studies were further assessed for evaluating the contribution of serum SFTPD level with AECOPD. Meta-analysis was performed on 53 cases and 114 controls. Elevated level of serum SFTPD was found associated with AECOPD [MD = 130.41(114.62, 46.20); p_Z < 0.00001] (**Table 1** and **Figure 2B**).

Association of *SFTPD* Genotype With COPD

All of these three reports on the association of SFTPD genetic variations with overall COPD were carried out on Asian populations. Cases in Chinese and Lebanese populations were diagnosed according to both American Thoracic Society (ATS) and GOLD Criteria, while the Pakistani population was diagnosed solely on the basis of GOLD criteria (**Supplementary Table 2**). For rs721917; allelic and genotypic (dominant and recessive) associations are summarized in **Table 2**. Under the allelic model of association, "T" allele was identified to confer risk for both COPD [OR = 1.34 (1.07–1.67); p_Z = 0.01] and AECOPD [OR = 1.41 (1.09–1.83); p_Z = 0.009] (**Figure 3**). Similarly, dominant model identified association of "TT" genotype with both COPD [OR = 1.41 (1.00–1.99); p_Z = 0.05] and AECOPD [OR = 1.50 (1.01–2.23); p_Z = 0.04] with marginal significance. Recessive model confirmed the protective role of "CC" genotype for both COPD [OR = 0.60 (0.39–0.94); p_Z = 0.02] and AECOPD [OR = 0.55 (0.33–0.92); p_Z = 0.02]. Dominant role and risk confers by the "T" allele was further confirmed by additive models of association (**Table 2**).

DISCUSSION

Surfactant protein-D is a key innate immunity molecule with significant role in host defense. It has been reported as associated with several health conditions including COPD as well as various associated manifestations, such as AECOPD

TABLE 2 | Results of meta-analysis for alleles and genotypes of rs721917 under different genetic models.

Sr. no.	Genetic model	Stratification	Study numbers	Cases/Control	OR (95 % CI)	Z_p	I^2 (%)	Heterogeneity *P*-value
1.	Allelic model (T vs. C)	COPD	3	668/656	1.34 [1.07,1.67]	0.01	0	0.73
		AECOPD	2	544/426	1.41 [1.09, 1.83]	0.009	0	0.98
2.	Dominant model (TT vs. CT/CC)	COPD	3	334/328	1.41 [1.00, 1.99]	0.05	0	0.72
		AECOPD	2	272/213	1.50 [1.01, 2.23]	0.04	0	0.58
3.	Recessive model (CC vs. CT/TT)	COPD	3	334/328	0.60 [0.39, 0.94]	0.02	22	0.28
		AECOPD	2	272/213	0.55 [0.33, 0.92]	0.02	54	0.14
4a.	Additive model Homozygote comparison (TT vs. CC)	COPD	3	162/150	1.88 [1.15, 3.09]	0.01	0	0.49
		AECOPD	2	135/99	2.10 [1.18, 3.75]	0.01	0	0.34
4b.	Additive model Heterozygote comparison (TT vs. CT)	COPD	3	293/264	1.30 [0.90, 1.86]	0.16	0	0.62
		AECOPD	2	241/172	1.36 [0.90, 2.06]	0.14	0	0.14
4c.	Additive model Heterozygote comparison (CT vs. CC)	COPD	3	213/140	1.51 [0.96, 2.38]	0.08	37	0.20
		AECOPD	2	168/153	1.63 [0.95, 2.78]	0.08	66	0.08

All analyses were done using Fixed effect model.

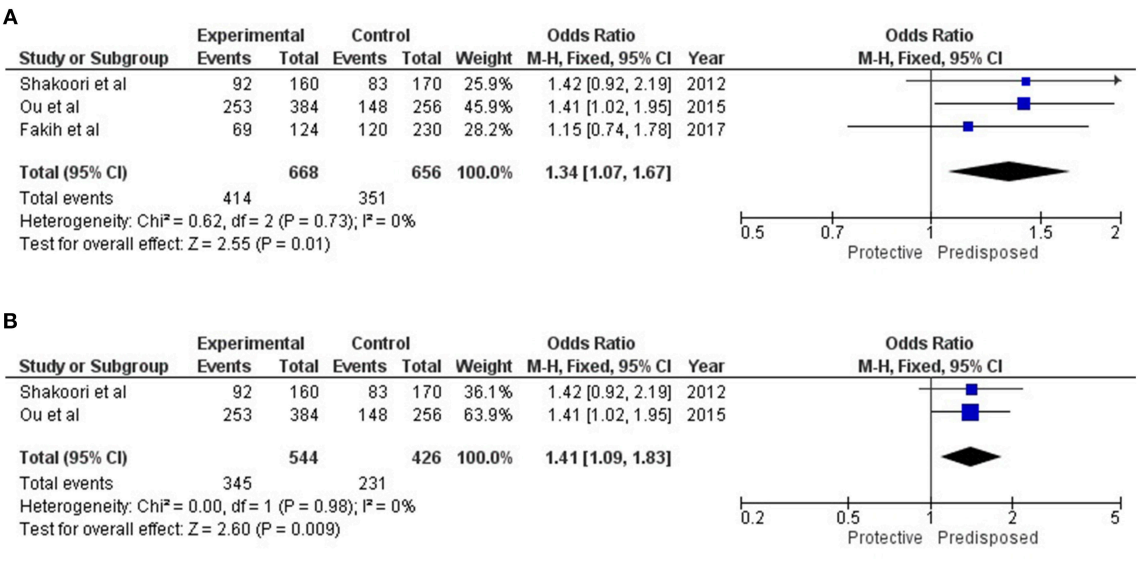

FIGURE 3 | Forest plots showing results of meta-analysis to assess the association of SFTPD alleles (rs721917-T) with **(A)** Overall COPD, and **(B)** AECOPD; where meta-Odds Ratio was calculated based on the previously reported data. Fixed model statistics was used for meta-analysis as no study heterogeneity ($I^2 = 0$) was observed.

(Hartl and Griese, 2006). Recent genome-wide association studies had identified SFTPD as one of the most significant and well-replicated gene associated with COPD and it's allied complications (Kim et al., 2012; Hobbs et al., 2017). Significant difference in SFTPD serum concentration among the COPD and healthy controls can be used as criteria for disease diagnosis and/or prognosis. So far, except for emphysema (alpha-1 antitrypsin) no other biomarker is available for the diagnosis or evaluation the COPD prognosis and associated lung function. COPD and AECOPD are reported to be associated with elevated serum concentration of SFTPD (Lomas et al., 2009; Shakoori et al., 2009, 2012; Ju et al., 2012; El-Deek et al., 2013; Ozyurek et al., 2013). In contrary emphysema patients are found to have lower level of serum SFTPD than healthy people (Ishii et al., 2012).

This study is the first attempt to review and meta-analyze the existing published literature to assess the predictive value of SFTPD serum concentration or genotypes for COPD and AECOPD. Due to stringent study inclusion and exclusion criteria limited articles were found eligible for this study. Present study identified strong associations of elevated serum SFTPD level with both COPD and AECOPD. Extracted data from all the eligible studies were homogenous and no study selection bias was observed. As expected, serum SFTPD was observed

more significantly associated with AECOPD ($p < 0.00001$; MD = 130.41) compared to COPD ($p < 0.0001$; MD = 39.26) when compared to healthy controls. However, as this study was performed on reported case-control based cross-sectional studies, causal effect relationship between the serum SFTPD level and COPD could not be affirmed. Recent evidences confirmed that elevated serum SFTPD can be used as a prognostic marker for COPD, as its serum concentration has been found significantly elevated among AECOPD compared to COPD (Shakoori et al., 2012; Ou et al., 2015).

Present study suggests the use of SFTPD as a biomarker to evaluate COPD. Since different range of SFTPD concentrations are found for different COPD and AECOPD conditions, single biomarker can be used for the diagnosis of COPD, and it's prognosis. Range of scale (SFTPD serum concentration) can be made to access the diagnosis and prognosis.

Limitations of this study include, less population numbers due to stringent inclusion and exclusion criteria. Population as a whole has been considered and not further stratified on their ethnicities. Studies with larger cohorts need to be conducted to confirm the association of serum SFTPD and its allelic conformation with COPD and AECOPD. Furthermore, to generalize these findings large-scale population based replication studies are warranted.

AUTHOR CONTRIBUTIONS

SS conceptualized the study. DN and NS performed the systematic review and meta-analysis. SS, DN, and NS wrote the manuscript. All authors reviewed and finalized the manuscript for submission.

FUNDING

Department of Science and Technology—Science and Engineering Research Board (ECR/2016/001660), New Delhi, India and University Grants Commission (F.30-4/2014(BSR), New Delhi, India.

ACKNOWLEDGMENTS

Dr. Kavita Singh, Public Health Foundation of India, Gurugram, India, for helping in data analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00339/full#supplementary-material

REFERENCES

Akiki, Z., Fakih, D., Jounblat, R., Chamat, S., Waked, M., Holmskov, U., et al. (2016). Surfactant protein D, a clinical biomarker for chronic obstructive pulmonary disease with excellent discriminant values. *Exp. Ther. Med.* 11, 723–730. doi: 10.3892/etm.2016.2986

Carolan, B. J., Hughes, G., Morrow, J., Hersh, C. P., O'Neal, W. K., Rennard, S., et al. (2014). The association of plasma biomarkers with computed tomography-assessed emphysema phenotypes. *Respir. Res.* 15:127. doi: 10.1186/s12931-014-0127-9

Celli, B. R., Locantore, N., Yates, J., Tal-Singer, R., Miller, B. E., Bakke, P., et al. (2012). Inflammatory biomarkers improve clinical prediction of mortality in chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* 185, 1065–1072. doi: 10.1164/rccm.201110-1792OC

Crouch, E., and Wright, J. R. (2001). Surfactant proteins A and D and pulmonary host defense. *Annu. Rev. Physiol.* 63, 521–554. doi: 10.1146/annurev.physiol.63.1.521

Dickens, J. A., Miller, B. E., Edwards, L. D., Silverman, E. K., Lomas, D. A., and Tal-Singer, R. (2011). COPD association and repeatability of blood biomarkers in the ECLIPSE cohort. *Respir. Res.* 12:146. doi: 10.1186/1465-9921-12-146

El-Deek, S. E., Makhlouf, H. A., Saleem, T. H., Mandour, M. A., and Mohamed, N. A. (2013). Surfactant protein D, soluble intercellular adhesion molecule-1 and high-sensitivity C-reactive protein as biomarkers of chronic obstructive pulmonary disease. *Med. Princ. Pract.* 22, 469–474. doi: 10.1159/000349934

Fakih, D., Akiki, Z., Junker, K., Medlej-Hashim, M., Waked, M., Salameh, P., et al. (2018). Surfactant protein D multimerization and gene polymorphism in COPD and asthma. *Respirology.* 23, 298–305. doi: 10.1111/resp.13193

Foreman, M. G., Kong, X., DeMeo, D. L., Pillai, S. G., Hersh, C. P., Bakke, P., et al. (2011). Polymorphisms in surfactant protein–D are associated with chronic obstructive pulmonary disease. *Am. J. Respir. Cell Mol. Biol.* 44, 316–322. doi: 10.1165/rcmb.2009-0360OC

Guo, X., Lin, M. H., Lin, Z., Montano, M., Sansores, R., Wang, G., et al. (2001). Surfactant protein gene A, B, D marker alleles in chronic obstructive pulmonary disease of a Mexican population. *Eur. Respir. J.* 18, 482–490. doi: 10.1183/09031936.01.00043401

Hartl, D., and Griese, M. (2006). Surfactant protein D in human lung diseases. *Eur. J. Clin. Invest.* 36, 423–435. doi: 10.1111/j.1365-2362.2006.01648.x

Hobbs, B. D., de Jong, K., Lamontagne, M., Bossé, Y., Shrine, N., Artigas, M. S., et al. (2017). Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nat. Genet.* 49, 426–432. doi: 10.1038/ng.3752

Hoegh, S. V., Sorensen, G. L., Tornoe, I., Lottenburger, T., Ytting, H., Nielsen, H. J., et al. (2010). Long-term stability and circadian variation in circulating levels of surfactant protein D. *Immunobiology* 215, 314–320. doi: 10.1016/j.imbio. 2009.05.001

Holmskov, U., Thiel, S., and Jensenius, J. C. (2003). Collectins and ficolins: humoral lectins of the innate immune defense. *Annu. Rev. Immunol.* 21, 547–578. doi: 10.1146/annurev.immunol.21.120601.140954

Horimasu, Y., Hattori, N., Ishikawa, N., Tanaka, S., Bonella, F., Ohshimo, S., et al. (2014). Differences in serum SP-D levels between German and Japanese subjects are associated with SFTPD gene polymorphisms. *BMC Med. Genet.* 15:4. doi: 10.1186/1471-2350-15-4

Ishii, T., Hagiwara, K., Kamio, K., Ikeda, S., Arai, T., Mieno, M. N., et al. (2012). Involvement of surfactant protein D in emphysema revealed by genetic association study. *EJHG.* 20, 230–235. doi: 10.1038/ejhg.2011.183

Jakel, A., Qaseem, A. S., Kishore, U., and Sim, R. B. (2013). Ligands and receptors of lung surfactant proteins SP-A and SP-D. *Front. Biosci.* 18, 1129–1140. doi: 10.2741/4168

Ju, C. R., Liu, W., and Chen, R. C. (2012). Serum surfactant protein D: biomarker of chronic obstructive pulmonary disease. *Dis. Markers.* 32, 281–287. doi: 10.3233/DMA-2011-0887.

Kim, D. K., Cho, M. H., Hersh, C. P., Lomas, D. A., Miller, B. E., Kong, X., et al. (2012). Genome-wide association analysis of blood biomarkers in chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Car.* 186, 1238–1247. doi: 10.1164/rccm.201206-1013OC

Liu, W., Ju, C. R., Chen, R. C., and Liu, Z. G. (2014). Role of serum and induced sputum surfactant protein D in predicting the response to treatment in chronic obstructive pulmonary disease. *Exp. Ther. Med.* 8, 1313–1317. doi: 10.3892/etm.2014.1865

Lomas, D. A., Silverman, E. K., Edwards, L. D., Locantore, N. W., Miller, B. E., Horstman, D. H., et al. (2009). Serum surfactant protein D is steroid

sensitive and associated with exacerbations of COPD. *Eur. Respir. J.* 34, 95–102. doi: 10.1183/09031936.00156508

Moreno, D., Garcia, A., Lema, D., and De Sanctis, B. J. (2014). Surfactant protein D in chornic obstructive pulmonary disease (COPD). *Recent Pat. Endocr. Metab.Immune Drug Discov.* 8, 42–47. doi: 10.2174/1872214808666140209142640

Ou, C. Y., Chen, C. Z., Hsiue, T. R., Lin, S. H., and Wang, J. Y. (2015). Genetic variants of pulmonary SP-D predict disease outcome of COPD in a Chinese population. *Respirology* 20, 296–303. doi: 10.1111/resp.12427

Ozyurek, B. A., Ulasli, S. S., Bozbas, S. S., Bayraktar, N., and Akcay, S. (2013). Value of serum and induced sputum surfactant protein-D in chronic obstructive pulmonary disease. *Multi. Discip. Respir. Med.* 8:36. doi: 10.1186/2049-6958-8-36

Shakoori, T. A., Sin, D. D., Bokhari, S. N., Ghafoor, F., and Shakoori, A. R. (2012). SP-D polymorphisms and the risk of COPD. *Dis. Markers.* 33, 91–100. doi: 10.3233/DMA-2012-0909

Shakoori, T. A., Sin, D. D., Ghafoor, F., Bashir, S., and Bokhari, S. (2009). Serum surfactant protein D during acute exacerbations of chronic obstructive pulmonary disease. *Dis. Markers.* 27, 287–294. doi: 10.3233/DMA-2009-0674

Sorensen, G. L., Hoegh, S. V., Leth-Larsen, R., Thomsen, T. H., Floridon, C., Smith, K., et al. (2009). Multimeric and trimeric subunit SP-D are interconvertible structures with distinct ligand interaction. *Mol. Immunol.* 46, 3060–3069. doi: 10.1016/j.molimm.2009.06.005

Zemans, R. L., Jacobson, S., Keene, J., Kechris, K., Miller, B. E., Tal-Singer, R., et al. (2017). Multiple biomarkers predict disease severity, progression and mortality in COPD. *Respir. Res.* 18:117. doi: 10.1186/s12931-017-0597-7

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 10 May 2019
doi: 10.3389/fgene.2019.00428

A Social Determinant of Health May Modify Genetic Associations for Blood Pressure: Evidence From a SNP by Education Interaction in an African American Population

Brittany M. Hollister[1], Eric Farber-Eger[2], Melinda C. Aldrich[3] and Dana C. Crawford[4]**

[1] Social and Behavioral Research Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, United States, [2] Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University Medical Center, Nashville, TN, United States, [3] Department of Thoracic Surgery, Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, United States, [4] Department of Population and Quantitative Health Sciences, Cleveland Institute for Computational Biology, Case Western Reserve University, Cleveland, OH, United States

African Americans experience the highest burden of hypertension in the United States compared with other groups. Genetic contributions to this complex condition are now emerging in this as well as other populations through large-scale genome-wide association studies (GWAS) and meta-analyses. Despite these recent discovery efforts, relatively few large-scale studies of blood pressure have considered the joint influence of genetics and social determinants of health despite extensive evidence supporting their impact on hypertension. To identify these expected interactions, we accessed a subset of the Vanderbilt University Medical Center (VUMC) biorepository linked to de-identified electronic health records (EHRs) of adult African Americans genotyped using the Illumina Metabochip (n = 2,577). To examine potential interactions between education, a recognized social determinant of health, and genetic variants contributing to blood pressure, we used linear regression models to investigate two-way interactions for systolic and diastolic blood pressure (DBP). We identified a two-way interaction between rs6687976 and education affecting DBP (p = 0.052). Individuals homozygous for the minor allele and having less than a high school education had higher DBP compared with (1) individuals homozygous for the minor allele and high school education or greater and (2) individuals not homozygous for the minor allele and less than a high school education. To our knowledge, this is the first EHR -based study to suggest a gene-environment interaction for blood pressure in African Americans, supporting the hypothesis that genetic contributions to hypertension may be modulated by social factors.

Keywords: electronic health records, social determinants of health, African Americans, blood pressure, gene-environment, education

OPEN ACCESS

Edited by:
C. Charles Gu, Washington University in St. Louis, United States

Reviewed by:
Tesfaye B. Mersha, Cincinnati Children's Hospital Medical Center, United States
Kenneth M. Weiss, The Pennsylvania State University, United States

****Correspondence:***
Melinda C. Aldrich
melinda.aldrich@vumc.org
Dana C. Crawford
dcc64@case.edu;
dana.crawford@case.edu

Specialty section:
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics

Received: *30 November 2018*
Accepted: *18 April 2019*
Published: *10 May 2019*

Citation:
Hollister BM, Farber-Eger E, Aldrich MC and Crawford DC (2019) A Social Determinant of Health May Modify Genetic Associations for Blood Pressure: Evidence From a SNP by Education Interaction in an African American Population. Front. Genet. 10:428. doi: 10.3389/fgene.2019.00428

INTRODUCTION

African Americans have a higher prevalence of hypertension, or chronically high blood pressure, compared with other racial/ethnic groups (Yoon et al., 2015; Writing Group Members et al., 2016). Despite this higher burden of disease in African Americans, early genome-wide association studies (GWAS) for hypertension and systolic blood pressure (SBP) and diastolic blood pressure (DBP) were limited to populations of European-descent (Levy et al., 2009; Newton-Cheh et al., 2009; Wang et al., 2009; International Consortium for Blood Pressure Genome-Wide Association Studies et al., 2011) or east Asian-descent (Kato et al., 2011). More recent GWAS have been performed in ancestrally diverse populations, including African Americans or African-descent populations (Adeyemo et al., 2009; Zhu et al., 2011, 2015; Kidambi et al., 2012; Franceschini et al., 2013; Hoffmann et al., 2017; Liang et al., 2017). Collectively, these associated common variants explain 3–6% of the variance for SBP and DBP, and in the largest European-descent study to date account for up to 27% of the estimated single nucleotide polymorphism (SNP)-wide heritability for these traits (Evangelou et al., 2018).

Current GWAS findings explain only a proportion of the expected contribution from additive genetic effects. Previous twin and family studies estimate these traits have moderate to high heritability (30–70%) (Fagard et al., 1995; Rotimi et al., 1999; Levy et al., 2000; Hottenga et al., 2005; Kupper et al., 2005), suggesting that additional genetic associations have yet to be discovered. Given that GWAS identify common single nucleotide variants (SNVs) for association, additional genetic associations may be found among rare SNVs (Doris, 2011; Russo et al., 2018). Importantly, most GWAS consider only main effects and do not consider interactions with relevant environmental exposures. Two recent and large GWAS of blood pressure have considered alcohol consumption (Feitosa et al., 2018) and smoking (Sung et al., 2018), both of which identified novel putative associations for these traits.

Here, we examine the modifying effects of education, a measure of socioeconomic status (SES) and recognized social determinant of health, on SBP and DBP traits among African Americans drawn from a clinical setting. Previous epidemiologic studies suggest that in addition to alcohol consumption and smoking, social environment and specifically SES has a strong influence on blood pressure and hypertension (Seeman et al., 2008; Cha et al., 2012; Non et al., 2012). Further, a GWAS in the Framingham Heart Study accounting for educational attainment identified novel associations for blood pressure traits among European Americans (Basson et al., 2014). Based on these prior findings, we hypothesized that educational attainment modifies associations between genetic variants and blood pressure among African Americans. To test this hypothesis, we accessed a large biobank linked to electronic health records (EHRs) in a racially diverse clinical population. We identified two associated SNPs, *ARHGAP22* rs4593967 (SBP) and *IQCK* rs950928 (DBP), neither of which has been previously associated with blood pressure. We also identified a novel SNP-education interaction affecting DBP, suggesting social determinants of health may modify genetic effects contributing to complex human traits.

MATERIALS AND METHODS

Study Population and Data Collection

The study population is derived from BioVU, a DNA biobank of the Vanderbilt University Medical Center (VUMC) linked to de-identified EHRs. DNA samples are extracted from discarded blood samples drawn for routine clinical care (Roden et al., 2008). These samples are linked to the Synthetic Derivative (SD), the de-identified version of the VUMC EHR. Medical records within the SD are scrubbed of all Health Insurance Portability and Accountability Act (HIPAA) identifiers. This study was approved by the Vanderbilt University Institutional Review Board.

The study population consists of African American adults >18 years old drawn from a larger study of minority patients with DNA samples in BioVU (n = 15,863) (Crawford et al., 2015). We extracted relevant demographic variables, including race/ethnicity, sex, and age at data extraction available in the SD. Smoking status was extracted using International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) tobacco use codes as previously described (Wiley et al., 2013). Education was extracted from the free text of EHRs using a recently validated text-mining algorithm (Hollister et al., 2016). Education was modeled as a categorical variable: less than high school, high school, and some college or above. All weight and height measures were extracted from the EHR, and after extensive quality control, as described in Goodloe et al. (2017), median values were used to represent individual-level body mass index (BMI).

The median value of all blood pressure measurements within an individual's EHR prior to a recording of blood pressure-altering medications in the patient's medication list were used in analyses. Medications included in the keyword list of anti-hypertensives were angiotensin converting enzyme inhibitors, angiotensin II receptor blocker, beta blockers, non-dihydropyridine calcium channel blockers, hydralazine, Minoxidil, central alpha antagonists, direct renin antagonists, aldosterone antagonists, alpha antagonists, and diuretics including thiazides, K-sparing, and loop diuretics. Any blood pressure measurement found after any mention of these types of medications were excluded from analyses.

Genotyping and Quality Control

Genotyping of 15,863 DNA samples from non-European descent individuals was performed using the Metabochip, a custom Illumina genotyping array designed to target SNPs and surrounding genomic regions associated with metabolic traits and cardiovascular disease (Buyske et al., 2012; Voight et al., 2012). We restricted the following quality control and statistical analyses to DNA samples from African Americans in BioVU (n = 11,301). All genotyping quality control was performed using PLINK 1.9 (Chang et al., 2015). After the removal of SNPs with a minor allele frequency of less than 5%, SNPs with a Hardy-Weinberg Equilibrium exact test p-value of less than

1×10^{-7}, and SNPs with a genotyping call rate of less than 95%, a total of 115,834 variants remained (**Supplementary Figure S1**). We further removed 967 samples for either ambiguous sex, missing genotypes (>5%), or relatedness (twins, full siblings, parent/offspring) (**Supplementary Figure S1**). A total of 10,334 DNA samples passed genotyping quality control. After quality control, global ancestry was estimated using unsupervised ADMIXTURE analysis, assuming $K = 2$ (Alexander et al., 2009). Linkage disequilibrium (r^2) was calculated using 1000 Genomes Phase 3 data and an expectation-maximization algorithm adapted from Haploview (Barrett et al., 2005) available through rAggr (Edlund et al., 2017).

Local ancestry for rs6687976 was assigned as previously described (Fish et al., 2018). Briefly, SHAPEITv2 (Delaneau et al., 2013) and the 1000 Genomes Phase 3 reference panel[1] were used to phase the genotype data. RFMix (Maples et al., 2013) was used to assign local ancestry. Phased chromosomal haplotypes were matched to Yoruba and CEPH/European ancestral population panels from 1000 Genomes.

Statistical Analysis

Inclusion criteria included African American adults with available Metabochip genotyping data and complete information on age, sex, BMI, premedication SBP, premedication DBP, smoking status, and education level. A total of 2,577 African Americans met genotyping quality control and had relevant covariates (**Supplementary Figure S2**). All statistical analyses were performed using PLINK 1.9 (Chang et al., 2015) or R (R Core Team, 2008). Linear regression models were used to identify genetic variants associated with either premedication SBP or premedication DBP. A Bonferroni adjusted p-value of 4.32×10^{-7} was used to determine significance. A main effects model included covariates for age, age squared, sex, BMI, smoking status, and percent global African ancestry:

$$Premedication\ SBP\ or\ DBP = \beta_0 + \beta_{cov}{}^{*}X_{cov} + \beta_1{}^{*}SNP + e$$

A second main effects model included the same covariates, but also included education. To examine the interaction between genetic variants and education and how it may affect blood pressure, we modeled two-way interactions using a linear regression model and the same covariates as in our main effects model:

$$Premedication\ SBP\ or\ DBP = \beta_0 + \beta_{cov}{}^{*}X_{cov} + \beta_1{}^{*}SNP + \beta_2{}^{*}Education + \beta_3{}^{*}SNP{}^{*}Education + e$$

The decision was made to focus on a set of SNPs which had a p-value of less than 1.4×10^{-5} from the main effects model to reduce issues with multiple testing. This significance threshold was chosen based on a Bonferroni correction for the number of SNPs that would remain if SNPs with an r^2-value of greater than 0.1 were removed from our dataset. For this set of significant SNPs, we used a model which included the main effects of education and the SNP, as well as the interaction term between education and the genetic variants. The significance threshold for the interaction models was based on the number of SNPs tested for association with premedication SBP and DBP ($p < 0.01$ and $p < 0.003$, respectively).

[1]https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#reference

RESULTS

Population Characteristics

The final study population for analysis included 2,577 African American adults with Metabochip genotyping data and complete phenotype data (**Supplementary Figures S1, S2**). Among this study population, the majority were female (71%) with a median age of 38 years and median BMI of 26.8 kg/m^2 (**Table 1**). Compared with a larger African American BioVU population genotyped on the Metabochip (Crawford et al., 2015), this subset had proportionally more females, was younger, and had a lower median BMI. The median premedication SBP and DBP were within the normal clinical range (122 and 74 mmHg, respectively) and most of the population was never smokers (87%; **Table 1**). The median percent global African ancestry was 81.7%. The majority of participants had at least a high school degree (**Table 1**). The median premedication SBP and DBP in this final study sample were statistically different ($p < 0.05$) from the larger study sample missing education data in the EHR but varied by only 3 mmHg (**Supplementary Table S1**).

TABLE 1 | Study population characteristics representing African American adults from a biobank with electronic health record (EHR)-extracted blood pressure.

Characteristic	Number of individuals n = 2,577
Sex	
Male	753 (29%)
Female	1,824 (71%)
Median age in years (±SD)	38.2 (±15.4)
Median percent global African ancestry (range)	81.7% (1.0–99.9)
Education level	
Less than high school	328 (12.7%)
High school degree or equivalent	1518 (58.9%)
At least some college	731 (28.4%)
Smoking Status	
Never smokers	2,242 (87%)
Ever smokers	335 (13%)
Median body mass index (kg/m^2; ±SD)	26.8 (±6.8)
Median premedication systolic blood pressure (mmHg; ±SD)	122.0 (±13.6)
Median Premedication diastolic blood pressure (mmHg; ±SD)	74.4 (±9.0)

The population in this study was a subset of African Americans from the Vanderbilt University Medical Center (VUMC) biobank, BioVU. Samples were drawn from BioVU in 2011. All individuals had Metabochip genotype data which passed quality control measures. Individuals also had complete phenotype data which included age, sex, education level, smoking status, median body mass index (BMI), median premedication systolic blood pressure (SBP), and median diastolic blood pressure (DBP). These phenotypes were derived from the electronic health record. African ancestry was determined using ADMIXTURE. SD, standard deviation.

Predictors of Systolic and Diastolic Blood Pressure

In univariate analyses (**Table 2**), both premedication SBP and DBP were significantly associated with increasing age, male sex, and increasing BMI. SBP increased with age and DBP increased with age until around the age of 60, then began decreasing (**Supplementary Figures S3, S4**).

Neither premedication SBP nor premedication DBP was associated with smoking status or global African ancestry. Also, education was not significantly associated with either premedication SBP or premedication DBP (**Table 2** and **Supplementary Figures S5, S6**). Of all the variables tested, age and premedication DBP significantly co-varied with education (**Supplementary Table S2**).

Education as a Modifier of Genetic Associations With Systolic and Diastolic Blood Pressure

To test for possible interactive effects between education and genetic variants associated with SBP and DBP, we examined three models: (1) initial single SNP tests of association without education as a covariate, (2) single SNP tests of association with education as a covariate, and (3) single SNP tests of association with SNP × education interaction terms. In the first model, single SNP tests of association were performed for SBP and DBP using linear regression adjusting for age, age squared, sex, BMI, smoking status, and percent global African ancestry. For both SBP (**Supplementary Figure S7**) and DBP (**Supplementary Figure S8**), only a single SNP was statistically significant using a Bonferroni correction ($p < 4.32 \times 10^{-7}$): *ARHGAP22* rs4593967 and *IQCK* rs950928, respectively.

The second set of models included education in addition to other relevant covariates (**Supplementary Figures S9, S10**). The addition of education to the model did not change the most significantly associated SNPs for either SBP or DBP (**Table 3**). In the regression model for SBP that included education, rs4593967 again passed Bonferroni correction ($p < 4.32 \times 10^{-7}$), and two other SNPs (rs10921895 and rs3804485) were associated at a suggestive significance threshold ($p < 7.24 \times 10^{-6}$). For DBP, rs950928 and rs8056711 passed Bonferroni correction. However, these SNPs have the same effect size and are in perfect linkage disequilibrium ($r^2 = 1.0$), so they likely represent the same association.

In the final set of models, education × SNP interaction terms were examined using SNPs associated with SBP or DBP at $p < 1.4 \times 10^{-5}$, as described above. No interaction terms met a strict Bonferroni correction (**Supplementary Figures S11, S12**). However, we identified a potential SNP-education interaction affecting DBP, rs6687976 ($p = 0.052$; **Table 4**). This potential interaction remained with the addition of local ancestry to the model. Individuals homozygous for the minor allele and having less than a high school education had higher DBP compared with (1) individuals homozygous for the minor allele and high school education or greater and (2) individuals not homozygous for the minor allele and less than a high school education (**Supplementary Figure S13**). No statistically significant interactions were identified for SBP (**Table 4**).

DISCUSSION

We sought to determine if education, a measure of SES and a recognized social determinant of health, modified genetic

TABLE 2 | Univariate analyses between relevant covariates and median premedication blood pressure values among African American adults.

Variable	Effect estimate, β (standard error)	*p*-value	Effect estimate, β (standard error)	*p*-value
	Premedication SBP		Premedication DBP	
Education level				
Less than high school	REF		REF	
High school degree and equivalent	0.16 (0.83)	0.84	1.95 (0.54)	0.0003*
At least some college	1.26 (0.91)	0.17	2.54 (0.59)	<0.0001*
Median age, years	0.31 (0.01)	<0.0001*	0.19 (0.01)	<0.0001*
Sex				
Male	REF		REF	
Female	−4.52 (0.52)	<0.0001*	−1.48 (0.35)	0.0001*
Median BMI (kg/m^2)	0.38 (0.04)	<0.0001*	0.25 (0.02)	<0.0001*
Smoking status				
Never smoker	REF		REF	
Ever smoker	0.56 (0.72)	0.43	0.45 (0.48)	0.35
Median global African ancestry	−0.46 (1.89)	0.809	−1.24 (1.24)	0.32

Prior to genetic analyses, covariates were examined to determine their association with the outcomes, premedication systolic (SBP) and diastolic blood pressure (DBP) in the study population, a subset of African Americans drawn from the Vanderbilt University Medical Center biobank BioVU (n = 2,577). Each linear regression model had either median premedication SBP or median premedication DBP as the outcome. The covariates included in each model were education level, median age, sex, body mass index (BMI), smoking status, and global African ancestry. Both premedication SBP and DBP were significantly associated with age, sex, and BMI. Premedication DBP is also significantly associated with education level. The symbol "" indicates statistical significance.*

TABLE 3 | Characteristics of single nucleotide polymorphisms (SNPs) associated with premedication systolic and diastolic blood pressure with and without education in the model.

					Without education			With education		
SNP	**Trait**	**Location**	**Gene**	**MAF**	**Effect estimate**	**Standard error**	***p*-value**	**Effect estimate**	**Standard error**	***p*-value**
rs4593967	SBP	Intron	*ARHGAP22*	0.14	−2.51	0.46	1.45×10^{-7}	−2.53	0.48	**1.16×10^{-7}**
rs10921895	SBP	Intergenic		0.37	−1.52	0.37	5.31×10^{-6}	−1.55	0.36	3.92×10^{-6}
rs3804485	SBP	Intron	*LY86*	0.41	1.48	0.32	7.11×10^{-6}	1.51	0.33	5.20×10^{-6}
rs950928	DBP	Intron	*IQCK*	0.36	−1.04	0.24	3.13×10^{-6}	−1.10	0.22	**4.53×10^{-7}**
rs8056711	DBP	Intron	*IQCK*	0.36	−1.04	0.24	3.13×10^{-6}	−1.10	0.22	**4.53×10^{-7}**

In the both sets of linear regression models, median premedication systolic blood pressure (SBP) and median premedication diastolic blood pressure (DBP) were the outcomes. Additionally, both sets of linear regression models included age, age squared, sex, median body mass index (BMI), smoking status, and median percent global African ancestry as covariates. The first set of models did not include education level. The second set of models included education. The addition of education to the model did not change which SNPs were most associated with SBP or DBP. Bolded p-values are considered statistically significant after Bonferroni correction.

TABLE 4 | Single nucleotide polymorphisms (SNPs) examined for interactions with education level impacting median premedication systolic and diastolic blood pressure.

SNP in education interaction term	***p*-value for the interaction between high school and SNP**	***p*-value for interaction between college and SNP**
Systolic Blood Pressure		
rs4593967_A	0.886	0.858
rs10921895_G	0.896	0.746
rs3804485_C	0.260	0.863
rs11066700_A	0.178	0.200
Diastolic Blood Pressure		
rs950928_G	0.175	0.298
rs8056711_C	0.175	0.297
rs9931167_A	0.927	0.503
rs11648873_T	0.940	0.478
rs28681321_A	0.941	0.461
rs6687976_A	**0.052**	0.996
chr16.19642355_G	0.094	0.110
rs3095994_A	0.738	0.726
rs1273518_G	0.218	0.648
rs1858975_A	0.076	0.091
rs4593967_A	0.512	0.768
rs8046628_C	0.181	0.316
rs6497402_A	0.062	0.085
chr16.19641087_A	0.064	0.076

Median premedication systolic blood pressure (SBP) and diastolic blood pressure (DBP) were outcomes in the linear regression models. Covariates included in the models were age, age squared, sex, body mass index, smoking status, and African ancestry. The main effect of education and the SNP, as well as the SNP × education interaction term were also included in the model. Less than high school was the reference group within the regression models. The p-value for the potential SNP-education interaction is bolded.

associations with SBP and DBP in African Americans. A previous study suggested gene × education interactions occur with blood pressure, but this study was conducted in a European-descent population (Basson et al., 2014). Associations between premedication SBP or premedication DBP and genetic variants from the Metabochip were examined, while including known predictors of blood pressure (age, BMI, sex, percent African ancestry, and smoking status) in the model. Results were compared with models which included a main effect for education, and a main effect for education plus a SNP-education interaction term. We observed a suggestive SNP by education interaction affecting DBP, a result not explained by local genetic ancestry. This potential interaction requires statistical replication and further investigation.

Models Without Interaction

In univariate analyses the associations between premedication SBP and DBP and increasing age, male sex as well as increasing BMI were consistent with previous reports (August, 1999; Wright et al., 2011; Dua et al., 2014). The patterns of associations between SBP and DBP and age across the age continuum are also consistent with previous reports (Liang et al., 2017; Evangelou et al., 2018).

Intronic *ARHGAP22* rs4593967 was significantly associated with SBP and has not been previously reported as associated with blood pressure or hypertension. The minor allele frequency for *ARHGAP22* rs4593967 in this African American sample was 0.14, consistent with frequencies reported for African-descent populations included in The Genome Aggregation Database (0.148; Lek et al., 2016) and the 1000 Genomes Project (0.176; 1000 Genomes Project Consortium et al., 2015). Conversely, the minor allele is less frequently observed among populations of European (∼0.08) or East Asian-descent (<0.01). No other common (MAF > 1%) variants within 500 kb are in strong linkage disequilibrium ($r^2 \geq 0.80$) with rs4593967 in African-descent populations from the 1000 Genomes Project. *ARHGAP22* encodes the rho GTPase activating protein 22 and is widely expressed with highest expression levels in the brain. Variants within *ARHGAP22* have been associated with diabetic retinopathy, conduct disorder, daytime sleep, and self-employment (Dick et al., 2011; Huang et al., 2011; Van Der Loos et al., 2013; Spada et al., 2016), but these associations have not been replicated.

Intronic *IQCK* rs950928 was significantly associated with DBP after adjusting for multiple testing. Like *ARHGAP22* rs4593967, the minor allele frequency for *IQCK* rs950928 is

higher among populations of African-descent (∼0.40) compared with European-descent populations (∼0.15). *IQCK* rs950928 is in perfect or strong linkage disequilibrium with rs8056711 and rs59009734 in African-descent populations, neither of which has been previously associated with human disease or traits. *IQCK*, which overlaps with several genes including *KNOP1*, encodes for IQ motif containing K and serves as an EF hand protein binding site. Like *ARHGAP22*, *IQCK* is highly expressed in the brain. A search within the Genotype-Tissue Expression (GTEx Consortium, 2013) database suggests that both rs8056711 and rs59009734 may be expression quantitative loci (eQTL), where each addition of the minor allele is associated with higher gene expression for several tissues including the right atrium auricular region of the heart and the aorta. While *IQCK* rs950928 and its associated SNPs rs8056711 and rs59009734 have not been previously associated with any phenotypes, common variants within *IQCK* have previously been associated with blood pressure, BMI, bone density, heart rate, chronic obstructive pulmonary disease, bipolar disorder, and a BMI-education interaction (Cho et al., 2009; Liu et al., 2010; Wan et al., 2011; Boardman et al., 2014; Winham et al., 2014).

Despite the present study's small sample size (n = 2,577), there was sufficient power (80%) to detect significant associations with moderate effect size of 1.0 and a minor allele frequency of 0.20. For less common variants (MAF = 0.10), the study was powered to detect alleles with an effect size of 1.5 or greater. For variants with a MAF of 0.05, an effect size of 2.0 was needed in order to detect the variant's effect. This study was not powered to detect any of the variants reported in the recent one million-person GWAS of blood pressure, as the variant with the largest effect size in that study was less than 1.0, with a median effect size of 0.219 mmHg (Evangelou et al., 2018). The limited power due to small sample size and limited directly genotyped variants likely contributed to the lack of replication of SNPs known to be associated with blood pressure in African Americans from previous GWAS.

SNP × Education Interactions

We identified a possible SNP-education interaction affecting DBP for rs6687976 (p = 0.052). As the addition of local ancestry to the model did not alter the association, we expect that this observation is a result of true modifying effects of SES rather than ancestry. Individuals with two minor alleles and less than a high school education had higher blood pressure compared to those with two minor alleles and a high school education or those with less than a high school education and fewer minor alleles (**Supplementary Figure S13**). SNP rs6687976 is located within an intergenic region of chromosome 1 (Chr1:105674536 in GRCh37.p13) and has not been previously associated with any human traits within the literature. It is also not identified as an eQTL in GTEx (GTEx Consortium, 2013). Despite the limited information known about rs6687976, this result suggests that interactions between markers of social determinants of health and genetic variants affecting blood pressure likely exist, consistent with the findings of other studies that have observed interactions between genetic variants and social factors such as depression (Smith et al., 2017), perceived discrimination (Taylor et al., 2017), and cigarette smoking (Taylor et al., 2016).

Limitations

The present study has several limitations. Primarily, the sample size is limited driven by the inclusion criteria of complete phenotype data for a specific racial/ethnic group within the larger clinical dataset. Therefore, we are unable to detect any variants of smaller effect sizes. The requirement for complete data may have also introduced biases that limit the interpretation and generalizability of these data.

In addition to the limited sample size, the study population was also different compared with previously published studies of blood pressure in African American populations. While the proportion of females, median BMI, percent African ancestry, median SBP, and median DBP were comparable with previous studies (Parra et al., 1998; Dumitrescu et al., 2015; Baharian et al., 2016; Franceschini et al., 2016; Jones et al., 2018; Restrepo et al., 2018), the population in this study did have a much lower median age, over 15 years younger. Given that blood pressure increases with age, this younger study population may have reduced variability in blood pressure measurements compared with the older published study populations with right-skewed distributions (Wright et al., 2011).

Another limitation was the lack of a replication dataset; therefore, all associations reported here are putative pending statistical replication or corroborative functional data. To date, other studies comparable or larger in sample size have not yet reported associations between these SNPs and blood pressure (Hoffmann et al., 2017). Furthermore, the genotyping array used here was also designed to include rare variation collected from the African ancestry samples as part of the 1000 Genomes Project. Therefore, many of the variants on the Metabochip were rare in African ancestry populations (Buyske et al., 2012) and filtered out during the quality control process as the present study was not powered to detect associations for rare variation.

There were also limitations regarding the phenotype data. All the variables were extracted from EHRs. While these records have extensive amounts of data, the data recorded by healthcare providers are not always accurate and the ability to extract the data can be limited. Furthermore although the positive predictive value of our algorithm was 80% (Hollister et al., 2016), there may have been inaccurate education information for the individuals within the dataset.

Determining which blood pressure measurements to use in the study is also a challenge, as measurements can vary widely across the EHR. The median blood pressure measurements were chosen for our study to reduce the influence of this variation. Beyond the inaccuracies and decisions to be made regarding the information within the EHR, blood pressure is difficult to measure within the clinic. Measurements of blood pressure can vary due to the calibration of instruments, the time of day it is measured, and due to illness (Jones et al., 2003). Patients also tend to have higher blood pressure within a clinical setting due to stress (Jones et al., 2003). To avoid these potential biases as much as possible, we chose median premedication blood pressure

values for analysis, thereby avoiding outlier measurements and the changes introduced by blood pressure medications.

Finally, while education is a recognized social determinant of health, it is not a perfect proxy for social experiences. Still, evidence suggests that educational attainment can be a reflection of earning potential and social status (Shavers, 2007; Tamborini et al., 2015). Education has been shown to be associated with life expectancy, numerous biomarkers, and other health outcomes such as obesity and smoking (Seeman et al., 2008; National Center for Health Statistics, 2012). Low educational attainment itself is not the cause of poor health outcomes, but rather a variable often associated with individual-level behavioral determinants (e.g., smoking) or community-level determinants (e.g., racial segregation) that may influence blood pressure. Neither of these determinants is routinely recorded with the EHR; in contrast, educational attainment is often mentioned in the EHR. The availability of these data coupled with the observation that individual educational attainment is often stable over time make this variable a robust albeit imperfect proxy for social experiences.

Strengths

Despite the limitations within the study, there were also several strengths. Primarily, this is the first study to incorporate EHR-derived education information into a large-scale genetic investigation. This study is a proof of principle that EHR-derived social determinant information can be investigated in a GWAS setting, thus breaking new ground to incorporate social factors in genetic studies among biobank populations. This is also the first analysis to observe an interaction between education and a common genetic variant with blood pressure in an African American population.

Despite the consistent association between social environment and health, social determinants of health are typically not included in genetic studies of health outcomes. For studies that access biobanks, the lack of social determinant data is likely related to the difficulty in accessing these data within the EHR, where they are not usually recorded in structured fields. The algorithms used in our study are the first to extract these important data from EHRs for research purposes (Hollister et al., 2016).

This study paves the road for the incorporation of education, as well as other social determinants of health, into genetic studies using biobank populations. The SNP-by-education interaction we observed affecting DBP (rs6687976) could suggest an example of a possible biological impact of the adversity experienced due to lower educational achievement. Only individuals homozygous for the minor allele who had less than a high school education experienced an increase in DBP. This association needs to be replicated; however, it suggests a potential pathway for the biological imbedding of stress experiences (represented by lower educational attainment) affecting blood pressure and risk for hypertension. Further studies are needed to support this hypothesis. We anticipate that this research will encourage other investigators to continue to study the genetics of health outcomes associated with racial health disparities and to incorporate social determinants of health within these studies.

AUTHOR CONTRIBUTIONS

BH conducted the analyses and wrote manuscript. EF-E helped to extract the phenotype data from the electronic health record. MA and DC contributed to guidance on project, and manuscript writing and editing.

FUNDING

This work was supported by the National Institutes of Health (NIH) U01 HG004798 and its ARRA supplements (DC), as well as National Cancer Institute 1K07CA172294 (MA). This publication was also made possible by the Clinical and Translational Science Collaborative of Cleveland, 4UL1TR0002548, from the National Center for Advancing Translational Sciences (NCATS) component of the National Institutes of Health and NIH Roadmap for Medical Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. The dataset (s) used for the analyses described were obtained from Vanderbilt University Medical Center's BioVU, which is supported by institutional funding and the National Center for Research Resources, grant UL1 RR024975–01 (now at NCATS, grant 2UL1 TR000445–06).

ACKNOWLEDGMENTS

We would like to thank Drs. Alex Fish and William Bush for access to the local genetic ancestry data, and we further thank Dr. Bush for helpful comments during the revision process.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00428/full#supplementary-material

REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Adeyemo, A., Gerry, N., Chen, G., Herbert, A., Doumatey, A., Huang, H., et al. (2009). A genome-wide association study of hypertension and blood pressure in African Americans. *PLoS Genet.* 5:e1000564. doi: 10.1371/journal.pgen.1000564

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

August, P. (1999). Hypertension in men. *J. Clin. Endocrinol. Metab.* 84, 3451–3454.

Baharian, S., Barakatt, M., Gignoux, C. R., Shringarpure, S., Errington, J., Blot, W. J., et al. (2016). The great migration and African-American genomic diversity. *PLoS Genet.* 12:e1006059. doi: 10.1371/journal.pgen.1006059

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263–265. doi: 10.1093/bioinformatics/bth457

Basson, J., Sung, Y. J., Schwander, K., Kume, R., Simino, J., De Las Fuentes, L., et al. (2014). Gene-education interactions identify novel blood pressure loci in the Framingham Heart Study. *Am. J. Hypertens.* 27, 431–444. doi: 10.1093/ajh/hpt283

Boardman, J. D., Domingue, B. W., Blalock, C. L., Haberstick, B. C., Harris, K. M., and Mcqueen, M. B. (2014). Is the gene-environment interaction paradigm relevant to genome-wide studies? The case of education and body mass index. *Demography* 51, 119–139. doi: 10.1007/s13524-013-0259-4

Buyske, S., Wu, Y., Carty, C. L., Cheng, I., Assimes, T. L., and Dumitrescu, L. (2012). Evaluation of the metabochip genotyping array in African Americans and implications for fine mapping of GWAS-identified loci: the PAGE study. *PLoS One* 7:e35651. doi: 10.1371/journal.pone.0035651

Cha, S. H., Park, H. S., and Cho, H. J. (2012). Socioeconomic disparities in prevalence, treatment, and control of hypertension in middle-aged Koreans. *J. Epidemiol.* 22, 425–432. doi: 10.2188/jea.je20110132

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4:7. doi: 10.1186/s13742-015-0047-8

Cho, Y. S., Go, M. J., Kim, Y. J., Heo, J. Y., Oh, J. H., Ban, H. J., et al. (2009). A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat. Genet.* 41, 527–534. doi: 10.1038/ng.357

GTEx Consortium. (2013). The genotype-tissue expression (GTEx) project. *Nat. Genet.* 45, 580–585. doi: 10.1038/ng.2653

Crawford, D. C., Goodloe, R., Farber-Eger, E., Boston, J., Pendergrass, S. A., Haines, J. L., et al. (2015). Leveraging epidemiologic and clinical collections for genomic studies of complex traits. *Hum. Hered.* 79, 137–146. doi: 10.1159/000381805

Delaneau, O., Zagury, J. F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* 10, 5–6. doi: 10.1038/nmeth.2307

Dick, D. M., Aliev, F., Krueger, R. F., Edwards, A., Agrawal, A., Lynskey, M., et al. (2011). Genome-wide association study of conduct disorder symptomatology. *Mol. Psychiatry* 16, 800–808. doi: 10.1038/mp.2010.73

Doris, P. A. (2011). The genetics of blood pressure and hypertension: the role of rare variation. *Cardiovasc. Ther.* 29, 37–45. doi: 10.1111/j.1755-5922.2010.00246.x

Dua, S., Bhuker, M., Sharma, P., Dhall, M., and Kapoor, S. (2014). Body mass index relates to blood pressure among adults. *N. Am. J. Med. Sci.* 6, 89–95. doi: 10.4103/1947-2714.127751

Dumitrescu, L., Restrepo, N. A., Goodloe, R., Boston, J., Farber-Eger, E., Pendergrass, S. A., et al. (2015). Towards a phenome-wide catalog of human clinical traits impacted by genetic ancestry. *Biodata Min.* 8:35. doi: 10.1186/s13040-015-0068-y

Edlund, C. K., Conti, D. V., and Van Den Berg, D. J. (2017). *rAggr*. Available at: http://raggr.usc.edu/ (accessed November 7, 2018).

Evangelou, E., Warren, H. R., Mosen-Ansorena, D., Mifsud, B., Pazoki, R., Gao, H., et al. (2018). Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nat. Genet.* 50, 1412–1425. doi: 10.1038/s41588-018-0205-x

Fagard, R., Brguljan, J., Staessen, J., Thijs, L., Derom, C., Thomis, M., et al. (1995). Heritability of conventional and ambulatory blood pressures. A study in twins. *Hypertension* 26, 919–924. doi: 10.1161/01.hyp.26.6.919

Feitosa, M. F., Kraja, A. T., Chasman, D. I., Sung, Y. J., Winkler, T. W., Ntalla, I., et al. (2018). Novel genetic associations for blood pressure identified via gene-alcohol interaction in up to 570K individuals across multiple ancestries. *PLoS One* 13:e0198166. doi: 10.1371/journal.pone.0198166

Fish, A. E., Crawford, D. C., Capra, J. A., and Bush, W. S. (2018). Local ancestry transitions modify SNP-trait associations. *Pac. Symp. Biocomput.* 23, 424–435.

Franceschini, N., Carty, C. L., Lu, Y., Tao, R., Sung, Y. J., Manichaikul, A., et al. (2016). Variant discovery and fine mapping of genetic loci associated with blood pressure traits in Hispanics and African Americans. *PLoS One* 11:e0164132. doi: 10.1371/journal.pone.0164132

Franceschini, N., Fox, E., Zhang, Z., Edwards, T. L., Nalls, M. A., Sung, Y. J., et al. (2013). Genome-wide association analysis of blood-pressure traits in African-ancestry individuals reveals common associated genes in African and non-African populations. *Am. J. Hum. Genet.* 93, 545–554. doi: 10.1016/j.ajhg.2013.07.010

Goodloe, R., Farber-Eger, E., Boston, J., Crawford, D. C., and Bush, W. S. (2017). Reducing clinical noise for body mass index measures due to unit and transcription errors in the electronic health record. *AMIA Jt. Summits Transl. Sci. Proc.* 2017, 102–111.

Hoffmann, T. J., Ehret, G. B., Nandakumar, P., Ranatunga, D., Schaefer, C., Kwok, P. Y., et al. (2017). Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. *Nat. Genet.* 49, 54–64. doi: 10.1038/ng.3715

Hollister, B. M., Restrepo, N. A., Farber-Eger, E., Crawford, D. C., Aldrich, M. C., and Non, A. (2016). Development and performance of text-mining algorithms to extract socioeconomic status from de-identified electronic health records. *Pac. Symp. Biocomput.* 22, 230–241. doi: 10.1142/9789813207813_0023

Hottenga, J. J., Boomsma, D. I., Kupper, N., Posthuma, D., Snieder, H., Willemsen, G., et al. (2005). Heritability and stability of resting blood pressure. *Twin Res. Hum. Genet.* 8, 499–508. doi: 10.1375/183242705774310123

Huang, Y. C., Lin, J. M., Lin, H. J., Chen, C. C., Chen, S. Y., Tsai, C. H., et al. (2011). Genome-wide association study of diabetic retinopathy in a Taiwanese population. *Ophthalmology* 118, 642–648. doi: 10.1016/j.ophtha.2010.07.020

International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret, G. B., Munroe, P. B., Rice, K. M., Bochud, M., Johnson, A. D., et al. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* 478, 103–109. doi: 10.1038/nature10405

Jones, C. C., Mercaldo, S. F., Blume, J. D., Wenzlaff, A. S., Schwartz, A. G., Chen, H., et al. (2018). Racial disparities in lung cancer survival: the contribution of stage, treatment, and ancestry. *J. Thorac. Oncol.* 13, 1464–1473. doi: 10.1016/j.jtho.2018.05.032

Jones, D. W., Appel, L. J., Sheps, S. G., Roccella, E. J., and Lenfant, C. (2003). Measuring blood pressure accurately: new and persistent challenges. *JAMA* 289, 1027–1030.

Kato, N., Takeuchi, F., Tabara, Y., Kelly, T. N., Go, M. J., and Sim, X. (2011). Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in east Asians. *Nat. Genet.* 43, 531–538. doi: 10.1038/ng.834

Kidambi, S., Ghosh, S., Kotchen, J. M., Grim, C. E., Krishnaswami, S., Kaldunski, M. L., et al. (2012). Non-replication study of a genome-wide association study for hypertension and blood pressure in African Americans. *BMC Med. Genet.* 13:27. doi: 10.1186/1471-2350-13-27

Kupper, N., Willemsen, G., Riese, H., Posthuma, D., Boomsma, D. I., and De Geus, E. J. (2005). Heritability of daytime ambulatory blood pressure in an extended twin design. *Hypertension* 45, 80–85. doi: 10.1161/01.hyp.0000149952.84391.54

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291. doi: 10.1038/nature19057

Levy, D., Destefano, A. L., Larson, M. G., O'Donnell, C. J., Lifton, R. P., Gavras, H., et al. (2000). Evidence for a gene influencing blood pressure on chromosome 17. Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the Framingham Heart Study. *Hypertension* 36, 477–483. doi: 10.1161/01.hyp.36.4.477

Levy, D., Ehret, G. B., Rice, K., Verwoert, G. C., Launer, L. J., and Dehghan, A. (2009). Genome-wide association study of blood pressure and hypertension. *Nat. Genet.* 41, 677–687. doi: 10.1038/ng.384

Liang, J., Le, T. H., Edwards, D. R. V., Tayo, B. O., Gaulton, K. J., and Smith, J. A. (2017). Single-trait and multi-trait genome-wide association analyses identify novel loci for blood pressure in African-ancestry populations. *PLoS Genet.* 13:e1006728. doi: 10.1371/journal.pgen.1006728

Liu, J. Z., Medland, S. E., Wright, M. J., Henders, A. K., Heath, A. C., Madden, P. A., et al. (2010). Genome-wide association study of height and body mass index in Australian twin families. *Twin Res. Hum. Genet.* 13, 179–193. doi: 10.1375/twin.13.2.179

Maples, B. K., Gravel, S., Kenny, E. E., and Bustamante, C. D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* 93, 278–288. doi: 10.1016/j.ajhg.2013.06.020

National Center for Health Statistics (2012). *Health, United States, 2011: With Special Feature on Socioeconomic Status and Health*. Hyattsville, MD: National Center for Health Statistics. Available at: https://www.cdc.gov/nchs/data/hus/hus11.pdf

Newton-Cheh, C., Johnson, T., Gateva, V., Tobin, M. D., Bochud, M., Coin, L., et al. (2009). Genome-wide association study identifies eight loci associated with blood pressure. *Nat. Genet.* 41, 666–676. doi: 10.1038/ng.361

Non, A. L., Gravlee, C. C., and Mulligan, C. J. (2012). Education, genetic ancestry, and blood pressure in African Americans and Whites. *Am. J. Public Health* 102, 1559–1565. doi: 10.2105/AJPH.2011.300448

Parra, E. J., Marcini, A., Akey, J., Martinson, J., Batzer, M. A., Cooper, R., et al. (1998). Estimating African American admixture proportions by use of population-specific alleles. *Am. J. Hum. Genet.* 63, 1839–1851. doi: 10.1086/302148

R Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Restrepo, N. A., Laper, S. M., Farber-Eger, E., and Crawford, D. C. (2018). Local genetic ancestry in CDKN2B-AS1 is associated with primary open-angle glaucoma in an African American cohort extracted from de-identified electronic health records. *BMC Med. Genomics* 11:70. doi: 10.1186/s12920-018-0392-4

Roden, D. M., Pulley, J. M., Basford, M. A., Bernard, G. R., Clayton, E. W., Balser, J. R., et al. (2008). Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin. Pharmacol. Ther.* 84, 362–369. doi: 10.1038/clpt.2008.89

Rotimi, C. N., Cooper, R. S., Cao, G., Ogunbiyi, O., Ladipo, M., Owoaje, E., et al. (1999). Maximum-likelihood generalized heritability estimate for blood pressure in Nigerian families. *Hypertension* 33, 874–878. doi: 10.1161/01.hyp.33.3.874

Russo, A., Di Gaetano, C., Cugliari, G., and Matullo, G. (2018). Advances in the genetics of hypertension: the effect of rare variants. *Int. J. Mol. Sci.* 19:E688. doi: 10.3390/ijms19030688

Seeman, T., Merkin, S. S., Crimmins, E., Koretz, B., Charette, S., and Karlamangla, A. (2008). Education, income and ethnic differences in cumulative biological risk profiles in a national sample of US adults: NHANES III (1988-1994). *Soc. Sci. Med.* 66, 72–87. doi: 10.1016/j.socscimed.2007.08.027

Shavers, V. L. (2007). Measurement of socioeconomic status in health disparities research. *J. Natl. Med. Assoc.* 99, 1013–1023.

Smith, J. A., Zhao, W., Yasutake, K., August, C., Ratliff, S. M., Faul, J. D., et al. (2017). Gene-by-psychosocial factor interactions influence diastolic blood pressure in European and African ancestry populations: meta-analysis of four cohort studies. *Int. J. Environ. Res. Public Health* 14:E1596. doi: 10.3390/ijerph14121596

Spada, J., Scholz, M., Kirsten, H., Hensch, T., Horn, K., Jawinski, P., et al. (2016). Genome-wide association analysis of actigraphic sleep phenotypes in the LIFE adult study. *J. Sleep Res.* 25, 690–701. doi: 10.1111/jsr.12421

Sung, Y. J., Winkler, T. W., De Las Fuentes, L., Bentley, A. R., Brown, M. R., Kraja, A. T., et al. (2018). A large-scale multi-ancestry genome-wide study accounting for smoking behavior identifies multiple significant loci for blood pressure. *Am. J. Hum. Genet.* 102, 375–400. doi: 10.1016/j.ajhg.2018.01.015

Tamborini, C. R., Kim, C., and Sakamoto, A. (2015). Education and lifetime earnings in the United States. *Demography* 52, 1383–1407. doi: 10.1007/s13524-015-0407-0

Taylor, J. Y., Schwander, K., Kardia, S. L., Arnett, D., Liang, J., Hunt, S. C., et al. (2016). A Genome-wide study of blood pressure in African Americans accounting for gene-smoking interaction. *Sci. Rep.* 6:18812. doi: 10.1038/srep18812

Taylor, J. Y., Sun, Y. V., Barcelona De Mendoza, V., Ifatunji, M., Rafferty, J., Fox, E. R., et al. (2017). The combined effects of genetic risk and perceived discrimination on blood pressure among African Americans in the Jackson Heart Study. *Medicine* 96:e8369. doi: 10.1097/MD.0000000000008369

Van Der Loos, M. J., Rietveld, C. A., Eklund, N., Koellinger, P. D., Rivadeneira, F., and Abecasis, G. R. (2013). The molecular genetic architecture of self-employment. *PLoS One* 8:e60542. doi: 10.1371/journal.pone.0060542

Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., and Chines, P. S. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* 8:e1002793. doi: 10.1371/journal.pgen.1002793

Wan, E. S., Cho, M. H., Boutaoui, N., Klanderman, B. J., Sylvia, J. S., and Ziniti, J. P. (2011). Genome-wide association analysis of body mass in chronic obstructive pulmonary disease. *Am. J. Respir. Cell Mol. Biol.* 45, 304–310. doi: 10.1165/rcmb.2010-0294OC

Wang, Y., O'Connell, J. R., Mcardle, P. F., Wade, J. B., Dorff, S. E., Shah, S. J., et al. (2009). From the cover: whole-genome association study identifies STK39 as a hypertension susceptibility gene. *Proc. Natl. Acad. Sci. U.S.A.* 106, 226–231. doi: 10.1073/pnas.0808358106

Wiley, L. K., Shah, A., Xu, H., and Bush, W. S. (2013). ICD-9 tobacco use codes are effective identifiers of smoking status. *J. Am. Med. Inform. Assoc.* 20, 652–658. doi: 10.1136/amiajnl-2012-001557

Winham, S. J., Cuellar-Barboza, A. B., Oliveros, A., Mcelroy, S. L., Crow, S., Colby, C., et al. (2014). Genome-wide association study of bipolar disorder accounting for effect of body mass index identifies a new risk allele in TCF7L2. *Mol. Psychiatry* 19, 1010–1016. doi: 10.1038/mp.2013.159

Wright, J. D., Hughes, J. P., Ostchega, Y., Yoon, S. S., Nwankwo, T. (2011). Mean systolic and diastolic blood pressure in adults aged 18 and over in the United States, 2001–2008. *Natl. Health Stat. Report*. 35, 1–22.

Writing Group Members, Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., and Blaha, M. J. (2016). Heart disease and stroke statistics-2016 update: a report from the American heart association. *Circulation* 133, e38–e360.

Yoon, S. S., Gu, Q., Nwankwo, T., Wright, J. D., Hong, Y., and Burt, V. (2015). Trends in blood pressure among adults with hypertension: United States, 2003 to 2012. *Hypertension* 65, 54–61. doi: 10.1161/HYPERTENSIONAHA.114.04012

Zhu, X., Feng, T., Tayo, B. O., Liang, J., Young, J. H., Franceschini, N., et al. (2015). Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am. J. Hum. Genet.* 96, 21–36. doi: 10.1016/j.ajhg.2014.11.011

Zhu, X., Young, J. H., Fox, E., Keating, B. J., Franceschini, N., and Kang, S. (2011). Combined admixture mapping and association analysis identifies a novel blood pressure genetic locus on 5p13: contributions from the CARe consortium. *Hum. Mol. Genet.* 20, 2285–2295. doi: 10.1093/hmg/ddr113

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling Editor declared a past co-authorship with one of the authors DC.

frontiers in Genetics

SYSTEMATIC REVIEW
published: 16 May 2019
doi: 10.3389/fgene.2019.00413

Systematic Review and Meta-Analysis to Establish the Association of Common Genetic Variations in Vitamin D Binding Protein With Chronic Obstructive Pulmonary Disease

*Ritesh Khanna, Debparna Nandy and Sabyasachi Senapati**

Department of Human Genetics and Molecular Medicine, Central University of Punjab, Bathinda, India

OPEN ACCESS

Edited by:
William Scott Bush, Case Western Reserve University, United States

Reviewed by:
Lijun Ma, Wake Forest University, United States
Renata Ferrari, Universidade Estadual Paulista (UNESP), Brazil
Suzana Erico Tanni, Universidade Estadual Paulista (UNESP), Brazil

****Correspondence:***
Sabyasachi Senapati
sabyasachi1012@gmail.com

Specialty section:
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics

Received: *04 June 2018*
Accepted: *16 April 2019*
Published: *16 May 2019*

Citation:
Khanna R, Nandy D and Senapati S (2019) Systematic Review and Meta-Analysis to Establish the Association of Common Genetic Variations in Vitamin D Binding Protein With Chronic Obstructive Pulmonary Disease. Front. Genet. 10:413. doi: 10.3389/fgene.2019.00413

Background: Vitamin-D binding protein (DBP) also known as GC protein, is a major determinant for vitamin- D metabolism and transport. GC1F, GC1S, and GC2 are the three allelic variants (denoted as rs4588 and rs7041) of GC, and known to be associated with chronic obstructive pulmonary disease (COPD). However, contradictory reports and population specific risk attributed by these alleles warranted detailed genetic epidemiology study to establish the association between GC variants and COPD. In this study we performed a meta-analysis and investigated the genetic architecture of GC locus to establish the association and uncover the plausible reason for allelic heterogeneity.

Methods: Published cross-sectional case control studies were screened and meta-analysis was performed between GC variants and COPD outcome. RevMan-v5.3 software was used to perform random and/or fixed models to calculate pooled odds ratio (Meta-OR). Linkage disequilibrium (LD) and haplotypes at GC locus were evaluated using 1000 Genomes genotype data. *In silico* functional implications of rs4588 and rs7041 was tested using publicly available tools.

Results: GC1F allele and GC1F/1F genotype were found to confer COPD risk in overall meta-analysis. GC1S/1S was found to confer risk only among Europeans. *In silico* investigation of rs4588 and rs7041 identified strong eQTL effects and potential role in regulation of GC expression. Large differences in allele frequencies, linkage disequilibrium (LD) and haplotypes were identified at GC locus across different populations (Japanese, African, Europeans, and Indians), which may explain the variable association of different GC alleles in different populations.

Conclusion: GC1F and GC1F/1F impose significant genetic risk for COPD, among Asians. Considerable differences in allele frequencies and LD structure in GC locus may impose population specific risk.

Keywords: vitamin D-binding protein, COPD, meta-analysis, linkage disequilibrium, genetic polymorphisms, allelic heterogeneity

INTRODUCTION

Chronic obstructive pulmonary disease (COPD) is a complex disease affecting the lung function. Genetically susceptible individuals develop the COPD while they get exposed to environmental triggers, such as noxious gases or suspended particles. Decreased level of vitamin-D in serum is associated with COPD among individuals with a history of smoking (Janssens et al., 2010). Besides environmental and genetic factors, metabolic factors are also critical and do cross talk with each other for the pathogenesis of COPD (Rabe et al., 2007).

Vitamin-D binding protein (DBP), also known as group-specific component (GC), belongs to a gene cluster family which is expressed in liver and other tissues (Chishimba et al., 2010). As the name suggests, it is known for its binding to circulating vitamin-D_3 and its transportation from liver to other tissues during its metabolism (Daiger et al., 1975). *GC* is a highly polymorphic gene and three of its allelic variants, namely *GC*1F, *GC*1S, and *GC*2, have been studied extensively for their association with vitamin-D deficiency (VDD) and other diseases including COPD (Chishimba et al., 2010; Wood et al., 2011). These variants correspond to different allelic arrangements of rs7041 and rs4588 (**Table 1**). These GC protein variants are reported to have a different affinity to bind to vitamin-D_3 i.e., 25(OH)D_3, and thus affect its serum concentration (Arnaud and Constans, 1993; Janssens et al., 2010). Circulating level of vitamin-D_3 is regulated by the synthesis and enzymatic degradation of 25(OH)D_3 by catabolizing enzymes. More than 90% of the circulating 25(OH)D_3 present in tightly bound ($K_d \sim 10^{-9}$ M) form with GC proteins (Arnaud and Constans, 1993). Therefore, different GC isoforms influence the serum concentration/bioavailability of 25(OH)D_3.

A study performed on north Indian cohort has shown homozygous *GC*1F variant to confer risk, and the disease severity is observed in a variant specific dose dependent manner, where the geometric mean of serum 25(OH)D_3 was observed in the ascending order of *GC* genotypes 1F/1F<1S/1F<1S/1S<1S/2<2/2 among COPD patients (Maheswari et al., 2014). Recent reports also indicate the protective role of *GC*2 variant among healthy individuals. Similar reports were published by the studies done among Caucasian and north Indian cohorts (Schellenberg et al., 1998; Berg et al., 2013; Maheswari et al., 2014; Chen et al., 2015). Azzawi et al. confirmed similar study outcomes in an Egyptian cohort where *GC*1F and *GC*1F/1S variants were found to be associated with low serum vitamin-D_3 concentration (Al-Azzawi et al., 2017). While studies done among Korean population have shown an association of *GC*2 variants with COPD progression, where *GC*2 and *GC*1F/1S variants were shown to be associated with higher emphysema index, irrespective of VDD. These studies also identified an association of *GC*2 and *GC*1F/1S variants with lower and higher serum concentration of vitamin-D_3, respectively. *GC2* showed significant association with VDD (Jung et al., 2014; Park et al., 2016).

GC protein (or DBP) is also involved in the inflammation by getting converted into MAF (Macrophage Activating Factor) in the presence of enzymes secreted by leucocytes. It has been found that the conversion of GC into GC-MAF is a deglycosylation process. Absence of glycosylated *Lys* residue at 420 in *GC*2 variants makes it an inappropriate reactant for the deglycosylation process, which makes them protective for COPD (Maheswari et al., 2014). Vitamin-D_3 is known to inhibit the expression of MMPs (Matrix Metalloproteinases), which are responsible for the emphysema degradation of lung alveoli. Thus, optimal serum concentration of Vitamin-D_3 is very critical among emphysema patients and a trial for such serum Vitamin D_3 intervention among a large participant group can further elucidate its role in COPD progression (Berg et al., 2013). Serum Vitamin-D_3 is also found to have seasonal and geographical variations, which depend on the amount of sunlight reaching the skin (Jung et al., 2014; Al-Azzawi et al., 2017). ECLIPSE Cohort study did not find an association between serum DBP and emphysema or lung function, although a negative correlation was found among DBP and serum 25(OH)D_3 level (Berg et al., 2013). While another study in an alpha1-antitrypsin deficient Caucasian population showed the association of serum DBP with COPD conditions (Wood et al., 2011). A recent report indicated a strong relationship between serum 25(OH)D_3 and pulmonary function (FEV1 and FVC) in a well-defined COPD cohort (Janssens et al., 2010).

It is evident that GC is a major determinant for several health parameters including those associated with COPD. However, contradictory findings of association of different alleles with COPD and non-replication across different populations warranted further meta-analysis and detailed population genetics studies. In the present study, we anticipated to explain the association of known *GC* alleles with COPD and investigate the genetic and functional aspects of *GC* alleles. Locus architecture of different populations was also investigated to explain the non-replication/differential replication of *GC* alleles in different populations. We hypothesized that genetics architecture at *GC* locus leads to population specific allelic variation in *GC* and its association with COPD. The study was performed with the following specific objectives: (i) perform meta-analysis to establish association of commonly studied *GC* alleles with COPD, and (ii) investigate the genetic heterogeneity at a functionally relevant *GC* locus, that explain variability in GC protein and COPD.

TABLE 1 | Allelic arrangements correspond to three different GC variants implicated in COPD.

GC variants	Allele of rs4588 (amino acid)	Allele of rs7041(amino acid)
GC-1F	C (Thr436)	T (Asp432)
GC-1S	C (Thr436)	A/G (Glu432)
GC-2	A (Lys436)	T (Asp432)

MATERIALS AND METHODS

Literature Retrieval

Our objective was to identify research articles where genetic association of *GC* has been tested with COPD. We restricted our study to three major genetic polymorphisms of *GC*, namely *GC*1F, *GC*1S, and *GC*2 alleles. Literature was searched online in the National Center for Biotechnology Information (NCBI-PubMed), Google Scholar and Medline. The major search language for the literature was English, papers in other languages were translated for further review. To obtain the best quality outcome, we include only peer reviewed scientific literature. Literature were searched until May 2018. The keywords used for the search for literature were as follows: Vitamin D binding protein and chronic obstructive pulmonary disease, DBP and COPD, *GC* alleles and COPD, COPD association *GC*. Cross references were also reviewed and references from the retrieved articles were also checked manually so as to find any relevant articles.

Inclusion and Exclusion Criteria

Only case-control studies were included for this meta-analysis. Only those studies were included where different alleles (1F, 1S, and 2) and genotypes (1F/1F, 1S/1S, 2/2, 1F/1S, 1F/2, 1S/2) of *GC* were studied for their association with COPD. Included studies clearly mentioned either the actual numbers, or the percentage of cases and controls with different genotypes and alleles of *GC*. Included studies have both smokers and non-smokers among both cases and controls.

Data Extraction

Data was extracted from eligible articles by two investigators independently and differences and controversies were resolved by group discussions. We first validated the study types and then extracted author names, year of publication, details of genotypes/alleles and their frequencies in COPD patients and controls.

Statistical Analysis

Results of association of three distinct alleles have been included in this study. These alleles were *GC*1F, *GC*1S and *GC*2, represented in NCBI dbSNP as rs4588 and rs7041, respectively (**Table 1**). Therefore, a total of six different genotypic combinations were studied, such as, *GC*1F/1F, *GC*1F/1S, *GC*1S/1S, *GC*1F/2, *GC*1S/2, and *GC*2/2. Independently these genotypes and three allelic associations were evaluated by meta-analysis. In each analysis, the experimental allele or genotypes were tested against the total allele or genotype counts. Meta-analysis was performed using Review Manager (RevMan-v5.3) Copenhagen: The Nordic Cochrane Center, The Cochrane Collaboration, 2014. Additive genetic model with 95% confidence interval (CI) was used in each of these independent analyses. Heterogeneity between studies was calculated by the I^2 and chi^2 test, where $I^2 > 50\%$ and chi^2 $p < 0.05$ was considered as significant heterogeneity. Meta-analysis of odds ratios were performed using a random effect model where significant heterogeneity was observed, otherwise a fixed effect model was used. Overall effect size (Meta-OR) was calculated by Z-test with 5% alpha level. A sensitivity analysis was performed to access whether meta-analysis results were substantially influenced by the presence of any study. This was done by systematically excluding one study at a time and recalculating the significance (p-value of the χ^2 and Z-test) of the results. The funnel plot was used to analyze the publication bias. Subgroup analysis between Asian and Caucasian studies was also performed to identify any significant differences due to individual group stratification.

Linkage Disequilibrium, Haplotypes, and Comparative Allele Frequency

Genetic architecture of *GC* locus was evaluated to explain population specific effects (if any) of *GC* alleles on its association with different human traits/diseases. To analyze the linkage disequilibrium, LD plots and haplotypes were reconstructed using Haploview (Barrett et al., 2004). LD calculations and manipulation of genotype files were done using Plink 1.07 (Purcell et al., 2007). 1000 Genomes genotype information for four major populations, such as CEU (Utah residents with northern and western European ancestry), GIH (Gujarati Indians in Houston, USA), YRI (Yoruba in Ibadan, Nigeria), and JPT (Japanese in Tokyo), were evaluated for LD analysis. Raw genotype data for these populations were obtained from 1,000 Genomes ftp through Ensembl. Genotype data were obtained for a 50 kb window on both the sides around rs7041 i.e., chr4:71702617-71802617 (GRCh38.p12). Comparative allele frequencies for *GC*1F, *GC*1S, and *GC*2 corresponding to rs4588 and rs7041 were evaluated from Ensembl (https://asia.ensembl.org/index.html), HaploReg (http://archive.broadinstitute.org/mammals/haploreg/haploreg.php).

In silico Functional Implication Assessment

Functional implications of rs4588 and rs7041 were analyzed using open source browsers. RegulomeDB (http://www.regulomedb.org/index) was used to analyze the regulatory function and GTEx portal (https://gtexportal.org/home/) was used to analyze single tissue or gene eQTL.

RESULTS

Characteristics of Eligible Studies

A total of 71 studies were identified initially after online literature search. After screening and proper reviewing for the eligible papers 48 papers were excluded. There were two duplicate studies, six studies were for asthma, and 11 were for diseases other than asthma and COPD, such as osteomalacia, type II diabetes, adenocarcinoma, pulmonary tuberculosis and other non-relevant diseases. There was a non-human study done on mice, which was also excluded from the meta-analysis. Meta-analysis ($n = 5$), which was done previously on COPD and *GC,* was also excluded but was used to identify cross references. Fourteen studies were excluded because they were either cohort studies or random clinical trials done on supplementation of vitamin-D_3. Eleven studies were found irrelevant, either due to less information for cases or control subjects, and one study was in other language and was excluded from the meta-analysis. A further seven studies were excluded as adequate/complete

genotype and study participant information were not given. After this screening based on our inclusion/exclusion criteria, a total of 14 studies were found eligible for meta-analysis (Kueppers et al., 1977; Home et al., 1990; Ishii et al., 2001; Ito et al., 2004; Laufs et al., 2004; Lu et al., 2004; Korytina et al., 2006; Huang et al., 2007; Janssens et al., 2010; Shen et al., 2010; Jung et al., 2014; Li et al., 2014; Maheswari et al., 2014; Al-Azzawi et al., 2017) (**Figure 1**).

Genotypic and Allelic Association

A total of 14 studies were included in this meta-analysis where genotypes for different above-mentioned *GC* alleles, in both COPD patients and healthy controls, were reported. Out of these 14 studies, nine studies were performed on different Asian populations and five were on European populations. Details of the study participants and haplotypes or allele frequencies are given in the **Supplementary Table 1**. Random effect model was performed to find out the pooled effect size for *GC*1F/1F, *GC*1F/S, *GC*1F/2, and *GC*2/2 genotypes, and *GC*1F, *GC*1S and *GC*2 alleles in COPD. For remaining analyses, fixed effect model was used due to insignificant study heterogeneity (chi^2 $p > 0.05$ and I^2 <50%) (**Figure 2** and **Supplementary Figure 1**). Meta-analysis was performed separately for reports on Asians and Europeans to identify significant differences in effect size, if any.

Allelic Association

*GC*1F allele has been found significantly predisposing for COPD outcome in combined analysis (Meta-OR = 1.29; 95% CI = 1.09–1.55; Z p-val = 0.004). Independently, *GC*1F allele has been found strongly associated among Asians (OR_{Asia} = 1.45; 95% CI = 1.24–1.68; Z p-val< 0.00001), but not among Europeans (OR_{Europe} = 1.02; 95% CI = 0.73–1.42; Z p-val = 0.92) (**Figures 2A,B**). Both *GC*1S and *GC*2 alleles were not found significant in conferring risk or protection with COPD outcome (**Supplementary Figure 1**). However, considering the trend of association, both these alleles were found protective in combined analyses.

Genotypic Association

Homozygous *GC*1F/1F was found significantly predisposing genotype with COPD outcome (Meta-OR = 1.61; 95% CI = 1.18–2.20; Z p-val = 0.002). Independent analysis found significant association of this genotype among Asians (OR_{Asia} = 1.93; 95% CI = 1.38–2.70; Z p-val = 0.0001), but it remains insignificant among Europeans (OR_{Europe} = 1.11 with 95% CI = 0.64–1.95; Z p-val = 0.71). Significant predisposition was observed for *GC*1S/1S genotype among Europeans (OR_{Europe} = 1.29; 95% CI = 1.00–1.68; Z p-val = 0.05), however it remains insignificant among Asians (**Figure 2**). Further, no significant associations were observed for any of the alleles or genotypes, either in combined or independent analyses in Asians and Europeans (**Supplementary Figure 1**).

Sensitivity Analysis and Publication Bias

Sensitivity analysis was performed for each study. No significant deviation in heterogeneity and study significance (p-value of the χ^2 and Z-test) was observed. Subgroup analyses did not identify any significant (p<0.05) subgroup stratification (**Figure 2** and **Supplementary Figures 1A–G**). Further, manual investigation of funnel plots did not identify any publication bias, where shapes of the funnel plots were symmetrical (**Supplementary Figure 2**).

Linkage Disequilibrium

Comparative LD analysis of GC locus showed substantial differences in the background LD structure between four reference populations. Comparatively similar LD structure was observed in CEU and GIH, however structure is further broken in JPT and YRI. Both the variations, rs4588 and rs7041, do not constitute any likely haplo-blocks in JPT and YRI (**Supplementary Figure 3**). Haplotypes for *GC*1F, *GC*1S, and *GC*2 were found to be present with relatively equal frequency among CEU (0.19, 0.57, and 0.24) and GIH (0.21, 0.46, and 0.32), however, these haplotypes were not found in JPT and YRI. Furthermore, moderate yet similar LD was observed between these two markers in CEU ($r^2 = 0.42$; D' = 1) and GIH ($r^2 = 0.41$; D' = 1), however, LD is completely broken in JPT (r^2 = 0.10; D' = 1) and YRI (r^2 = 0.00; D' = 0.53). Notable haplotypic variations were observed across the genomic region, whereas in JPT and YRI, these two variations are not in tight linkage with neighboring markers (**Supplementary Figure 3**). Allele frequencies of rs4588 and rs7041 and LD between them were seen to be very heterogeneous across 26 different populations, as documented in 1000 Genomes Project. Absolutely no LD (r^2 =0) was observed among different African populations, whereas the highest degree of LD was observed among Europeans and South Asian populations followed by Americans (**Supplementary Table 2**).

In silico Functional Implications

GC (ENSG00000145321) expresses in the liver despite very negligible expression in the pancreas and stomach. For two missense SNPs, rs4588, and rs7041, no evidence was observed for significant eQTL on *GC* in liver tissue, however, significant eQTL was observed in subcutaneous adipose (p = 6.55E-6), sun exposed skin (p = 1.67E-6), and stomach (p = 5.46E-9) tissues. SNP rs4588 was identified to alter motif-binding sites of transcription factors SP1 and SP3; and transcription factor binding element (KLF16). rs4588 and rs7041 were both identified: (a) to localize in DNase hypersensitivity regions in a common set of cell types and tissues, and (b) potentially alter histone modification in liver (strongly) and skin (quiescent/low) tissue.

DISCUSSION

In this systematic review, we performed meta-analysis and evaluated linkage disequilibrium at GC locus, in order to investigate the association of common *GC* polymorphisms with COPD. This meta-analysis established that *GC*1F allele and *GC*1F/1F genotype confers risk of COPD. However, association is majorly restricted to Asians and not in Europeans. On the contrary, *GC*1S/1S genotype was observed to confer risk to Europeans only (with borderline significance). At least one copy of *GC2* has been found to confer protection from COPD among both Asians and Europeans. Previous meta-analysis studies

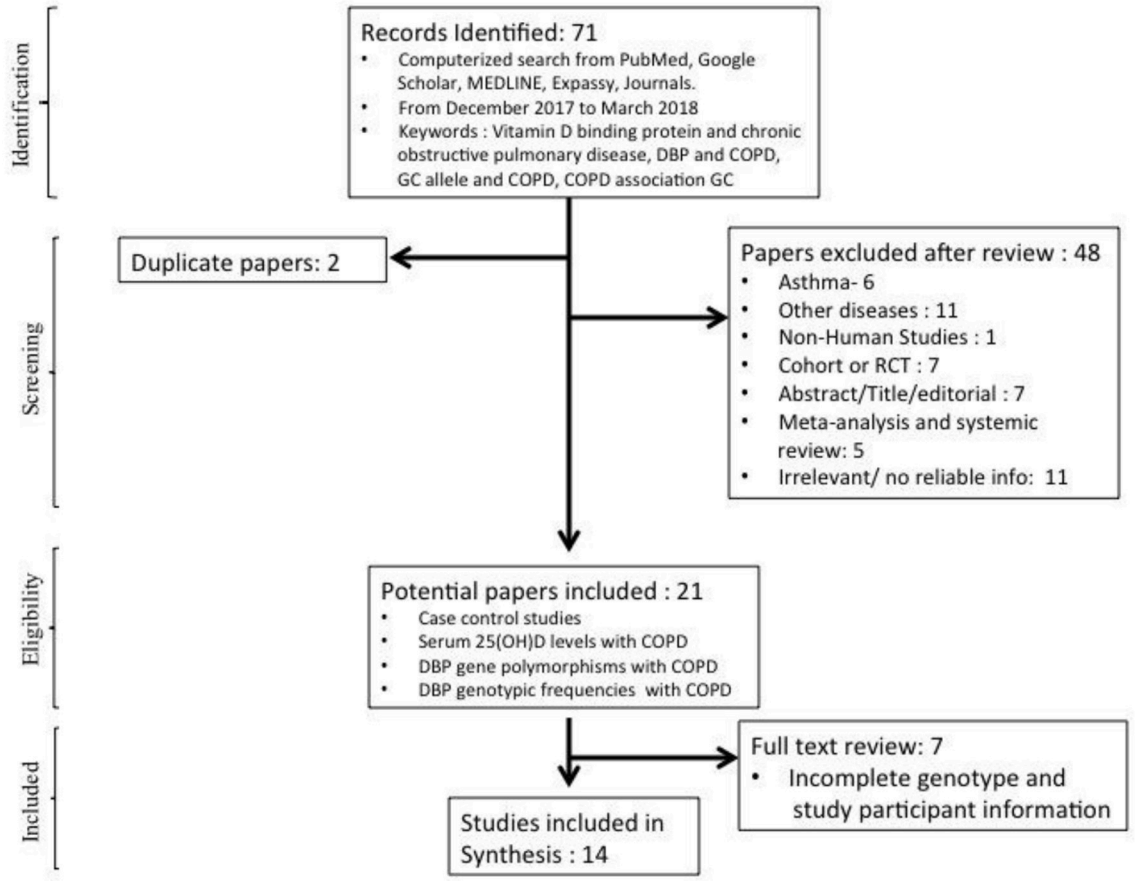

FIGURE 1 | Step-wise flow chart showing identification and screening of studies included in the meta-analysis.

and independent reports have shown different results from Europeans and Asians, which could be due to differences in allelic segregation and haplotypic heterogeneity at a population level (Chen et al., 2015; Horita et al., 2015; Wang et al., 2015; Xiao et al., 2015; Xie et al., 2015). Large differences in LD structure and haplotypes were observed in different ethnic populations, such as CEU GIH, JTP, and YRI. Although no reports are available (on association of *GC* variants and COPD) from African countries, we have included their representative genotypes for comparative genetic studies (**Supplementary Figure 3**). Notable differences in allele frequencies and LD between rs4588 and rs7041, among different populations, suggest significant population specific genetic contribution in *GC* variants (**Supplementary Table 2**). These major differences in LD between these two critical variants resulted into different haplotype frequencies and an absence of any quantifiable haplotypes in JPT and YRI. This indicates that perhaps different haplotypes are associated with different ethnic populations, which requires further large-scale genetic studies to uncover the novel alleles or haplotypes, if any. The overall trend shows relative similarity between CEU and GIH and distinct differences were observed in YRI and JPT. Different allelic arrangements of GC result into different GC variants, which vary in their isoelectric points and binding efficiency to vitamin D_3 (Braun et al., 1992; Arnaud and Constans, 1993; Speeckaert et al., 2006). Furthermore, in different populations, rs4588 and rs7041 may tag different sets of regulatory and structural SNPs (in haplotypes) across *GC*, and thus could play critical role in regulating expression and function of the GC protein.

Although VDD is found to be associated with COPD (Jolliffe et al., 2018), the underlying causes for such mechanisms remain unanswered. Recent GWAS studies on COPD were unable to identify GC or vitamin D receptor (VDR) as a significantly associated gene (Wain et al., 2017). However, genetic polymorphisms from these genes are found to be associated with VDD (Yousefzadeh et al., 2014; Zaki et al., 2017). In most of the studies, low level of serum Vitamin-D_3 is reported to be associated with the severity of COPD condition. Particularly, rs4588 has been shown to influence GC binding to Vitamin-D_3 (Nimitphong et al., 2013). It can be argued that, along with sufficient vitamin-D_3 intake/supplementation, a functionally more potent form of GC is necessary to maintain optimal serum bioavailability of vitamin-D_3. Therefore, inter individual differences in the GC protein may act as a predisposing factor for COPD. Further genetic epidemiological studies are warranted to identify novel risk alleles from *GC* that are associated with *GC*

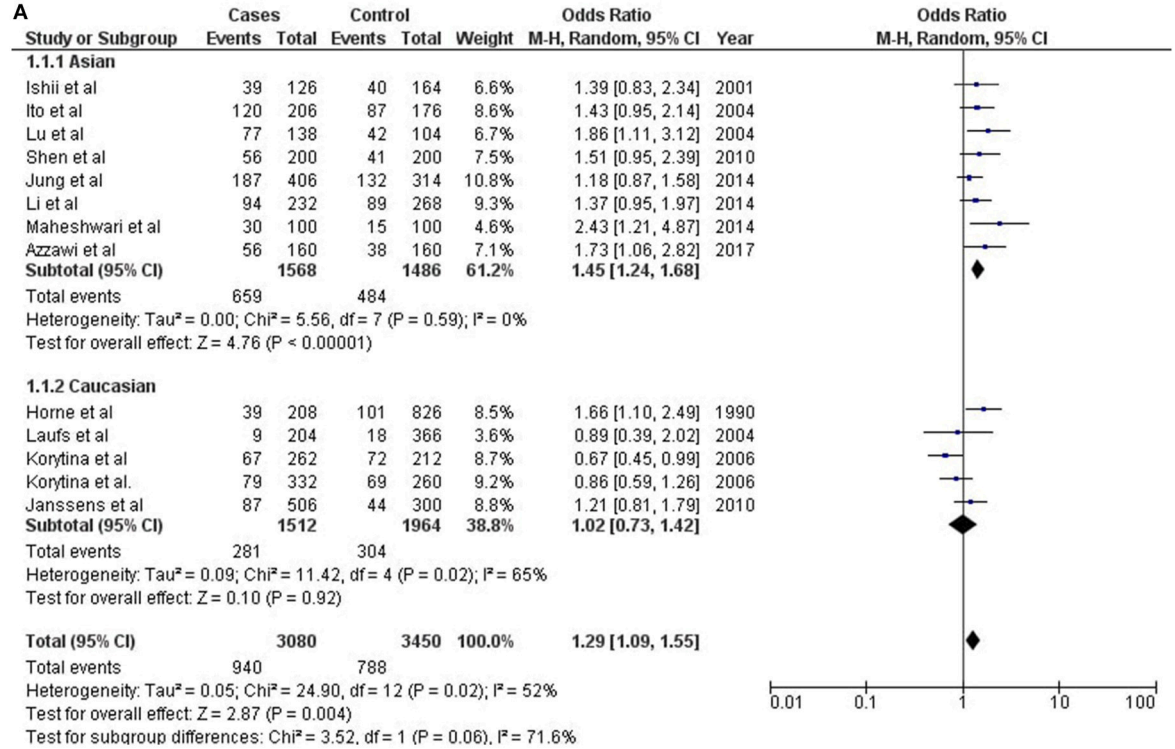

FIGURE 2 | Assessment of risk for meta-analysis of **(A)** *GC1F* allele, and **(B)** GC1F/1F genotype with COPD.

function, and thus implication in COPD. However, the presence of differential LD structure of GC locus needs to be considered as a major confounding factor.

AUTHOR CONTRIBUTIONS

SS conceptualized and designed the study. RK and DN performed literature screening and meta-analysis. SS performed *in silico* genetic study. RK, DN, and SS contributed in writing the manuscript and interpreted the results. All the authors reviewed the manuscript and finalized for submission.

FUNDING

We acknowledge financial supports from DST-SERB (#ECR/2016/001660), UGC-BSR grant (30-4/2014-BSR), and research grant from Central University of Punjab (GP.25).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00413/full#supplementary-material

REFERENCES

Al-Azzawi, M. A., Ghoneim, A. H., and Elmadbouh, I. (2017). Evaluation of vitamin D, vitamin D binding protein gene polymorphism with oxidant-antioxidant profiles in chronic obstructive pulmonary disease. *J. Med. Biochem.* 36, 331–340. doi: 10.1515/jomb-2017-0012

Arnaud, J., and Constans, J. (1993). Affinity differences for vitamin D metabolites associated with the genetic isoforms of the human serum carrier protein (DBP). *Hum. Genet.* 92, 183–188. doi: 10.1007/BF00219689

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2004). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263–265 doi: 10.1093/bioinformatics/bth457

Berg, I., Hanson, C., Sayles, H., Romberger, D., Nelson, A., Meza, J., et al. (2013). Vitamin D binding protein, lung function and structure in COPD. *Respir. Med.* 107, 1578–1588. doi: 10.1016/j.rmed.2013.05.010

Braun, A., Bichlmaier, R., and Cleve, H. (1992). Molecular analysis of the gene for the human vitamin-D-binding protein (group-specific component): allelic differences of the common genetic GC types. *Hum. Genet.* 89, 401–406. doi: 10.1007/BF00194311

Chen, H., Zhang, L., He, Z., Zhong, X., Zhang, J., Li, M., et al. (2015). Vitamin D binding protein gene polymorphisms and chronic obstructive pulmonary disease: a meta-analysis. *J. Thoracic Dis.* 7:1423–1440. doi: 10.3978/j.issn.2072-1439.2015.08.16.

Chishimba, L., Thickett, D. R., Stockley, R. A., and Wood, A. M. (2010). The vitamin D axis in the lung: a key role for vitamin D-binding protein. *Thorax* 65, 456–462. doi: 10.1136/thx.2009.128793

Daiger, S. P., Schanfield, M. S., and Cavalli-Sforza, L. L. (1975). Group-specific component (GC) proteins bind vitamin D and 25-hydroxyvitamin D. *Proc. Nat. Acad. Sci. U.S.A.* 72, 2076–2080.

Home, S. L., Cockcroft, D. W., and Dosman, J. A. (1990). Possible protective effect against chronic obstructive airways disease by the GC 2 allele. *Hum. Hered.* 40, 173–176. doi: 10.1159/000153926

Horita, N., Miyazawa, N., Tomaru, K., Inoue, M., Ishigatsubo, Y., and Kaneko, T. (2015). Vitamin D binding protein genotype variants and risk of chronic obstructive pulmonary disease: a meta-analysis. *Respirology* 20, 219–225. doi: 10.1111/resp.12448

Huang, P., Ma, Y., Du, X., Chen, J., Song, B., Hong, Y., et al. (2007). The vitamin D-binding protein gene polymorphism in chronic obstructive pulmonary disease patients. *Zhonghua Jie He He Hu Xi ZaZhi.* 30, 780–781.

Ishii, T., Keicho, N., Teramoto, S., Azuma, A., Kudoh, S., Fukuchi, Y., et al. (2001). Association of Gc-globulin variation with susceptibility to COPD and diffuse panbronchiolitis. *Eur. Respir. J.* 18, 753–757. doi: 10.1183/09031936.01.00094401

Ito, I., Nagai, S., Hoshino, Y., Muro, S., Hirai, T., Tsukino, M., et al. (2004). Risk and severity of COPD is associated with the group-specific component of serum globulin 1F allele. *Chest* 125, 63–70. doi: 10.1378/chest.125.1.63

Janssens, W., Bouillon, R., Claes, B., Carremans, C., Lehouck, A., Buysschaert, I., et al. (2010). Vitamin D deficiency is highly prevalent in COPD and correlates with variants in the vitamin D-binding gene. *Thorax* 65, 215–220. doi: 10.1136/thx.2009.120659

Jolliffe, D. A., James, W. Y., Hooper, R. L., Barnes, N. C., Greiller, C. L., Islam, K., et al. (2018). Prevalence, determinants and clinical correlates of vitamin D deficiency in patients with Chronic Obstructive Pulmonary Disease in London, U. K. *J. Steroid Biochem. Mol. Biol.* 175, 138–145. doi: 10.1016/j.jsbmb.2017.01.019

Jung, J. Y., Choi, D. P., Won, S., Lee, Y., Shin, J. H., Kim, Y. S., et al. (2014). Relationship of vitamin D binding protein polymorphisms and lung function in Korean chronic obstructive pulmonary disease. *Yonsei Med. J.* 55, 1318–1325. doi: 10.3349/ymj.2014.55.5.1318

Korytina, G. F., Akhmadishina, L. Z., Ianbaeva, D. G., and Viktorova, T. V. (2006). Genotypes of vitamin-D-binding protein (DBP) in patients with chronic obstructive pulmonary disease and healthy population of Republic Bashkortostan. *Mol. Biol.* 40, 231–238. doi: 10.1134/S002689330602004X

Kueppers, F., Miller, R. D., Gordon, H., Hepper, N. G., and Offord, K. (1977). Familial prevalence of chronic obstructive pulmonary disease in a matched pair study. *Am. J. Med.* 63, 336–342. doi: 10.1016/0002-9343(77)90270-4

Laufs, J., Andrason, H., Sigvaldason, A., Halapi, E., Thorsteinsson, L., Jónasson, K., et al. (2004). Association of vitamin D binding protein variants with chronic mucus hypersecretion in Iceland. *Am. J. Pharmacogenom.* 4, 63–68. doi: 10.2165/00129785-200404010-00007

Li, X., Liu, X., Xu, Y., Xiong, W., Zhao, J., Ni, W., et al. (2014). The correlation of vitamin D level and vitamin D-binding protein gene polymorphism in chronic obstructive pulmonary disease. *Zhonghuaneikezazhi* 53, 303–307.

Lu, M., Yang, B., and Cai, Y. Y. (2004). The relationship between vitamin D binding protein gene polymorphism and chronic obstructive pulmonary disease. *ZhonghuaNeiKeZaZhi* 43, 117–120.

Maheswari, K., Choudhary, M., Javid, S. (2014). Association of Vitamin D Binding protein gene polymorphism with serum 25-hydroxy Vitamin D levels in COPD. *Online Int. Interdiscipl. Res. J.* 4, 46–55.

Nimitphong, H., Saetung, S., Chanprasertyotin, S., Chailurkit, L. O., and Ongphiphadhanakul, B. (2013). Changes in circulating 25-hydroxyvitamin D according to vitamin D binding protein genotypes after vitamin D3or vitamin D2 supplementation. *Nutr. J.* 4:1239. doi: 10.1186/1475-2891-12-39

Park, Y., Kim, Y. S., Kang, Y. A., Shin, J. H., Oh, Y. M., Seo, J. B., et al. (2016). Relationship between vitamin D-binding protein polymorphisms and blood vitamin D level in Korean patients with COPD. *Int. J. Chron. Obstruct. Pulmon Dis.* 11, 731–738. doi: 10.2147/COPD.S96985

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Rabe, K. F., Hurd, S., Anzueto, A., Barnes, P. J., Buist, S. A., Calverley, P., et al. (2007). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *Am J RespirCrit Care Med.* 176, 532–555. doi: 10.1164/rccm.200703-456SO

Schellenberg, D., Paré, P. D., and Weir, T. D., Spinelli, J. J., Walker, B. A., and Sandford, A. J. (1998). Vitamin D binding protein variants and the risk of COPD. *Am. J. Respir. Crit. Care Med.* 157, 957–961. doi: 10.1164/ajrccm.157.3.9706106

Shen, L. H., Zhang, X. M., Su, D. J., Yao, S. P., Yu, B. Q., Wang, H. W., et al. (2010). Association of vitamin D binding protein variants with susceptibility to chronic obstructive pulmonary disease. *J. Int. Med. Res.* 38, 1093–1098. doi: 10.1177/147323001003800337

Speeckaert, M., Huang, G., Delanghe, J. R., and Taes, Y. E. (2006). Biological and clinical aspects of the vitamin D binding protein)Gc-globulin) and its polymorphism. *Clin. Chim. Acta*. 372, 33–42. doi: 10.1016/j.cca.2006.03.011

Wain, L. V., Shrine, N., Artigas, M. S., Erzurumluoglu, A. M., Noyvert, B., Bossini-Castillo, L., et al. (2017). Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. *Nat. Genet*. 49, 416. doi: 10.1038/ng.3787

Wang, Y. L., Kong, H., Xie, W. P., and Wang, H. (2015). Association of vitamin D-binding protein variants with chronic obstructive pulmonary disease: a meta-analysis. *Genet. Mol. Res.* 14, 10774–10785. doi: 10.4238/2015.September.9.16

Wood, A. M., Bassford, C., Webster, D., Newby, P., Rajesh, P., Stockley, R. A., et al. (2011). Vitamin D-binding protein contributes to COPD by activation of alveolar macrophages. *Thorax* 66, 205–210. doi: 10.1136/thx.2010.140921

Xiao, M., Wang, T., Zhu, T., and Wen, F. (2015). Dual role of vitamin D-binding protein 1F allele in chronic obstructive pulmonary disease susceptibility: a meta-analysis. *Genet. Mol. Res*. 14, 3534–3540. doi: 10.4238/2015.April.17.1

Xie, X., Zhang, Y., Ke, R., Wang, G., Wang, S., Hussain, T., et al. (2015). Vitamin D-binding protein gene polymorphisms and chronic obstructive pulmonary disease susceptibility: a meta-analysis. *Biomed. Rep.* 3, 183–188. doi: 10.3892/br.2014.392

Yousefzadeh, P., Shapses, S. A., and Wang, X. (2014). Vitamin D binding protein impact on 25-hydroxyvitamin D levels under different physiologic and pathologic conditions. *Int. J. Endocrinol.* 2014:981581. doi: 10.1155/2014/981581

Zaki, M., Kamal, S., Basha, W. A., Youness, E., Ezzat, W., El-Bassyouni, H., et al. (2017). Association of vitamin D receptor gene polymorphism (VDR) with vitamin D deficiency, metabolic and inflammatory markers in Egyptian obese women. *Genes Dis.* 4, 176–182. doi: 10.1016/j.gendis.2017.07.002

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

REVIEW
published: 14 June 2019
doi: 10.3389/fgene.2019.00548

A Review of African Americans' Beliefs and Attitudes About Genomic Studies: Opportunities for Message Design

Courtney L. Scherr[1], Sanjana Ramesh[1], Charlotte Marshall-Fricker[1] and Minoli A. Perera[2]*

[1] Department of Communication Studies, Center for Communication and Health, Northwestern University, Chicago, IL, United States, [2] Department of Pharmacology, Center for Pharmacogenomics, Feinberg School of Medicine, Chicago, IL, United States

Precision Medicine, the practice of targeting prevention and therapies according to an individual's lifestyle, environment or genetics, holds promise to improve population health outcomes. Within precision medicine, pharmacogenomics (PGX) uses an individual's genome to determine drug response and dosing to tailor therapy. Most PGX studies have been conducted in European populations, but African Americans have greater genetic variation when compared with most populations. Failure to include African Americans in PGX studies may lead to increased health disparities. PGX studies focused on patients of African American descent are needed to identify relevant population specific genetic predictors of drug responses. Recruitment is one barrier to African American participation in PGX. Addressing recruitment challenges is a significant, yet potentially low-cost solution to improve patient accrual and retention. Limited literature exists about African American participation in PGX research, but studies have explored barriers and facilitators among African American participation in genomic studies more broadly. This paper synthesizes the existing literature and extrapolates these findings to PGX studies, with a particular focus on opportunities for message design. Findings from this review can provide guidance for future PGX study recruitment.

Keywords: African American, genomics, health communication, pharmacogenomics, precision medicine, recruitment

OPEN ACCESS

Edited by:
Jessica Nicole Cooke Bailey, Case Western Reserve University, United States

Reviewed by:
Satyanarayana M. R. Rao, Jawaharlal Nehru Centre for Advanced Scientific Research, India
Suzette J. Bielinski, Mayo Clinic, United States

****Correspondence:***
Courtney L. Scherr
courtney.scherr@northwestern.edu

Specialty section:
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics

Received: *02 October 2018*
Accepted: *24 May 2019*
Published: *14 June 2019*

Citation:
Scherr CL, Ramesh S, Marshall-Fricker C and Perera MA (2019) A Review of African Americans' Beliefs and Attitudes About Genomic Studies: Opportunities for Message Design. Front. Genet. 10:548. doi: 10.3389/fgene.2019.00548

INTRODUCTION

Precision Medicine (PM) refers to the targeting of therapies according to an individual's, genetics, lifestyle or environment and holds immense promise to improve population health outcomes (Khoury et al., 2016). A branch of precision medicine, pharmacogenomics (PGX) is the study of genetic information to determine individual response (e.g., efficacy/toxicity) to pharmaceutical agents with the goal of developing safe and effective medications and dosage that can be tailored based on an individual's genetics (Lee, 2003; Empey, 2016). In order to draw conclusions about gene interactions and genetic variation within and across ancestries, substantial and diverse patient data are needed (Jaffe, 2015; Khoury et al., 2016). To date, most PGX participants are of European ancestry (Perera et al., 2014). However, African Americans have greater genetic variation than European populations, therefore, results from existing PGX studies may not be as predictive in

African American populations (Johnson et al., 2011; Perera et al., 2014). Under-representation of African American populations impairs the ability to translate PGX findings into clinical care, and will ultimately result in increased health disparities (Perera et al., 2014).

The challenge of recruiting minority populations likely stems from historic and contemporary mistreatment. For example, the Tuskegee Syphilis study has had a lingering effect on African Americans trust of medical institutions and research (Gamble, 1997). In addition to historic mistrust related to clinical research more broadly, genomic studies are further problematized due to concerns about personal identification, disenfranchisement stemming from genomic-based policies, and the potential threat of eugenics (Jackson, 1999). Furthermore, concerns about the inability for genomic research to address issues of social justice, and potentially exacerbate issues of health disparities remain (Jackson, 1999). Although few studies have examined the recruitment of African Americans to PGX studies, several have reported African American recruitment for genetic studies or biobanks (which we hereinafter refer to as genomic studies for simplicity).

Prior studies have reported demographic differences, for example, that African Americans are less likely to participate in research that includes a DNA sample or a biopsy compared with whites (Dye et al., 2016; Moledina et al., 2018). However, other studies have reported conflicting findings related to demographic factors influencing participation. One study related to prostate cancer genomics compared African American participants with white participants and found African American participants were younger, less educated, lower income, and less likely to be married compared with white participants (Patel et al., 2012). However, a different study found that African American women who provided a saliva sample for genomic research were older, regularly took a multivitamin, had a physician visit in the previous year, and reported a history of breast colorectal, or cervical screening compared with African American women who did not provide a saliva sample (Adams-Campbell et al., 2016). While demographic differences are useful in the categorization of participants, they do not provide useful insight for recruitment efforts.

Literature on recruitment efforts often describe community-based approaches (CBA) to engage participants in genomic studies by emphasizing intentional and meaningful community member engagement throughout the research process (Israel et al., 1998; Vadaparampil and Pal, 2010; Kiviniemi et al., 2013; Ochs-Balcom et al., 2015; McNeill et al., 2018). However, CBA focus on broad methods for recruitment and less on message content. Existing studies also have reported on the use of educational materials and seminars to improve African American recruitment (Skinner et al., 2008; Halverson and Ross, 2012; Rodriguez et al., 2016; Radecki Breitkopf et al., 2018). Studies found pre-post increases in knowledge about genomic studies, more favorable attitudes (Patel et al., 2018) and less negative affect (Kiviniemi et al., 2013) after receiving an educational intervention. However, random control trials and other studies employing pre-post assessment found no changes in attitudes about genomic research because of educational interventions (Skinner et al., 2008; Halverson and Ross, 2012). Such findings are not surprising because attitudes do not correlate with knowledge, but are shaped by values and beliefs (Grimshaw et al., 2002; Marteau et al., 2002; Fishbein and Yzer, 2003). Therefore, recruitment messages which address beliefs and attitudes related to participation in PGX studies, in addition to providing education, may speak more directly to African Americans' concerns, and may more consistently improve recruitment efforts (Scherr et al., 2017).

Existing literature regarding African Americans' beliefs and attitudes about genomic studies is disparate, and sometimes conflicting. Aggregating existing information provides an opportunity to reflect on current findings and potentially guide recruitment message strategies. Therefore, the objective of this paper is to systematically review qualitative and quantitative literature on African Americans' beliefs and attitudes about genomic studies that may influence their decision to participate. We synthesized results from this review to highlight opportunities for the design of genomic study recruitment messages.

MATERIALS AND METHODS

Study Design

Studies that provided insight regarding African Americans' beliefs and attitudes toward participation in biobanks or genomic studies (inclusive of genetic or PGX) were included in this review. We focused on biobank and genomic studies because, to the best of our knowledge, no studies have exclusively explored African Americans' beliefs and attitudes about PGX. Qualitative and quantitative studies with original empirical data were included, but conference abstracts, reviews, commentaries, editorials, legal opinions, letters to the editors, case studies, dissertations, and thesis studies were excluded. Given the potential influence of historical context, we excluded studies conducted outside the United States. We were interested in genetic studies that may be able to provide information on the treatment of chronic adult onset conditions; therefore, we excluded studies related to behavioral, developmental, or mental health genomics because we believed contextual factors (e.g., stigma, environment) could impact the results of such studies. We also excluded studies that explored medical professionals' attitudes or beliefs about genomic studies because, while valuable, their attitudes and beliefs may be influenced by their additional education and training. We excluded studies that included less than 13% African Americans as a proportion of the total sample, which is consistent with the proportion of African Americans in the United States population. Finally, we excluded studies in which we could not distinguish African Americans' responses from the responses of other study participants. Genomic studies have been conducted over a relatively limited period; therefore, we included all studies accepted for publication up to July 25, 2018 in this review.

Information Sources and Search

A study team member worked with a University librarian and searched PubMed, Scopus, Web of Science, Embase, and

Google Scholar for relevant citations. The search string was as follows: "African American" OR Black AND "genetic research" OR "pharmacogenomics research" OR "genomic research" OR "personalized medicine" OR "precision medicine" AND "study recruitment" OR "research participation." The initial search returned 1,179 total citations: 15 from PubMed, 14 from Scopus, 133 from Web of Science, 26 from Embase, and 990 from Google Scholar. After consolidating the lists, we removed 109 duplicate citations, for a final sample of 1,070 citations.

Study Selection

We screened studies for eligibility by conducting a review of the study titles, followed by an abstract review, and finally a full text review. Reviewers were instructed to be conservative in their exclusion; when uncertain, the study was retained. One study team member conducted the review of titles and excluded those that did not meet eligibility criteria. A second team member reviewed 20% of the titles to confirm exclusion criteria reliability. Kripendorf's $\alpha = 0.73$ was achieved, an acceptable level of reliability (Krippendorff, 2004). Next, two study team members split the remaining abstracts evenly for review, and excluded those which did not meet eligibility criteria. Twenty percent of the abstracts overlapped for reliability calculation, and $\alpha = 0.86$ was achieved. Finally, one study team member reviewed 92% and another study team member reviewed 28% of full text and excluded those that did not meet eligibility criteria. Twenty percent of the full text overlapped to calculate reliability, and $\alpha = 0.85$ was achieved.

Data Analysis

One study team member reviewed the final studies included in the analysis to extrapolate information including the study design, the population setting, the total sample size, the sample race, and age. Two study team members conducted thematic analysis of the articles using MAXQDA to manage the data (VERBI Software, 2018).

RESULTS

Of the 1,070 total titles screened, we removed 292 based on the title review, 558 based on the abstract review, and 197 based on the full text review, for a final sample of 24 articles (see **Figure 1**).

Review of Studies

Our review of the literature (**Table 1**), identified tensions in African Americans' beliefs and attitudes about genomic research. The overarching theme of trust (or lack thereof) was present across studies, and influenced subsequent attitudes about genomic research and participation. However, even with concerns about trust, African Americans believed their participation in genomic studies was critical. These negative and positive beliefs informed their attitudes about participation in genomic studies. What follows is a summary of the literature highlighting tensions between distrust and the value of their participation.

Distrust

We found a shadow of historic and continued injustice cast across studies. Distrust was ubiquitous in all facets of the research enterprise and extended from members of the research and medical communities (Skinner et al., 2015; Drake et al., 2017; Kraft et al., 2018), to medical or research institutions (Drake et al., 2017; Kraft et al., 2018), and the conduct of research and science in general (Skinner et al., 2015). The Tuskegee Study of Untreated Syphilis frequently functioned as a historical referent for the distrust of biomedical research, particularly among African Americans (Hoyo et al., 2003; Bates and Harris, 2004; Cohn et al., 2015; Kraft et al., 2018). One study found African Americans were significantly more concerned that something like Tuskegee could happen again than white participants (Hagiwara et al., 2014). More specific to genetics, revelations about Henrietta Lacks, and more recent and local race-related abuses by researchers, raised concerns about trust, privacy and the benefits of genomic studies (Buseh et al., 2013; Drake et al., 2017; Kraft et al., 2018; Lee et al., 2019). The impact of race-related injustice was apparent in two multi-race studies that found distrust was more salient among African American participants compared with their white counterparts (Bussey-Jones et al., 2010; Hagiwara et al., 2014). The salience of race in historic injustices in the United States raised suspicions about researchers' intentions, and the potential for race-based research to be used for maleficence ranging from racial discrimination to eugenics, or even genocide (Buseh et al., 2013; Isler et al., 2013; Kraft et al., 2018).

Distrust often was tied to fears about study processes and outcomes. Most frequently mentioned were fears of being experimented on or treated as a "guinea pig" or "lab rat" (Hoyo et al., 2003; Ochs-Balcom et al., 2011; Luque et al., 2012; Buseh et al., 2013; Erwin et al., 2013; Hagiwara et al., 2014; Walker et al., 2014), as was fear of exploitation (McDonald et al., 2012; Buseh et al., 2013). Several studies revealed beliefs that research is conducted at the expense of African Americans for the financial profit of those in power (Kraft et al., 2018; Lee et al., 2019), or to provide more effective treatments to white or privileged individuals (Luque et al., 2012; Halbert et al., 2016). Both African American and white participants in one study raised concerns about the possibility that genetic research could be used to discriminate against certain groups of people, with significantly more African Americans reporting that their concern about potential discrimination would influence their willingness to provide a blood sample for research (Goldenberg et al., 2011). Personal experiences with racial discrimination, and witnessing expanding health disparities in spite of medical advancements, added to beliefs that the medical and research communities were not trustworthy (Buseh et al., 2013). Among African Americans, increased distrust was significantly associated with reduced likelihood of biobank participation (McDonald et al., 2014; Halbert et al., 2016).

Despite concerns about trust and associated fears about participation, participants' relationship with medical research was complicated (McDonald et al., 2012). Tensions existed

TABLE 1 | Studies included in Review.

References	Study design	Setting and population	Sample size	n(%) AA*
Bates and Harris, 2004	Qualitative focus group study	Southeastern United States General population	*N* = 215	118(55)
Brewer et al., 2014	Quantitative cross-sectional survey study	Orlando, Florida at The Links, Incorporated 38th National Assembly Female Links Members	*N* = 381	381(100)
Buseh et al., 2013	Mixed methods; CBPR and focus group study	Wisconsin Genomic Initiative Community members	*N* = 21	21(100)
Bussey-Jones et al., 2010	Mixed methods telephone survey	North Carolina North Carolina Colorectal Cancer Study Database	*N* = 801	153(19%)
Cain et al., 2016	Quantitative cross-sectional survey study	Washington DC Metro Area Community members	*N* = 304	304(100)
Cohn et al., 2015	Qualitative exploratory study	Central Harlem, New York Community members	*N* = 46	39(89) 4(9) AA/ Hispanic 1(2) AA/ Native American
Dash et al., 2014	Mixed methods; focus group and cross-sectional survey study	Southeast/Southwest Washington, DC Community members	Focus groups (*n* = 41) Surveys (*n* = 321)	Focus groups 41(100) Surveys 234(73)
Diaz et al., 2008	Quantitative cross-sectional survey study	South Carolina State University Students	*N* = 200	200(100)
Drake et al., 2017	Qualitative focus group study	St. Louis, Missouri; Prostate Cancer Community Partnership Men with prostate cancer	*N* = 70	70(100)
Erwin et al., 2013	Mixed methods; focus group and cross-sectional survey study	Niagara Falls, New York Community members and Key informants	Key informant interviews (*n* = 9) Community focus groups (*n* = 21) Staff focus group (*n* = 5) Surveys (*n* = 64)	Community focus groups 13(62) Surveys 34(53)
Goldenberg et al., 2011	Mixed methods; computer assisted telephone interviewing system	Patients from Duke University, Johns Hopkins, University of Arizona, University of North Carolina, University of Utah	*N* = 1,193	192(16)
Hagiwara et al., 2014	Quantitative survey study	Detroit, Michigan Community members	*N* = 78	78(100)
Halbert et al., 2016	Quantitative survey study using vignettes	National sample of AA	*N* = 510	510(100)
Hoyo et al., 2003	Qualitative semi-structured interview and focus group study	North Carolina Community members	Focus groups (*n* = 46) Interviews (*n* = 9)	55(100)
Isler et al., 2013	Qualitative semi-structured interview study	North Carolina Community members	*N* = 91	72(79)
Jones et al., 2017	Quantitative cross-sectional survey study	Kansas City, Kansas Community members	*N* = 169	169(100)
Kraft et al., 2018	Qualitative focus group study using trigger videos	Northern California Patients at a large multispecialty practice	*N* = 122	23(18.9)
Lee et al., 2019	Qualitative focus group study using trigger videos	Northern California Patients at a large multispecialty practice	*N* = 122	23(18.9)
Luque et al., 2012	Qualitative focus group study	Tampa, Florida Community members	*N* = 95	33(34.7)
McDonald et al., 2014	Quantitative survey study	National sample of AA	*N* = 1,033	1,033(100)
McDonald et al., 2012	Qualitative focus group study	Philadelphia, Pennsylvania	*N* = 91	91(100)
Ochs-Balcom et al., 2011	Qualitative focus group study	Buffalo, New York Female breast cancer survivors	*N* = 14	14(100)
Skinner et al., 2015	Qualitative focus group study	Lenoir County, North Carolina Community members	*N* = 25	19(76)
Walker et al., 2014	Qualitative focus group study	Jackson, Mississippi Community members	*N* = 140	140(100)

**AA, African American.*

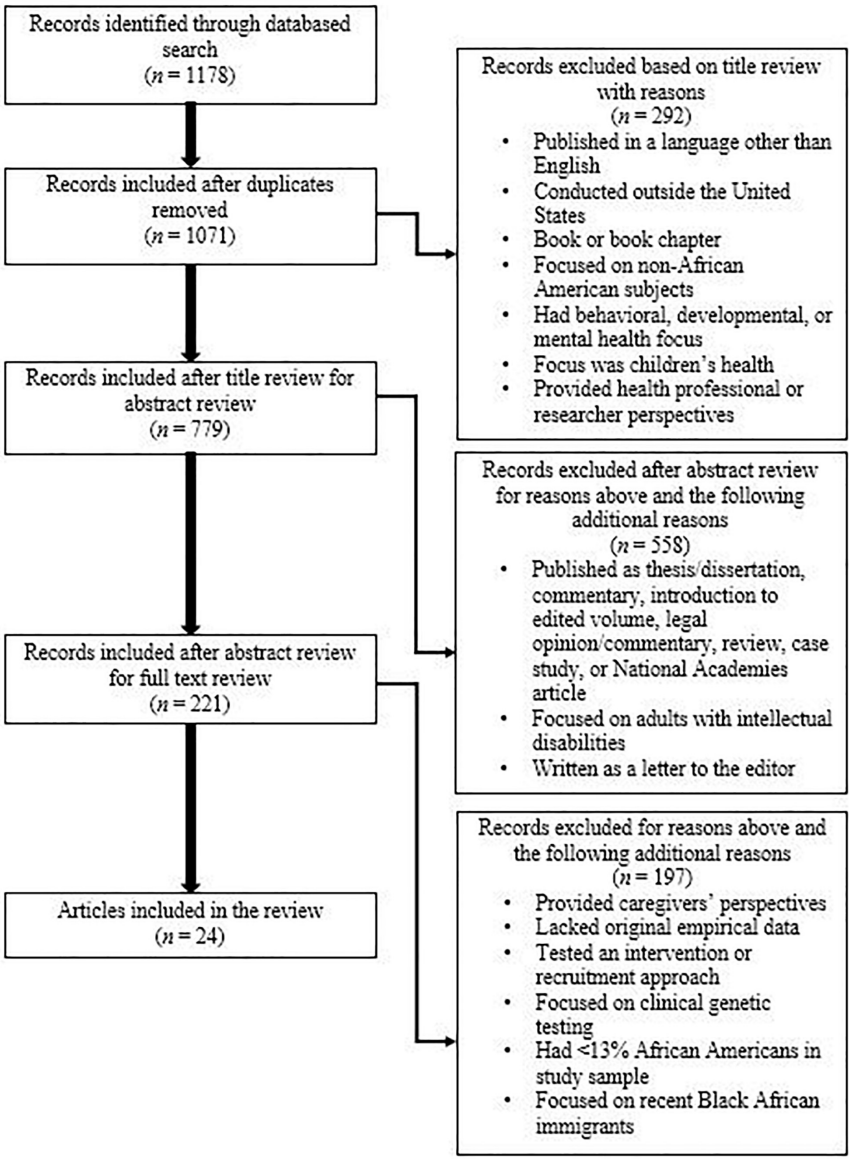

FIGURE 1 | Exclusion process.

between distrust of medical research and beliefs that African American participation in research is imperative (Bates and Harris, 2004; Ochs-Balcom et al., 2011; McDonald et al., 2012; Erwin et al., 2013; Hagiwara et al., 2014). In particular, participants described the necessity of African American participation in order to determine the efficacy and optimal dosing (i.e., PGX) and find more effective ways to treat and prevent diseases which frequently impact their race (Bates and Harris, 2004; Buseh et al., 2013; Erwin et al., 2013). In one study, neither concerns about exploitation nor distrust of medical research were associated with willingness to donate biological specimens for research (Hagiwara et al., 2014).

In contrast, some studies found African American participants trusted medical research and biobanks, and were favorable toward medical research (Hagiwara et al., 2014; Walker et al., 2014; Cain et al., 2016). More recent studies assessing African American community members' knowledge, beliefs, and attitudes about medical and genomic research found study participants did not believe they would be taken advantage of or harmed by research focused on minorities (Cain et al., 2016; Jones et al., 2017). Female members of The Links Incorporated (a not-for-profit African American service organization) who believed research conducted in the United States was ethical were more willing to participate in genomic studies (Brewer et al., 2014).

The overarching theme of distrust was present in most, but not all studies. Even among those with high levels of distrust, the importance of African Americans' participation in medical and genomic research was recognized. This dichotomy may explain why some studies found high levels of distrust and others did not. Participants' divergent views may underlie an attempt to reconcile beliefs about distrust of medical research with the importance of their participation in medical research to avoid cognitive dissonance.

Community Engagement

Participants described community engagement as one strategy to overcome distrust. Community members and leaders described how researchers often entered their community to obtain something from them, and then simply left (Buseh et al., 2013). Such interactions left the community feeling used, disrespected and engendered continued distrust (Buseh et al., 2013). Failing to engage community members prior to conducting studies was viewed as a barrier (Hoyo et al., 2003), whereas genuine engagement, care and communication were viewed as facilitators that created trust (Walker et al., 2014). "Authentic collaboration" is desired which means that researchers: (1) engage with community leaders and the community at the start of the project before major decisions are made, (2) ensure proper resources are available, (3) give credit to the communities, (4) maintain community engagement beyond the study, and (5) share study outcomes (Buseh et al., 2013; Cohn et al., 2015). Participants did not desire frequent contact, but they wanted to know how their participation contributed to the advancement of science (Cohn et al., 2015). Similarly, participants in a focus group study recommended working early on in the research process to improve relationships between institutions and community members citing existing strong relationships with local community hospitals as an example (Kraft et al., 2018).

Awareness and Knowledge

Awareness and knowledge of genomics, or a desire to learn more were associated with favorable attitudes toward genomic studies and/or intentions to participate (Hoyo et al., 2003; Ochs-Balcom et al., 2011; Cohn et al., 2015; Jones et al., 2017). Conversely, lack of education, understanding, awareness or knowledge were associated with less favorable attitudes and lower intentions to participate (Hoyo et al., 2003; Bates and Harris, 2004; Ochs-Balcom et al., 2011; Skinner et al., 2015; Drake et al., 2017). Participants noted that information about research studies was not readily available in their communities, or that African Americans are often not approached or asked to participate (Drake et al., 2017).

Participants described opportunities to overcome low levels of awareness, such as providing educational sessions to ensure informed participation of African Americans (Buseh et al., 2013). Participants in another study suggested that researchers could learn as much from the community as the community could learn from researchers, and advocated for bidirectional educational efforts be bidirectional (Buseh et al., 2013). Similarly, research targeting the African American community was viewed as an opportunity for collaboration between researchers and community members (Cohn et al., 2015). Tying together trust and education, participants suggested that one way to prevent mistreatment of African Americans was for them to request additional information about research studies during recruitment (Bates and Harris, 2004; McDonald et al., 2012). Given this finding, researchers should anticipate that African Americans will have a greater need for information about study procedures than white participants do.

Process of Study Conduct

Across studies, African Americans described their attitudes and beliefs about particular aspects of the research process including research team members and/or the associated institution, study procedures and safeguards, participation risk and compensation. We describe each category next.

Face of the Study

African Americans reported in two studies that they were more likely to participate in research conducted by Historically Black Colleges (HBC) (Hoyo et al., 2003; Diaz et al., 2008). HBCs were viewed as more trustworthy, and participants believed the involvement of HBCs would ensure results and benefits from their participation would be returned to the African American community (Hoyo et al., 2003). Additionally, African Americans want to see African American physicians and/or researchers in leadership roles on the research team (Hoyo et al., 2003; Bates and Harris, 2004; Buseh et al., 2013; McDonald et al., 2014; Cain et al., 2016). It was believed researchers from shared racial backgrounds would be more likely to understand relevant cultural beliefs and experiences, and were viewed as more trustworthy (Hoyo et al., 2003; Bates and Harris, 2004; Buseh et al., 2013; McDonald et al., 2014; Cain et al., 2016; Kraft et al., 2018). In two studies African Americans reported that they were more likely to participate if the investigator was African American (Diaz et al., 2008; McDonald et al., 2014), and one study found a decreased likelihood of participation if the study was conducted by a predominately white college or a white investigator (Diaz et al., 2008).

Similarly, participants across several studies preferred information about genomic research or specific studies be delivered by African Americans (Diaz et al., 2008; Dash et al., 2014), particularly if the study was race specific (McDonald et al., 2014). Participants reported more favorable attitudes toward research, and an increased likelihood of enrollment when the study was introduced by a trusted other such as their physician, friends, family members, and/or community leaders (Hoyo et al., 2003; Diaz et al., 2008; Drake et al., 2017). Participants suggested that hearing about the research study within their community, and knowing others in their community who were involved in the study, would increase their likelihood of participation (Drake et al., 2017).

Study Procedures and Safeguards

Given past injustices, African Americans held significant concerns about the use and accessibility of their data by other individuals or institutions. Due to racism and possible malevolent intent, across studies African Americans wanted to know specifically how their biological material might be used (Buseh et al., 2013; Hagiwara et al., 2014). Not knowing

specifically how the specimen would be used was a barrier to participation (Dash et al., 2014). There were concerns about surreptitious use of genetic material for surveillance, to deny rights and privileges, in criminal investigations, and for other uses beyond the purpose of their original consent (Hoyo et al., 2003; Buseh et al., 2013; Cohn et al., 2015; Kraft et al., 2018). In addition to the aforementioned concerns, participants in one focus group study specifically mentioned concerns related to identity, cloning, and the use of their sample after death (Lee et al., 2019). In addition, not knowing who would have access to their personal information, and who might obtain access to their personal information (e.g., other medical entities like insurance companies) raised concerns, and in some cases, significantly decreased likelihood of participation (Ochs-Balcom et al., 2011; McDonald et al., 2014; Walker et al., 2014; Halbert et al., 2016). Across studies, transparency of study procedures and clear descriptions about safeguards to protect participant privacy were determined essential for participation. Specifically, African Americans want transparency and to know as much as possible about the purpose and rationale for the study, how their specimen would be used and by whom, and the safeguards in place to protect their privacy (Dash et al., 2014; Hagiwara et al., 2014; Skinner et al., 2015; Kraft et al., 2018). Furthermore, continued and ongoing communication about changes to study protocols, or changes to sample access, and the specific studies for which their sample would be used was important, as was maintaining the option to opt in or out of particular studies (Kraft et al., 2018; Lee et al., 2019).

Participation Risk

One study identified that beliefs about the risk of participation were negatively associated with willingness to participate (Brewer et al., 2014), but another study found concerns about the risk of participation was only a consideration when making participation decisions (McDonald et al., 2012). Another study found African Americans were specifically worried about the possible contamination of equipment used for biospecimen collection (Hagiwara et al., 2014). Aside from risk, concerns about procedures primarily focused on invasiveness. Studies found participants least preferred studies where methods were viewed as invasive (Cain et al., 2016), and were more favorable toward participating in studies they believed were less invasive in terms of procedure, privacy, and resources (Hoyo et al., 2003; Diaz et al., 2008; Cain et al., 2016). Although one study found blood donation for participation in a genomic study to be minimally invasive (McDonald et al., 2012), other studies identified fear of needles or the donation of blood as a barrier to study participation (Ochs-Balcom et al., 2011; Dash et al., 2014; Drake et al., 2017).

Concerns about invasiveness included the expenditure of resources, specifically, cost and time. Participants in one study raised concerns about the potential personal costs of participating including costs associated with blood draws and genetic analysis (Skinner et al., 2015). Possible sustained participation in a longitudinal study evoked questions about the number of tasks and time required of participants (Hoyo et al., 2003; McDonald et al., 2012); participants were more favorable about participating in studies which only lasted a short period of time (McDonald et al., 2014). Participants viewed the distance they had to travel for study participation as a barrier to participation (McDonald et al., 2012; Cain et al., 2016). Any perceived expense to the participant such as cost or time for participation, including time that would be taken from work (Walker et al., 2014; Skinner et al., 2015) and transportation issues (McDonald et al., 2014; Halbert et al., 2016) were barriers to participation, unless compensation could be provided (Cain et al., 2016).

Compensation

African Americans expected compensation for participants' time for any study that required any type of time commitment, including travel. Compensation for such expenses were believed to increase participation (Erwin et al., 2013; Skinner et al., 2015; Cain et al., 2016; Drake et al., 2017; Jones et al., 2017), and in some cases, African Americans suggested profit sharing as a means for compensation (Buseh et al., 2013; Jones et al., 2017). However, across studies it was noted that the form of compensation did not always need to be direct participant payment. African Americans suggested that food, gas cards, healthcare and/or medication (Hoyo et al., 2003; Hagiwara et al., 2014; Drake et al., 2017), and even individual research results could be provided as a form of compensation (Skinner et al., 2015; Jones et al., 2017). Indeed, some studies found failure to provide research results to participants would prevent African Americans from participating (McDonald et al., 2014; Halbert et al., 2016).

Individual Level Benefits and Drawbacks of Study Participation

African Americans' interest in participating in genomic studies often was driven by beliefs about benefits for themselves, family members, or future generations. In some cases, individual benefit was broadly or unclearly defined (McDonald et al., 2012, 2014; Skinner et al., 2015; Jones et al., 2017). In other studies, individual benefit included the belief that participation in research meant they would receive better health care (Brewer et al., 2014). Participants across several studies believed they would derive individual benefit by learning more about their genetic risk, which, depending on the results, could act as a motivator for making positive lifestyle changes (Buseh et al., 2013; Skinner et al., 2015). Studies conducted with affected participants, or those already at risk for a specific disease, found increased interest in participation when the study could provide knowledge about the particular condition, for example, cancer (Ochs-Balcom et al., 2011; McDonald et al., 2014; Halbert et al., 2016), asthma (Jones et al., 2017), cardiovascular disease, or type 2 diabetes (Skinner et al., 2015).

Aside from personal benefit, African Americans across studies believed participation in genomic or biobank studies could provide insight into disease that would ultimately benefit their family members or future generations (Ochs-Balcom et al., 2011; McDonald et al., 2012; Dash et al., 2014; Walker et al., 2014; Skinner et al., 2015; Drake et al., 2017; Kraft et al., 2018). They also suggested benefits to family members or future generations could be indirect or much further into the future, such as helping researchers develop medicine that may be used by future generations (Dash et al., 2014).

Notably, two studies found participants did not believe there would be a personal benefit from participating in a research study, and did not believe they would be a benefactor of research outcomes (Halbert et al., 2016; Drake et al., 2017). African Americans believed they were unlikely to benefit personally from medical advancements due to insurance discrimination and the out of pocket costs associated with new pharmaceutical treatments (Halbert et al., 2016; Lee et al., 2019). In some cases, African Americans believe harm could come from finding out about a medical condition that they did not want to know about. As a result, in some studies, learning about personal genetic information was identified as a barrier to participation (Ochs-Balcom et al., 2011; Walker et al., 2014; Skinner et al., 2015; Drake et al., 2017; Jones et al., 2017).

At the Community Level

The potential for genomic or biobank studies to improve health outcomes for their community was embraced by participants (Goldenberg et al., 2011; McDonald et al., 2012; Buseh et al., 2013; Walker et al., 2014; Cohn et al., 2015). Several studies highlighted participants' beliefs that African American participation in medical research, and genomic research in particular, is essential as a means to address health issue of traditionally underserved populations as a means to reduce health disparities (Ochs-Balcom et al., 2011; McDonald et al., 2012; Isler et al., 2013; Skinner et al., 2015). African Americans in one study held the belief that their participation in today's research would facilitate personalized medicine and more targeted prevention and treatment options for disease, for future generations of African Americans (Buseh et al., 2013). While African Americans were favorable toward race specific studies designed to improve health outcomes for their own race (Goldenberg et al., 2011; Ochs-Balcom et al., 2011; McDonald et al., 2014; Walker et al., 2014), results from one study found participants felt such studies were more likely to take advantage of or hurt minorities (Jones et al., 2017). Further, African Americans suggested that despite their participation and advances in medicine, they believed study results were unlikely to reach their community as a result of historic barriers to medical care (Luque et al., 2012). As a solution, African Americans suggested that any prevention or treatment innovations resulting from African American participation must be accessible and affordable for those community members (Buseh et al., 2013; Halbert et al., 2016). Yet, concerns were raised about whether genomic studies could address social determinants of health that are typically responsible for poor health outcomes, and are often ignored (Buseh et al., 2013).

Related to the belief that their participation could benefit their community, favorable views about participation in genomic studies or biobanks most frequently stemmed from altruistic beliefs. Participants believed participation in genomic studies would help future patients or people in general (McDonald et al., 2012; Skinner et al., 2015; Kraft et al., 2018). Caring for others and the benefit of participation to society were central to motivating

TABLE 2 | Summary of Beliefs and Attitudes and Message Design Opportunities.

Beliefs and attitudes	Message design opportunities
Barriers to recruitment	
Distrust – of researchers, universities or health care organizations, science and medicine at large	1. Establish relationship with community members prior to beginning research study and engage them in recruitment design efforts
	2. Consider engaging African American community members, including other research participants and community health care workers, as the senders/disseminators of recruitment messages
	3. Engage African American study team members as senders/disseminators of recruitment messages
	4. Provide a clear description of study purpose, procedures, who will be able to access their data and privacy safeguards in place
	5. Messages about the use of participant data should be clearly detailed
	6. Describe how information from the study may impact health care for the African American population
	7. Any and all forms of compensation should be clearly described in any study asking for participants' time, including travel time
Lack of Education – about research studies and genetics created less favorable attitudes about participation	1. Outreach efforts should focus on providing more information about genomic studies more broadly
	2. Delivering in-person education may be advantageous because researchers can address additional questions or concerns on the spot, and at the same time engage with and learn from the population
	3. Combine educational messages with messages that describe use of data and standard privacy protections that are in place
	4. Messages should provide detailed information about research purpose, processes and outcomes
Facilitators of recruitment	
Participation – beliefs that African American participation is necessary and essential	1. Messages should emphasize the importance of African American participation for their community
	2. When appropriate, messages should describe any potential individual level benefit from participation in the study
	3. When appropriate, messages should describe any potential future benefit to family members
	4. When appropriate, messages should describe any potential future benefit for the African American community
	5. Messages about altruism should be included in recruitment efforts

participation (Brewer et al., 2014; Jones et al., 2017), despite concerns about trust (Bates and Harris, 2004).

DISCUSSION AND CONCLUSION

Given favorable attitudes, but low participation rates, culturally appropriate and ethical messages about PGX studies that facilitate recruitment of African Americans are needed (Halbert et al., 2016). Trust has often been cited as the leading barrier to African American participation in health-related research (George et al., 2014; Luebbert and Perez, 2016; Hughes et al., 2017; Jones et al., 2017). Consistently, our review found that distrust in the healthcare system, medical research, organization, and researchers is a commonly held belief by many African Americans (Bates and Harris, 2004; Bussey-Jones et al., 2010; Hagiwara et al., 2014; McDonald et al., 2014; Cohn et al., 2015; Skinner et al., 2015; Halbert et al., 2016; Drake et al., 2017). We forward several suggestions to overcome distrust (see **Table 2**). First, meaningful and intentional community collaboration can demonstrate value and meaning for African American participants (Walker et al., 2014). Indeed, a systematic review conducted by Johnson et al. (2011) identified community-based strategies, such as engaging community leadership, as one method for improving recruitment of African Americans into genomic research. However, results from our review suggest researchers must move beyond simply contacting community leaders at the time of the study. Instead, researchers should engage in what participants called "authentic collaboration" from before the start of the research study and extending after the study as a means to foster trust, demonstrate respect and honor the value of community contributions (Buseh et al., 2013; Cohn et al., 2015). These findings are consistent with the success of other studies, which have used CBA as a method to improve recruitment of African Americans (Israel et al., 1998; Vadaparampil and Pal, 2010; Kiviniemi et al., 2013; Ochs-Balcom et al., 2015; McNeill et al., 2018).

Our review also identified lack of knowledge or awareness about genomic studies as an overarching barrier (Hoyo et al., 2003; James et al., 2008; Skinner et al., 2008; Ochs-Balcom et al., 2011; Drake et al., 2017). However, educational interventions have demonstrated little impact on attitudes or beliefs, thus suggesting messages that address existing attitudes and beliefs in addition to providing education may be more effective at addressing African Americans' concerns about participation in genomic studies (Skinner et al., 2008; Halverson and Ross, 2012). Furthermore, it could be argued that beliefs about the trustworthiness of research scientists or institutions (Luque et al., 2012; Erwin et al., 2013; Hagiwara et al., 2014; Walker et al., 2014) impact African Americans' expectations for research participation. For example, African Americans concerns about being experimented on or exploited explain why they want complete transparency about study protocols and data sharing practices (Dash et al., 2014; Hagiwara et al., 2014; Skinner et al., 2015). As such, messages that are transparent and clearly describe the study protocol may reduce mistrust as a barrier. Based on our review, messages for African Americans about genomic studies should provide substantial information about the study purpose and procedure and describe processes and measures in place to safeguard their privacy. Previous research found that messages which intentionally highlight procedures and security are more likely to overcome concerns related to privacy and outcomes (McQuillan et al., 2006; George et al., 2014; Luebbert and Perez, 2016; Hughes et al., 2017; Jones et al., 2017).

Contrary to the belief that minority populations are not interested in participating in research studies, our review found African Americans were highly interested in participating (Wendler et al., 2005; Horowitz et al., 2017; Jones et al., 2017). Studies in our review indicated African Americans believed their participation in medical research was crucial for the advancement of science (Bates and Harris, 2004; McDonald et al., 2012; Erwin et al., 2013; Hagiwara et al., 2014). Thus, researchers should devote more attention to facilitators of African American participation in medical research. Specifically, as identified in our review, messages that highlight altruism or benefit for one's community and recognize the importance of including minority populations may promote participation in clinical studies of African Americans (George et al., 2014; Hughes et al., 2017; Jones et al., 2017).

Ultimately, one goal of PM research is to reduce health disparities (Collins and Varmus, 2015; Khoury et al., 2016). In particular, PGX uses personal genomic data to inform optimal tailoring of pharmaceutical agents to prevent adverse drug interactions (Perera et al., 2014). Despite the individualized focus of PGX, efforts require a population-based approach to better understand inter-population and intrapopulation diversity (Bonham et al., 2016; Khoury et al., 2016). This review drew upon existing literature to provide a consolidated overview of African American's beliefs and attitudes toward genomic research. This information can inform recruitment strategies and messages that may increase African American participation in genomic studies, and PGX studies in particular. Future research testing the message strategies identified in this review are needed to continue to understand best practices for communicating genomic research with the African American population. Additionally, future studies should explore African Americans' beliefs and attitudes regarding PGX studies. Such knowledge may contribute to the advancement of PM among minority populations.

AUTHOR CONTRIBUTIONS

CS, SR, and MP conceptualized the study. CS, SR, and CM-F devised the methods, conducted the literature search, reviewed the literature, and conducted the analysis. CS and SR drafted the manuscript. MP reviewed and edited the manuscript.

FUNDING

This study was supported by U54 MD010723 African American Cardiovascular pharmacogenetic CONsorTium (ACCOUNT): discovery and translation.

REFERENCES

Adams-Campbell, L. L., Dash, C., Palmer, J. R., Wiedemeier, M. V., Russell, C. W., Rosenberg, L., et al. (2016). Predictors of biospecimen donation in the black women's health study. *Cancer Causes Control* 27, 797–803. doi: 10.1007/s10552-016-0747-0

Bates, B. R., and Harris, T. M. (2004). The tuskeegee study of untreated syphilis and public perceptions of biomedical research: a focus group study. *J. Natl. Med. Assoc.* 96, 1051–1064.

Bonham, V. L., Callier, S. L., and Royal, C. D. (2016). Will precision medicine move us beyond race? *N. Engl. J. Med.* 374, 2003–2005. doi: 10.1056/nejmp1511294

Brewer, L. C., Hayes, S. N., Parker, M. W., Balls-Berry, J. E., Halyard, M. Y., Pinn, V. W., et al. (2014). African American women's perceptions and attitudes regarding participation in medical research: the mayo clinic/the links, incorporated partnership. *J. Womens Health* 23, 681–687. doi: 10.1089/jwh.2014.4751

Buseh, A. G., Stevens, P. E., Millon-Underwood, S., Townsend, L., and Kelber, S. T. (2013). Community leaders' perspectives on engating African Americans in biobanks and otehr human geneitcs initiatives. *J. Commun. Genet.* 4, 483–494. doi: 10.1007/s12687-013-0155-z

Bussey-Jones, J., Garrett, J., Henderson, G., Moloney, M., Blumenthal, C., and Corbie-Smith, G. (2010). The role of race and trust in tissue/blood donation for genetic research. *Genet. Med.* 12, 116–121. doi: 10.1097/GIM.0b013e3181cd6689

Cain, G. E., Kalu, N., Kwagyan, J., Marshall, V. J., Ewing, A. T., Bland, W. P., et al. (2016). Beliefs and preferences for medical research among African-Americans. *J. Racial Ethn. Health Disparities* 3, 74–82. doi: 10.1007/s40615-015-0117-8

Cohn, E. G., Husamudeen, M., Larson, E. L., and Williams, J. K. (2015). Increasing participation in genomic research and biobanking through community-based capacity building. *J. Genet. Counsel.* 24, 491–502. doi: 10.1007/s10897-014-9768-6

Collins, F. S., and Varmus, H. (2015). A new initiative on precision medicine. *N. Engl. J. Med.* 372, 793–795. doi: 10.1056/nejmp1500523

Dash, C., Wallington, S. F., Muthra, S., Dodson, E., Mandelblatt, J., and Adams-Campbell, L. L. (2014). Disparities in knowledge and willingness to donate research biospecimens: a mixed-methods study in an underserved urban community. *J. Commun. Genet.* 5, 329–336. doi: 10.1007/s12687-014-0187-z

Diaz, V. A., Mainous Iii, A. G., Mccall, A. A., and Geesey, M. E. (2008). Factors affecting research participation in African American college students. *Family Med.* 40, 46–51.

Drake, B. F., Boyd, D., Carter, K., Gehlert, S., and Thompson, V. S. (2017). Barriers and strategies to participation in tissue research among African-American men. *J. Cancer Educ.* 32, 51–58. doi: 10.1007/s13187-015-0905-1

Dye, T., Li, D., Demment, M., Groth, S., Fernandez, D., Dozier, A., et al. (2016). Sociocultural variation in attitudes toward use of genetic information and participation in genetic research by race in the United States: implications for precision medicine. *J. Am. Med. Inform. Assoc.* 23, 782–786. doi: 10.1093/jamia/ocv214

Empey, P. E. (2016). Pharmacogenomics to achieve precision medicine. *Am. J. Health Syst. Pharm.* 73, 1906–1907. doi: 10.2146/ajhp160682

Erwin, D. O., Moysich, K., Kiviniemi, M. T., Saad-Harfouche, F. G., Davis, W., Clark-Hargrave, N., et al. (2013). Community-based partnership to identify keys to biospecimen research participation. *J. Cancer Educ.* 28, 43–51. doi: 10.1007/s13187-012-0421-5

Fishbein, M., and Yzer, M. C. (2003). Using theory to design effective health behavior interventions. *Commun. Theory* 13, 164–183. doi: 10.1093/ct/13.2.164

Gamble, V. N. (1997). Under the shadow of tuskegee: african americans and health care. *Am. J. Public Health* 87, 1773–1778. doi: 10.2105/ajph.87.11.1773

George, S., Duran, N., and Norris, K. (2014). A systematic review of barriers and facilitators to minority research participation among african americans, Latinos, Asian Americans, and Pacific Islanders. *Am. J. Public Health* 104, e16–e31. doi: 10.2105/AJPH.2013.301706

Goldenberg, A. J., Hull, S. C., Wilfond, B. S., and Sharp, R. R. (2011). Patient perspectives on group benefits and harms in genetic research. *Public Health Genomics* 14, 135–142. doi: 10.1159/000317497

Grimshaw, J. M., Shirran, L., Thomas, R., Mowatt, G., Fraser, C., Bero, L., et al. (2002). "Changing provider behavior: An overview of systematic reviews of interventions to promote implementation of research findings by healthcare professionals," in *Getting Research Findings into Practice*, ed. A. Haines (London: John Wiley & Sons), 29–65.

Hagiwara, N., Berry-Bobovski, L., Francis, C., Ramsey, L., Chapman, R. A., and Albrecht, T. L. (2014). Unexpected findings in the exploration of African American underrepresentation in biospecimen collection and biobanks. *J. Cancer Educ.* 29, 580–587. doi: 10.1007/s13187-013-0586-6

Halbert, C. H., Mcdonald, J., Vadaparampil, S., Rice, L., and Jefferson, M. (2016). Conducting precision medicine research with African Americans. *PLoS One* 11:e0154850. doi: 10.1371/journal.pone.0154850

Halverson, C. M., and Ross, L. F. (2012). Engaging African-Americans about biobanks and the return of research results. *J. Commun. Genet.* 3, 275–283. doi: 10.1007/s12687-012-0091-3

Horowitz, C. R., Ferryman, K., Negron, R., Sabin, T., Rodriguez, M., Zinberg, R. F., et al. (2017). Race, genomics and chronic disease: what patients with African ancestry have to say. *J. Health Care Poor Underserved* 28, 248–260. doi: 10.1353/hpu.2017.0020

Hoyo, C., Reid, M. L., Godley, P. A., Parrish, T., Smith, L., and Gammon, M. (2003). Barriers and strategies for sustained participation of African American men in cohort studies. *Ethn. Dis.* 13, 470–476.

Hughes, T. B., Varma, V. R., Pettigrew, C., and Albert, M. S. (2017). African Americans and clinical research: evidence concerning barriers and facilitators to participation and recruitment recommendations. *Gerontologist* 57, 348–358. doi: 10.1093/geront/gnv118

Isler, M. R., Sutton, K., Cadigan, R. J., and Corbie-Smith, G. (2013). Community perceptions of genomic research: implications for addressing health disparities. *N. C. Med. J.* 74, 470–476.

Israel, B. A., Schulz, A. J., Parker, E. A., and Becker, A. B. (1998). Review of community-based research: assessing partnership approaches to improve public health. *Annu. Rev. Public Health* 19, 173–202. doi: 10.1146/annurev.publhealth.19.1.173

Jackson, F. (1999). African-American responses to the human genome project. *Public Underst. Sci.* 8, 181–191. doi: 10.1088/0963-6625/8/3/303

Jaffe, S. (2015). Planning for US precision medicine initiative underway. *Lancet* 385, 2448–2449. doi: 10.1016/s0140-6736(15)61124-2

James, R. D., Yu, J. H., Henrikson, N. B., Bowen, D. J., and Fullerton, S. M. (2008). Strategies and stakeholders: minority recruitment in cancer genetics research. *Public Health Genomics* 11, 241–249. doi: 10.1159/000116878

Johnson, V. A., Powell-Young, Y. M., Torres, E. R., and Spruill, I. J. (2011). A systematic review of strategies that increase the recruitment and retention of African American adults in genetic and genomic studies. *ABNF J.* 22, 84–88.

Jones, B. L., Vyhlidal, C. A., Bradley-Ewing, A., Sherman, A., and Goggin, K. (2017). If we would only ask: how henrietta lacks continues to teach us about perceptions of research and genetic research among African Americans today. *J. Racial Ethn. Health Disparities* 4, 735–745. doi: 10.1007/s40615-016-0277-1

Khoury, M. J., Iademarco, M. F., and Riley, W. T. (2016). Precision public health for the era of precision medicine. *Am. J. Preven. Med.* 50, 398–401. doi: 10.1016/j.amepre.2015.08.031

Kiviniemi, M. T., Saad-Harfouche, F. G., Ciupak, G. L., Davis, W., Moysich, K., Hargrave, N. C., et al. (2013). Pilot intervention outcomes of an educational program for biospecimen research participation. *J. Cancer Educ.* 28, 52–59. doi: 10.1007/s13187-012-0434-0

Kraft, S. A., Cho, M. K., Gillespie, K., Halley, M., Varsava, N., Ormond, K. E., et al. (2018). Beyond consent: building trusting relationships with diverse populations in precision medicine research. *Am. J. Bioethics* 18, 3–20. doi: 10.1080/15265161.2018.1431322

Krippendorff, K. (2004). Reliability in content analysis: some common misconceptions and recommendations. *Hum. Commun. Res.* 30, 411–433. doi: 10.1111/j.1468-2958.2004.tb00738.x

Lee, S. S.-J. (2003). Race, distributive justice and the promise of pharmacogenomics. *Am. J. Pharm.* 3, 385–392. doi: 10.2165/00129785-200303060-00005

Lee, S. S.-J., Cho, M. K., Kraft, S. A., Varsava, N., Gillespie, K., Ormond, K. E., et al. (2019). I don't want to be Henrietta Lacks": diverse patient perspectives on donating biospecimens for precision medicine research. *Genet. Med.* 21, 107–113. doi: 10.1038/s41436-018-0032-6

Luebbert, R., and Perez, A. (2016). Barriers to clinical research participation among African Americans. *J. Transcult. Nurs.* 27, 456–463. doi: 10.1177/1043659615575578

Luque, J. S., Quinn, G. P., Montel-Ishino, F. A., Arevalo, M., Bynum, S. A., Noel-Thomas, S., et al. (2012). Formative research on perceptions of biobanking: what community members think. *J. Cancer Educ.* 27, 91–99. doi: 10.1007/s13187-011-0275-2

Marteau, T. M., Sowden, A. J., and Armstrong, D. (2002). "Implementing research findings into practice: Beyond the information deficit model," in *Getting Research Findings into Practice*, ed. A. Haines (London: John Wiley & Sons), 36–42.

McDonald, J. A., Barg, F. K., Weathers, B., Guerra, C. E., Troxel, A. B., Domchek, S., et al. (2012). Understanding participation by African Americans in cancer genetics research. *J. Natl. Med. Assoc.* 104, 324–330. doi: 10.1016/s0027-9684(15)30172-3

McDonald, J. A., Vadaparampil, S., Bowen, D., Magwood, G., Obeid, J. S., Jefferson, M., et al. (2014). Intentions to donate to a biobank in a national sample of African Americans. *Public Health Genomics* 17, 173–182. doi: 10.1159/000360472

McNeill, L. H., Reitzel, L. R., Escoto, K. H., Roberson, C. L., Nguyen, N., Vidrine, J. I., et al. (2018). Engaging black churches to address cancer health disparities: project CHURCH. *Front. Public Health* 6:191. doi: 10.3389/fpubh.2018.00191

McQuillan, G. M., Pan, Q., and Porter, K. S. (2006). Consent for genetic research in a general population: an update on the national health and nutrition examination survey experience. *Genet. Med.* 8:354. doi: 10.1097/01.gim.0000223552.70393.08

Moledina, D. G., Cheung, B., Kukova, L., Luciano, R. L., Peixoto, A. J., Wilson, F. P., et al. (2018). A survey of patient attitudes toward participation in biopsy-based kidney research. *Kidney Int. Rep.* 3, 412–416. doi: 10.1016/j.ekir.2017.11.008

Ochs-Balcom, H. M., Jandorf, L., Wang, Y., Johnson, D., Meadows Ray, V., Willis, M. J., et al. (2015). "It takes a village": multilevel approaches to recruit African Americans and their families for genetic research. *J. Commun. Genet.* 6, 39–45. doi: 10.1007/s12687-014-0199-8

Ochs-Balcom, H. M., Rodriguez, E. M., and Erwin, D. O. (2011). Establishing a community partnership to optimize recruitment of African American pedigrees for a genetic epidemiology study. *J. Commun. Genet.* 2, 223–231. doi: 10.1007/s12687-011-0059-8

Patel, K., Inman, W., Gishe, J., Johnson, O., Brown, E., Kanu, M., et al. (2018). A community-driven intervention for improving biospecimen donation in African American communities. *J. Racial Ethn. Health Disparities* 5, 15–23. doi: 10.1007/s40615-017-0336-2

Patel, Y. R., Carr, K. A., Magjuka, D., Mohammadi, Y., Dropcho, E. F., Reed, A. D., et al. (2012). Successful recruitment of healthy African American men to genomic studies from high-volume community health fairs: implications for future genomic research in minority populations. *Cancer* 118, 1075–1082. doi: 10.1002/cncr.26328

Perera, M. A., Cavallari, L. H., and Johnson, J. A. (2014). Warfarin pharmacogenetics: an illustration of the importance of studies in minority populations. *Clin. Pharmacol. Ther.* 95, 242–244. doi: 10.1038/clpt.2013.209

Radecki Breitkopf, C., Williams, K. P., Ridgeway, J. L., Parker, M. W., Strong-Simmons, A., Hayes, S. N., et al. (2018). Linking education to action: a program to increase research participation among African American women. *J. Womens Health* 27, 1242–1249. doi: 10.1089/jwh.2017.6791

Rodriguez, E. M., Saad-Harfouche, F. G., Miller, A., Mahoney, M. C., Ambrosone, C. B., Morrison, C. D., et al. (2016). Engaging diverse populations in biospecimen donation: results from the Hoy y Mañana study. *J. Commun. Genet.* 7, 271–277. doi: 10.1007/s12687-016-0275-3

Scherr, C. L., Dean, M., Clayton, M. F., Hesse, B. W., Silk, K., Street, R. L., et al. (2017). A research agenda for communication scholars in the precision medicine era. *J. Health Commun.* 22, 839–848. doi: 10.1080/10810730.2017.1363324

Skinner, C. S., Schildkraut, J. M., Calingaert, B., Hoyo, C., Crankshaw, S. S., Fish, L., et al. (2008). Factors associated with African Americans' enrollment in a national cancer genetics registry. *Commun. Genet.* 11, 224–233. doi: 10.1159/000116883

Skinner, H. G., Calancie, L., Vu, M. B., Garcia, B., Demarco, M., Patterson, C., et al. (2015). Using community-based participatory research principles to develop more understandable recruitment and informed consent documents in genomic research. *PLoS One* 10:e0125466. doi: 10.1371/journal.pone.0125466

Vadaparampil, S. T., and Pal, T. (2010). Updating and refining a study brochure for a cancer registry-based study of BRCA mutations among young African American breast cancer patients: lessons learned. *J. Commun. Genet.* 1, 63–71. doi: 10.1007/s12687-010-0010-4

VERBI Software (2018). *MAXQDA Analytics Pro*. Berlin: VERBI.

Walker, E. R., Nelson, C. R., Antoine-Lavigne, D., Thigpen, D. T., Puggal, M. A., Sarpong, D. F., et al. (2014). Research participants' opinions on genetic research and reseasons for participation: a jackson heart study focus group analysis. *Ethn. Dis.* 24, 290–297.

Wendler, D., Kington, R., Madans, J., Van Wye, G., Christ-Schmidt, H., Pratt, L. A., et al. (2005). Are racial and ethnic minorities less willing to participate in health research? *PLoS Med.* 3:e19. doi: 10.1371/journal.pmed.0030019

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 19 June 2019
doi: 10.3389/fgene.2019.00538

The Puerto Rico Alzheimer Disease Initiative (PRADI): A Multisource Ascertainment Approach

Briseida E. Feliciano-Astacio[1], Katrina Celis[2], Jairo Ramos[2], Farid Rajabli[2], Larry Deon Adams[2], Alejandra Rodriguez[1], Vanessa Rodriguez[2], Parker L. Bussies[2], Carolina Sierra[1], Patricia Manrique[2], Pedro R. Mena[2], Antonella Grana[2], Michael Prough[2], Kara L. Hamilton-Nelson[2], Nereida Feliciano[3], Angel Chinea[1], Heriberto Acosta[4], Jacob L. McCauley[2], Jeffery M. Vance[2], Gary W. Beecham[2], Margaret A. Pericak-Vance[2] and Michael L. Cuccaro[2]*

[1] Department of Internal Medicine, Universidad Central Del Caribe, Bayamón, PR, United States, [2] John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, FL, United States, [3] VA Caribbean Healthcare System, San Juan, PR, United States, [4] Clínica de la Memoria, San Juan, PR, United States

OPEN ACCESS

Edited by:
Kelli K. Ryckman,
The University of Iowa, United States

Reviewed by:
Phillip E. Melton,
Curtin University, Australia
Bethany Wolf,
Medical University of South Carolina, United States

***Correspondence:**
Margaret A. Pericak-Vance
MPericak@med.miami.edu

Specialty section:
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics

Received: 18 September 2018
Accepted: 17 May 2019
Published: 19 June 2019

Citation:
Feliciano-Astacio BE, Celis K, Ramos J, Rajabli F, Adams LD, Rodriguez A, Rodriguez V, Bussies PL, Sierra C, Manrique P, Mena PR, Grana A, Prough M, Hamilton-Nelson KL, Feliciano N, Chinea A, Acosta H, McCauley JL, Vance JM, Beecham GW, Pericak-Vance MA and Cuccaro ML (2019) The Puerto Rico Alzheimer Disease Initiative (PRADI): A Multisource Ascertainment Approach. Front. Genet. 10:538. doi: 10.3389/fgene.2019.00538

Introduction: Puerto Ricans, the second largest Latino group in the continental US, are underrepresented in genomic studies of Alzheimer disease (AD). To increase representation of this group in genomic studies of AD, we developed a multisource ascertainment approach to enroll AD patients, and their family members living in Puerto Rico (PR) as part of the Alzheimer's Disease Sequencing Project (ADSP), an international effort to advance broader personalized/precision medicine initiatives for AD across all populations.

Methods: The Puerto Rico Alzheimer Disease Initiative (PRADI) multisource ascertainment approach was developed to recruit and enroll Puerto Rican adults aged 50 years and older for a genetic research study of AD, including individuals with cognitive decline (AD, mild cognitive impairment), their similarly, aged family members, and cognitively healthy unrelated individuals age 50 and up. Emphasizing identification and relationship building with key stakeholders, we conducted ascertainment across the island. In addition to reporting on PRADI ascertainment, we detail admixture analysis for our cohort by region, group differences in age of onset, cognitive level by region, and ascertainment source.

Results: We report on 674 individuals who met standard eligibility criteria [282 AD-affected participants (42% of the sample), 115 individuals with mild cognitive impairment (MCI) (17% of the sample), and 277 cognitively healthy individuals (41% of the sample)]. There are 43 possible multiplex families (10 families with 4 or more AD-affected members and 3 families with 3 AD-affected members). Most individuals in our cohort were ascertained from the Metro, Bayamón, and Caguas health regions. Across health regions, we found differences in ancestral backgrounds, and select clinical traits.

Discussion: The multisource ascertainment approach used in the PRADI study highlights the importance of enlisting a broad range of community resources and providers. Preliminary results provide important information about our cohort that will be useful as we move forward with ascertainment. We expect that results from the PRADI study will lead to a better understanding of genetic risk for AD among this population.

Keywords: Alzheimer disease, ascertainment, PRADI, genetics, community resources, ADSP, diversity, health disparities

INTRODUCTION

Alzheimer disease (AD) is a progressive neurodegenerative disorder that affects 1 in 9 Americans over the age of 65. This disease has a significant impact on individuals with AD and their families and poses huge financial and social burden on society. To date, over 20 loci have been identified as risk factors for AD in non-Hispanic White (NHW), genome wide association studies (GWAS) with limited GWAS in other populations (Lambert et al., 2013). In addition, the only large AD sequencing effort to date, the Alzheimer's Disease sequencing project (ADSP) (Beecham et al., 2017), has focused its efforts on individuals of NHW descent, including a limited number of Hispanic (HI), and African American individuals. The importance of examining AD in other populations (Ramirez et al., 2008) is highlighted by findings that show Caribbean Hispanics from the Dominican Republic are twice as likely as NHW to have late-onset Alzheimer's Disease (LOAD) (Tang et al., 1998, 2001). Furthermore, the incidence of new LOAD cases in families from the Dominican Republic is three times larger than the incidence found in NHW families (Vardarajan et al., 2014) even though the genetic risk of LOAD is similar. Despite clear evidence that points to the importance of investigating AD in underserved populations, this work has lagged.

Although comparisons of risk among different ethnic groups are complicated by differences in the assessment of cognitive decline across studies and population differences in willingness to participate in medical research, there are several possible explanations for increased incidence in these specific ethnic groups (e.g., lower educational attainment, higher rates of cardio- and cerebrovascular disease, and metabolic syndrome). While the importance of diversity and inclusion in genomic research has been emphasized for more than two decades (NIH Revitalization Act of 1993, Public law 103–143) many groups, including Hispanics, are underrepresented in biomedical research (Shavers et al., 2002; Sheppard et al., 2005; Calderon et al., 2006), including genomic, and translational studies (Armstrong et al., 2005; Ricker et al., 2006; Armstrong et al., 2012). Further, this lack of participation has the potential to delay the application of novel treatments that may be relevant to these populations, exacerbating existing health disparities in a variety of diseases, including AD. Specifically, given the importance of genomic research in the development and implementation of precision medicine initiatives (Hampel et al., 2017), there is an urgency to engage with and include underserved and underrepresented groups in such research to enable access to these advanced treatments (Wilkins, 2018).

Alzheimer disease is the most common form of dementia and the fourth leading cause of death in Puerto Rico (PR) (Friedman et al., 2016). The population of PR was estimated at 3,474,182 individuals in 2015, with 617,007 over the age of 65, and AD prevalence of 12.5% (Puerto Rico Department of Health, 2015). Further, according to Perreira et al. (2017) the population of PR is aging and struggles with high rates of comorbid conditions (e.g., hypertension and diabetes) that contribute to dementia. These numbers underscore the need to investigate early risk factors and develop the necessary research to study the neurobiology of cognitive decline in Puerto Ricans and more broadly Hispanics. Furthermore, enriching AD genomic studies with Hispanic populations is fundamental for reducing health disparities, delivering precision medicine, and ultimately improving health outcomes for this community.

To address the range of disparities experienced by Hispanics due to under-representation in genomic studies of AD, we developed the Puerto Rico Alzheimer Disease Initiative (PRADI). The goal of this National Institute of Aging funded project is twofold. First, the PRADI study examines genomic risk for AD in Puerto Ricans and adds to the growing body of knowledge regarding Hispanic risk for AD. Second, the PRADI study makes comparisons using two types of controls: family-based (related controls) and case-control (unrelated controls), paralleling, and building on the ongoing work of the ADSP (Beecham et al., 2017). Furthermore, Puerto Ricans are an admixed population, enriched for at least three ancestries (European Caucasian, Western African, and Amerindian/Taino), resulting in complex population substructure (Claudio-Campos et al., 2015; Rajabli et al., 2018). The use of population substructure (i.e., global and local ancestry) can allow for adjustment of models to improve genetic analyses. The importance of examining ancestral contributions in Hispanics can be seen in studies of complex diseases, including asthma (Gignoux et al., 2019), multiple sclerosis (Amezcua et al., 2018), and cancer (Salgado-Montilla et al., 2017; Diaz-Zabala et al., 2018). The usefulness of understanding and incorporating genotypic and admixture information into the conceptualization and management of disease among Puerto Ricans is becoming increasingly apparent (Morales-Borges, 2017; Diaz-Zabala et al., 2018).

In contrast to other studies of Puerto Ricans (Tucker et al., 2010), the current study focuses exclusively on participants from the island of PR. We describe the design and implementation of our multisource method for recruiting individuals for the genetic study of AD and our corresponding work in the community to increase study participation among eligible Puerto Ricans. Equally important, we describe our cohort with respect to clinical

features and ancestral proportions by region. These results provide a preliminary picture of our PRADI cohort.

MATERIALS AND METHODS

A multisource ascertainment approach was implemented to recruit and enroll participants into the PRADI study. As described below, the approach consisted of different phases that revolved around community engagement and included: (a) identification and relationship building with key stakeholders from several organizations; (b) collaborative agreement on ascertainment methods and formalization using memorandums of understanding; (c) targeted actions and recruitment events; and (d) education and dissemination of information about AD to health professionals and the general public. This approach was designed to establish and strengthen collaborative relationships with key stakeholders to facilitate ascertainment for this study and future studies.

Ascertainment efforts were carried out in PR and encompassed all seven health regions (Arecibo, Bayamón, Caguas, Fajardo, Mayagüez, Metro, and Ponce) as defined by the Puerto Rico Department of Health. Only bilingual personnel were sent to the sites and plain Spanish was used for all verbal and written study-related communication (materials for public dissemination were developed for a third-grade reading level). Standard screening and evaluation activities were performed, which included collection of clinical, family, and medical history and neurocognitive testing. Individuals were determined to be cases or controls with further specification depending on whether they were family history positive or negative for AD.

Finally, to investigate potential differences among our participants from different parts of the island, we tested for differences in age of onset and 3MS scores by health region and ascertainment source (i.e., AD specialist, adult care center, or community event/activity). We also conducted admixture analysis to examine the population substructure of our Puerto Rican cohort by region to evaluate differences in ancestry proportions among the health regions.

Ascertainment Procedures

Ascertainment Phase One: Getting to Know the Field

Stakeholders From Multiple Sectors

In the initial phase of our multisource ascertainment approach, the local team identified potential sources of participants within PR communities by interacting with groups and providers that serve the AD population. There are multiple groups and ongoing community initiatives working to increase AD awareness in PR. Our goal was to establish collaborative relationships with stakeholders from different sectors (**Figure 1**). These interactions served as a starting point to disseminate information about the study, to identify sources for cases and controls, to build networks with potential collaborators, and to create opportunities for direct ascertainment. In addition, these initial meetings served as a venue for discussing the importance of inclusive recruitment in genetics research, especially how a lack of diversity can delay specific populations' access to personalized/precision medicine. The primary groups we approached included:

Governmental Stakeholders

We contacted central and local government representatives, including the PR office of the Ombudsman for the Elderly (OPPEA, for its Spanish acronym), a legal affairs office for older adults, and the AD Registry of the Health Department of PR. As an initial step, local team members joined the Health Department Alzheimer's Advisory Board. This process allowed us to meet with key stakeholders to discuss the PRADI study. Through these initial contacts, OPPEA provided us with additional contacts at the provider level to include various programs and adult care centers for older adults and those with AD and other cognitive problems. Through these contacts, we established ties with additional local government representatives of the

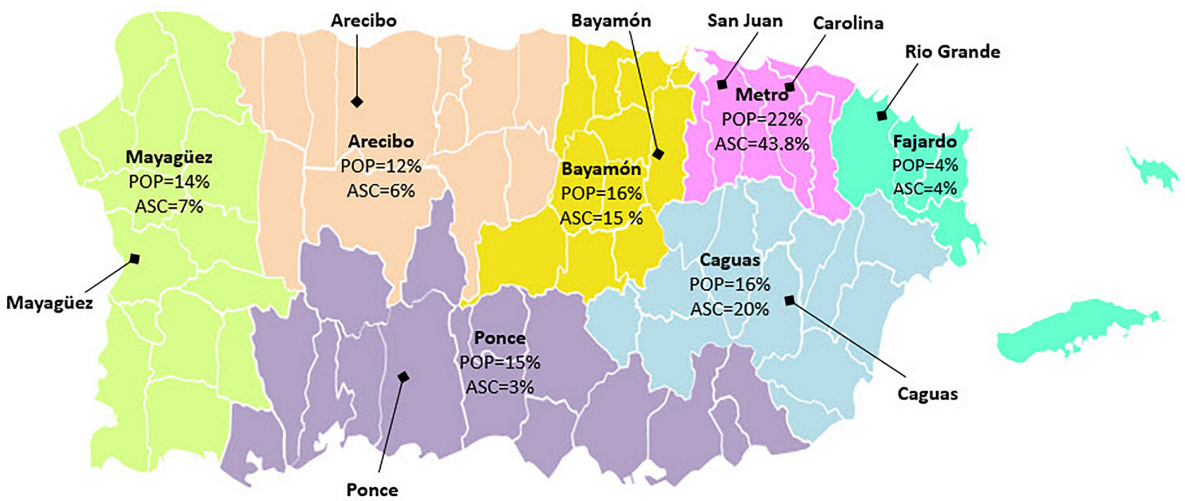

FIGURE 1 | Percentages of PRADI Ascertainment (*N* = 674) by Puerto Rico Department of Health Regions.

municipalities, including Cidra, Fajardo, Carolina, Aguadilla, Arecibo, among others.

Community Non-profit Organizations (NPO)

To establish community based collaborations in the non-profit sector, we contacted multiple groups that serve older adults in PR, including the Puerto Rican Chapter of the AARP; Mente Activa (Active Mind), which is a non-profit organization that promotes physical and mental activity for older adults and those with dementia; and Organización Pro Ayuda a Personas con Alzheimer (OPAPA), another non-profit organization that provides education and support to people with AD and their families. Our team met with leadership in these organizations to provide information about the PRADI study.

Religious Groups

Our primary religious contact was the Lutherans Social Service of PR, a non-profit faith-based organization involved in providing services to older adults. It is funded to provide programs to train dementia capable personnel and service providers as well as programs to identify older adults with early signs of AD. In addition, we contacted the Catholic Church, especially the Seminary of PR and the Caguas Cathedral. Both groups agreed to assist with the study by providing access to participants and disseminating information about our study during religious services and through print media.

Ascertainment Phase Two: Defining and Formalizing Collaborations With Stakeholders

The next phase in our multisource ascertainment approach was seeking and using input from the stakeholders and organizations about best practices for ascertainment. This process typically involved in person discussions between the local team (headed by Dr. Feliciano, a neurologist who specializes in the care of older individuals) and the organizations. This allowed us to define our ascertainment practices in alignment with accepted practices for the respective organizations, groups, etc. In addition, it allowed us to address any concerns at the outset. Based on these discussions, we constructed memorandums of understanding (MOUs) to specify the nature of the relationship and outline collaborative activities with the stakeholders from different sectors. MOUs were signed with OPPEA, the Puerto Rican Chapter of the AARP and Lutheran Social Services of PR. In addition, we established MOUs with Mayors and their staff from several municipalities, including Cidra, Fajardo, Carolina, Aguadilla, and Arecibo. As part of the MOU, the Universidad Central del Caribe provided insurance endorsements for the use of their venues during recruitment events.

Ascertainment Phase Three: Targeted Actions and Direct Recruitment

Working with the various groups with whom we had MOUs, we set up multiple recruitment events. Depending on the site, pre-recruitment conferences were scheduled to educate center personnel (e.g., primary doctors, nurses, social workers, psychologist, and others dementia specialists) or the public. These pre-recruitment meetings were used to provide general information about AD and to clarify aspects of the study in person to healthcare providers as well as potential participants and their families. At meetings involving the public, potential participants, or family members we gathered contact information for further follow up, leading to recruitment of interested individuals. This also allowed us to estimate the number of participants and to plan our ascertainment resources accordingly.

Ascertainment Phase Four: Giving Back: Dissemination and Education

We conducted a number of follow-up events to provide information for caregivers and center personnel at the various recruitment sites. For physicians, we were able to provide continuing medical education through the Puerto Rican College of Physicians and Surgeons; for health professional staff, we provided participation certificates for early detection of AD and culturally relevant adaptation of the comprehensive and evidence-based community support strategies.

This follow-up allowed us to disseminate information about AD to the community. The provision of information about AD to non-AD healthcare workers and general communities will help us build local resource networks and empower them with knowledge about dementia capabilities to improve the quality of life of the participants and their caregivers. In addition, at select venues we have also organized educational outreach activities where we served as expert speakers, providing information about dementia research and care. Typical audiences included healthcare providers (e.g., nurses, social workers, case managers, and primary care physicians) and the public. We have also engaged in dementia-related initiatives via social media, like "Un café por el Alzheimer" (A cup of coffee for Alzheimer) (Friedman et al., 2016), which shares our study information on their social media platforms.

Study Population

A convenience sampling method with a geographic distribution throughout the island was used. PRADI participants were self-reported Puerto Rican adults, aged 50 years, and older with no restrictions on gender or socioeconomic status. While the majority of participants were residents of PR, a small fraction of relatives of the Puerto Rican families living in the continental United States (Florida, New York, Connecticut, and Massachusetts) were enrolled. In addition, some individuals less than 50 years of age were enrolled. When conducting our analyses, we included only residents of PR who were 50 years of age or greater.

Our cohort is further specified based on seven health regions as defined by the PR Department of Health[1]. These seven regions contain multiple municipalities and place this cohort in the context of the previously established health related structure. Each of the health regions is labeled by the major municipality within each region (with the exception of the Metro region). As seen in **Figure 1**, the most heavily populated areas per the 2010 census are the Metro, Caguas, and Bayamón regions, containing 22, 16, and 16% of the total population, respectively.

[1]http://www.salud.gov.pr/Pages/Regiones-de-Salud-y-Servicios-Directos.aspx

Per the same census period, ∼15% of individuals in PR were over 65 years of age.

Ascertainment Sources

All participants were ascertained via three main sources: AD specialists, adult care centers, and community events. This approach allowed us to capture a wide range of AD cases from varied socioeconomic backgrounds and education levels. All individuals were recruited using site-specific IRB approved protocols.

AD Specialists

Several AD specialists (neurologists, psychiatrists, and geriatricians) served as collaborators and referred patients who met inclusion criteria and were interested in participating in the PRADI study. These included patients with AD, mild cognitive impairment (MCI), and dementia. As described below in the screening and evaluation section, we obtained clinical and medical records for patients who were recruited via AD specialists.

AD Centers and Adult Care Centers

To date, we have recruited participants from seven AD dedicated centers and advanced age nursing homes across the island, identified through the OPPEA directory of services website. The AD centers and nursing homes serve between 20 and 40 individuals who are typically older than 60 years of age (with or without the diagnosis of AD) on a daily basis. These centers focus on providing therapeutic, social, and recreational activities to improve quality of life, as well as educating, and supporting caregivers or family members.

Community Groups

We conducted recruitment events in various municipalities. Typically, these recruitment events were preceded by a pre-recruitment event. The actual recruitment visits were then conducted at various centers or in private spaces. During these events, our multi-disciplinary teams consented participants (or their proxies), conducted cognitive screenings, and drew blood samples. These events ranged in size from small venues that attracted 20 or so individuals to much larger events that drew 60 or more individuals. We were able to enroll cases and controls during these events.

Inclusion/Exclusion Criteria

Participants were enrolled in the following categories: cases (AD and MCI), unaffected family members of cases, or unrelated individuals with no cognitive problems. To be enrolled, participants had to meet basic inclusion criteria. All individuals had to: (a) be of Puerto Rican ancestry (with at least one grant-parent born on the island); (b) be $\geq$50 years of age; and (c) be willing to participate (or, in cases of serious cognitive impairment, have family members who consent on their behalf) and provide informed consent or have a proxy for consent.

To be included as a case, we required that individuals have a previous clinical diagnosis of AD, MCI, dementia, or show evidence of a memory disorder, and meet standard criteria for AD or MCI (McKhann et al., 1984; Albert et al., 2011; McKhann et al., 2011). We included cases from families (family history positive) as well as sporadic or isolated cases (family history negative). We excluded individuals whose memory and cognitive problems are secondary to other causes (e.g., stroke, psychoses, etc.) and those with a known mutation (e.g., PS1, PS2, or APP).

To be included as a control, individuals had to meet basic inclusion criteria, have no prior clinical diagnoses of a memory disorder or subjective memory complaints, demonstrate no cognitive problems on neurocognitive screening and assessment, and be unrelated to our cases. Unaffected family members had to meet the same inclusion criteria as the controls in addition to being a first- or second-degree relative of a case. For unaffected family members, we typically included the oldest available individual.

Screening and Evaluation

For participants enrolled as cases (i.e., with suspected memory problems or known dementias), we conducted a detailed chart review during which we corroborated clinical diagnoses and extracted current and past medical histories, current and past medications, family histories (pedigrees), and sociodemographic information. In addition, we collected clinical neurologic and neuropsychological test data, neuroimaging results, and pertinent lab values (e.g., hematology, thyroid function, lipid profile, vitamin D and B12 levels, and liver function tests, hypothyroidism, and vitamin deficiency).

For presumptive cases, we conducted an initial screening with the Modified Mini-Mental State Examination (3MS) (Folstein et al., 1975; Teng and Chui, 1987) followed by a cognitive evaluation that included the NIA-LOAD cognitive battery (Morris et al., 2006; Weintraub et al., 2009). In addition, we administered the Clinical Dementia Rating Scale (CDR) (Yesavage, 1988). Individuals who were deemed cognitively normal were screened with the 3MS (Folstein et al., 1975; Teng and Chui, 1987) and the CDR. For most cognitively normal individuals, we administered the NIA-LOAD battery.

Adjudication

All clinical, historical and screening/evaluation test data (e.g., laboratory tests, neurologic examination, neuroimaging, and neuropsychological screen and testing) from individuals with a known or suspected dementia were reviewed by a clinical adjudication panel consisting of a neurologist, neuropsychologist, and clinical staff. The panel reviewed all data and assigned best-estimate diagnoses. To be classified as AD individuals had to meet the current NIA-AA criteria (McKhann et al., 2011). They were further classified as definite (neuropathologic confirmation), probable, or possible AD. Diagnoses of MCI were assigned using the NIA-AA criteria (Albert et al., 2011). Cognitively normal individuals with no history of memory problems and MMSE or 3MS scores that fall above clinical cutoffs were designated as unrelated controls for the study. Family-based controls were evaluated similarly for inclusion in family-based analyses (Beecham et al., 2017). In the course of adjudication meetings, team members discussed cases until a diagnostic classification was determined. For those cases in

which the team was unable to arrive at a final decision, the team stipulated the reason and corrective actions were taken (e.g., obtaining a more detailed history, retesting, etc.) In the event of a disagreement, the team consulted with an independent dementia specialist.

Analysis

To test for possible differences in our cohort related to where participants live and how they were ascertained, we compared mean 3MS scores and mean age of onset (AAO) for cases by region and recruiting source. Cases consisted of both AD and MCI phenotypes. In addition, for our controls we were able to compare mean 3MS scores by region. All analyses were performed using one-way ANOVA in SAS and SPSS (SAS Institute Inc., 2011; SPSS, 2013). *P* values lower than 0.05 were considered statistically significant.

In addition, we conducted an admixture analysis to estimate the proportions of admixture (European, African, and Native American) in our cohort. Genotyping and quality control methods are described elsewhere (Alexander et al., 2009; Rajabli et al., 2018). Briefly, genotyping was performed on the Expanded Multi-Ethnic Genotyping Array and Global Screening Array (Illumina, San Diego, CA, United States) and quality was assessed using PLINK software, v.2. Using the reference panels (African, European, and Native American populations) from the Human Genome Diversity Project3, we conducted admixture analysis, using ADMIXTURE software (Alexander et al., 2009; Rajabli et al., 2018), to generate average ancestry proportions across PR's seven health regions.

RESULTS

We have enrolled 770 individuals over a 30-month period, 710 of which were from PR. After removing individuals <50 years of age (35 unaffected, 1 MCI), our current dataset consisted of 674 individuals. The distribution of enrollment across the seven health regions of PR, as seen in **Figure 1**, shows the heaviest ascertainment in the Metro (44%; N = 295), Caguas (20%; N = 134), and Bayamón (16%; N = 106) regions, which reflects the greater population densities of these regions and cities. Enrollment numbers for the seven health regions are presented in **Table 1**, which also provides the numbers for the respective municipalities within those health regions.

Among these 674 individuals, 282 (42%) were ascertained as AD, 115 (17%) were ascertained as MCI, and the remaining 277 (41%) were ascertained as unaffected. The majority of our cases (83%) had an age of onset ≥65 years of age. The greatest numbers of AD (N = 111; 39%) and MCI (N = 61; 53%) were ascertained in the Metro region. Equally high ascertainment numbers were also observed in Bayamón and Caguas (AD N = 54, 19%; MCI N = 17, 15%).

Participants were recruited from three sources: AD specialists (N = 261, 39%), adult care centers (N = 201, 30%), and community events (N = 202, 30%). Not surprisingly, as seen in **Table 2**, most of the AD cases were recruited via the AD specialist, while the largest number of MCI cases were ascertained through community events. **Figure 2** provides additional information regarding enrollment sources per the respective health regions.

TABLE 1 | Ascertainment by health regions and municipalities (N = 674).

	Ponce	Arecibo	Mayagüez	Metro	Bayamón	Caguas	Fajardo
Population (%)	565,683 (15%)	456,036 (12%)	535,488 (14%)	822,562 (22%)	620,110 (16%)	589,795 (16%)	136,115 (4%)
Ascertainment (%)	23 (3%)	41 (6%)	49 (7%)	295 (44%)	106 (16%)	134 (20%)	26 (4%)
Ascertainment by municipalities (n)	Ponce (n = 15)	Arecibo (n = 15)	Aguadilla (n = 31)	San Juan (n = 157)	Bayamón (n = 33)	Cidra (n = 45)	Fajardo (n = 15)
	Juana Díaz (n = 3)	Manatí (n = 10)	Mayagüez (n = 7)	Carolina (n = 98)	Naranjito (n = 12)	Cayey (n = 27)	Luquillo (n = 6)
	Yauco (n = 3)		San Germán (n = 3)	Guaynabo (n = 28)	Orocovis (n = 12)	Caguas (n = 16)	Ceiba (n = 2)
	Adjuntas (n = 1)		Lajas (n = 3)	Trujillo Alto (n = 11)	Toa Alta (n = 12)	Naguabo (n = 16)	Rio Grande (n = 2)
	Guayama (n = 1)		Hormigueros (n = 2)	Canovanas (n = 1)	Toa Baja (n = 12)	Humacao (n = 8)	Vieques (n = 1)
			Isabela (n = 2)	Loíza (n = 1)		San Lorenzo (n = 6)	
			San Sebastián (n = 1)	Rio Piedras (n = 1)		Gurabo (n = 4)	
						Juncos (n = 4)	
						Yabucoa (n = 4)	
						Las Piedras (n = 2)	
						Maunabo (n = 2)	

TABLE 2 | Ascertainment by source (*N* = 664)*.

	AD specialist	Adult care center	Community events	Total
AD *N*	136	86	55	277
%	(49%)	(31%)	(20%)	
MCI *N*	43	17	53	113
%	(38%)	(15%)	(47%)	
UNAFF *N*	82	98	94	274
%	(30%)	(36%)	(34%)	
Total	261	201	202	664

*Our total ascertainment=674; 10 individuals were missing source data, A = Alzheimer disease; MCI = Mild cognitive impairment; UNAFF = Unaffected.

Finally, our cohort can be further delineated by whether individuals were part of a family or ascertained as an isolated/sporadic case. Of the 43 multiplex families that have been completed to date, 10 families contain four or more living individuals with AD, 3 families contain 3 living individuals with AD, and 31 families contain 2 living individuals with AD. The mean number of LOAD cases per multiplex family is 3.9. Among the 198 individuals from those multiplex families 73 (37%) meet the criteria for LOAD, 19 (9%) meet the criteria for EOAD, 31 (16%) meet the criteria for MCI, and 75 (38%) meet the criteria for no cognitive problems.

Admixture Results

We examined the population structure of Puerto Ricans using the supervised ADMIXTURE analysis at $K = 3$. **Figure 3A** illustrates the results from the ADMIXTURE analysis in a bar-plot figure. Each vertical bar represents an individual and corresponding estimates of the fraction of continental ancestries (African, European, and Native American). On average, Puerto Ricans have mostly European ancestry with a mean value of 69.3% (*SD* = 12.2). Mean values for African and Native American ancestry are 17.3% (*SD* = 12.2) and 13.4% (*SD* = 4.2), respectively as seen in the box plots (**Figure 3B**).

Figure 4A illustrates the bar-plots of admixed individuals across the Puerto Rican health zones and shows heterogeneous admixture patterns. Results of the admixture analysis are in general agreement with recent genetic studies showing a three-way admixture (European, African, and Native American) structure in Puerto Ricans (Via et al., 2011).

We observed a non-uniform distribution of European and African ancestral backgrounds across the health regions with relatively high European and low African ancestral proportions in Mayagüez, Ponce, and Bayamón (**Figure 4B**). The average European and African ancestry fractions in these zones are 74.4% (*s* = 5.8), 74% (*SD* = 8.1), 73.3% (*SD* = 8.7) and 11.9% (*SD* = 6.6), 11.6% (*SD* = 5.9), 11.0% (*SD* = 4.9), respectively. In contrast, the Native American ancestral background shows nearly uniform distribution across the geographical zones (**Figure 4B**).

Clinical Comparisons

Separate one-way ANOVAs were conducted to test if mean values for AAO and the 3MS differed by (a) ascertainment region (i.e., the seven health regions of PR) and (b) ascertainment source (AD specialist, adult care center, and community).

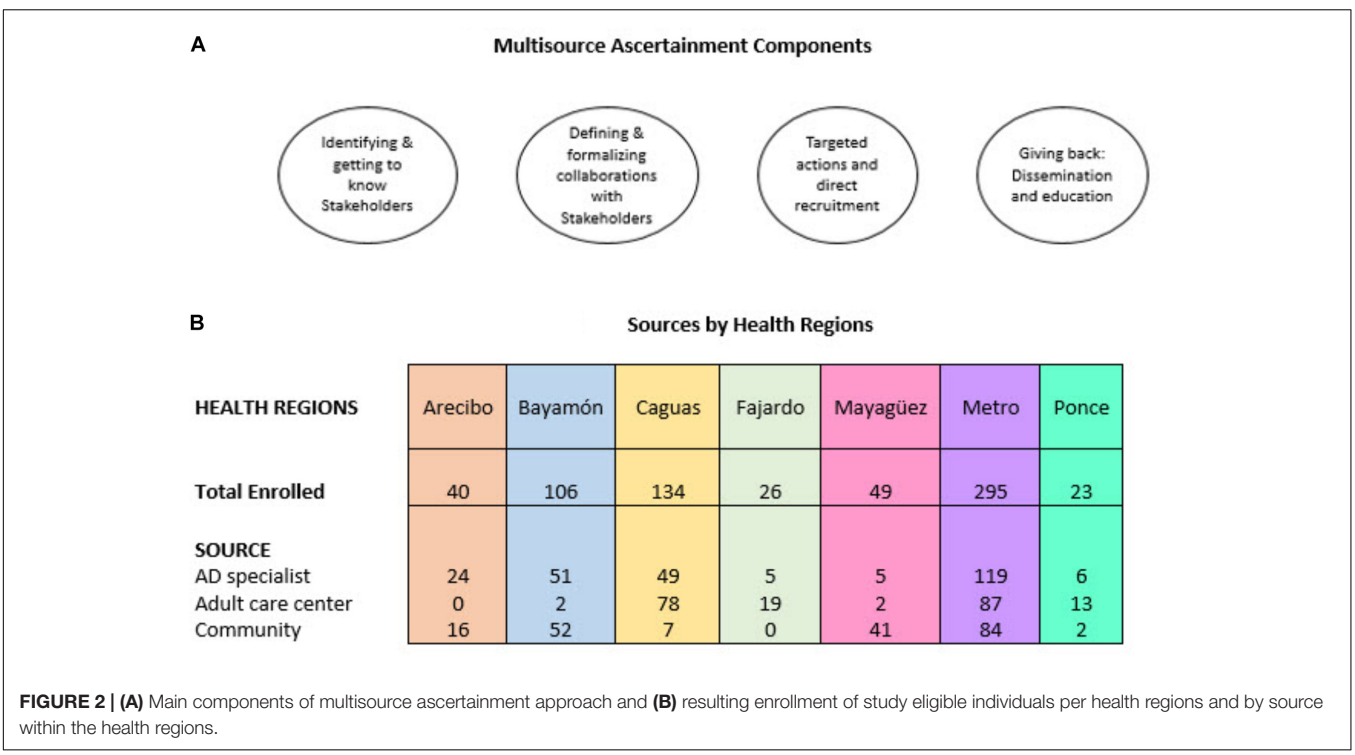

FIGURE 2 | (A) Main components of multisource ascertainment approach and **(B)** resulting enrollment of study eligible individuals per health regions and by source within the health regions.

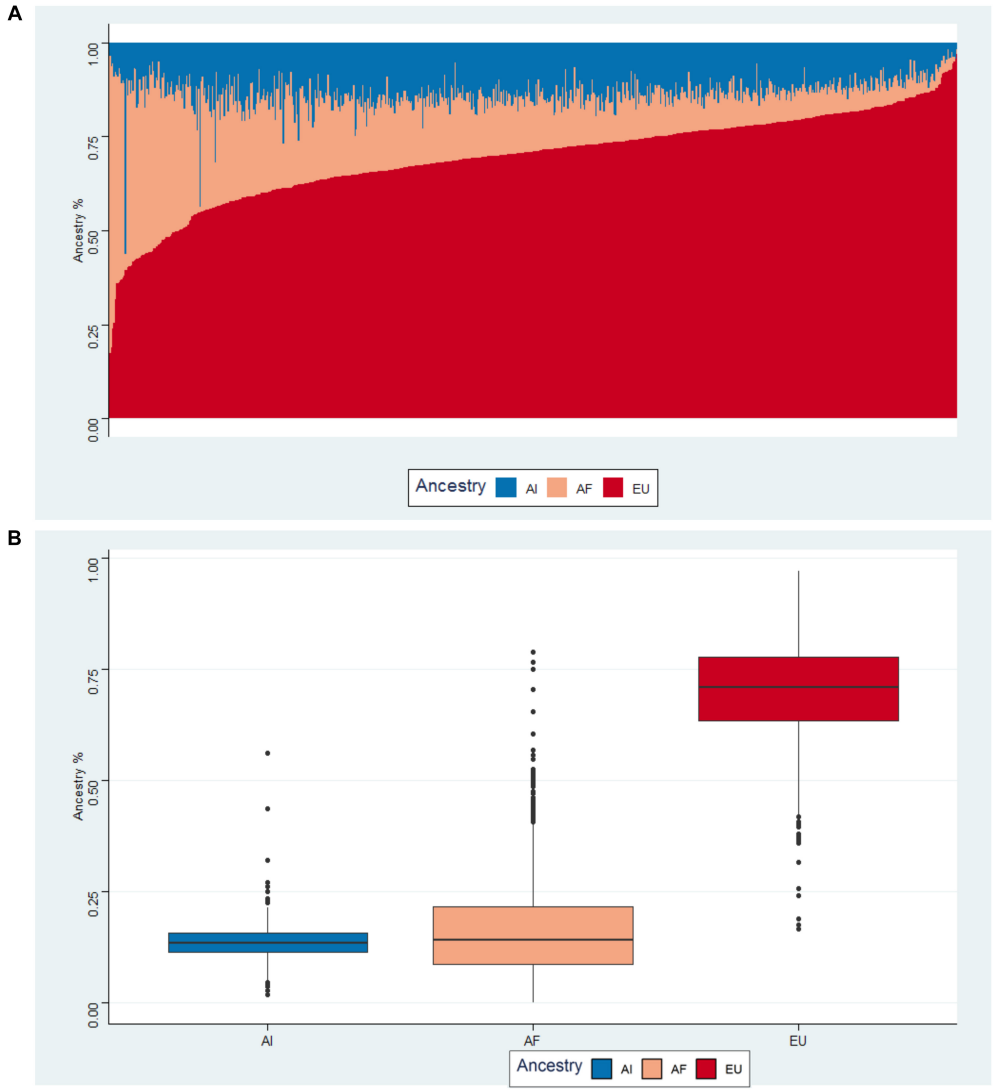

FIGURE 3 | (A) Bar-plot of three way admixed Puerto Rican individuals estimated using ADMIXTURE software at $K = 3$. **(B)** Box-plot of average ancestries by health region.

Age at Onset (AAO)

The mean AAO values for our AD and MCI case were 74.1 (*SD* = 9.4) and 71.2 (*SD* = 8.5), respectively. As noted above, for the purposes of analysis we combined these into one group (cases) which had a mean AAO value of 73.2 (*SD* = 9.2). The mean values for AAO for the seven ascertainment regions and sources are shown in **Table 3**.

Across the different regions, mean AAO values ranged from 70.3 (*SD* = 7.4) in Mayagüez to 75.9 (*SD* = 9.6) in Fajardo. Results of one-way ANOVA found no statistically significant differences in AAO across the different health regions $F(6,385) = 0.92$, $p = 0.48$. The mean AAO values for the three ascertainment sources ranged from 70.6 (*SD* = 4.6) for AD specialists to 76.4 (*SD* = 9.0) for cases ascertained through adult care centers. The results of the one-way ANOVA found significant group differences among the three ascertainment sources $F(2,382) = 16.29$ $p < 0.001$. *Post hoc* tests showed mean AAO was higher in patients recruited from the community sites (+4.1 years) and adult care centers (+6.0 years) than it was for patients ascertained from AD specialists.

Modified Mini Mental State Examination (3MS)

The mean 3MS scores for our AD and MCI cases were 52.6 (*SD* = 23.5) and 80.1 (*SD* = 12.2), respectively; the overall mean 3MS score for all cases was 63.5 (*SD* = 24). The mean 3MS scores for the seven ascertainment regions and sources are seen in **Table 3**.

Among the health regions, mean 3MS scores ranged from 46.3 (*SD* = 28) in Mayagüez to 69.5 in the Metro region (*SD* = 19.9). Note that we dropped the Fajardo region, as there were only three 3MS scores. For these comparisons, the

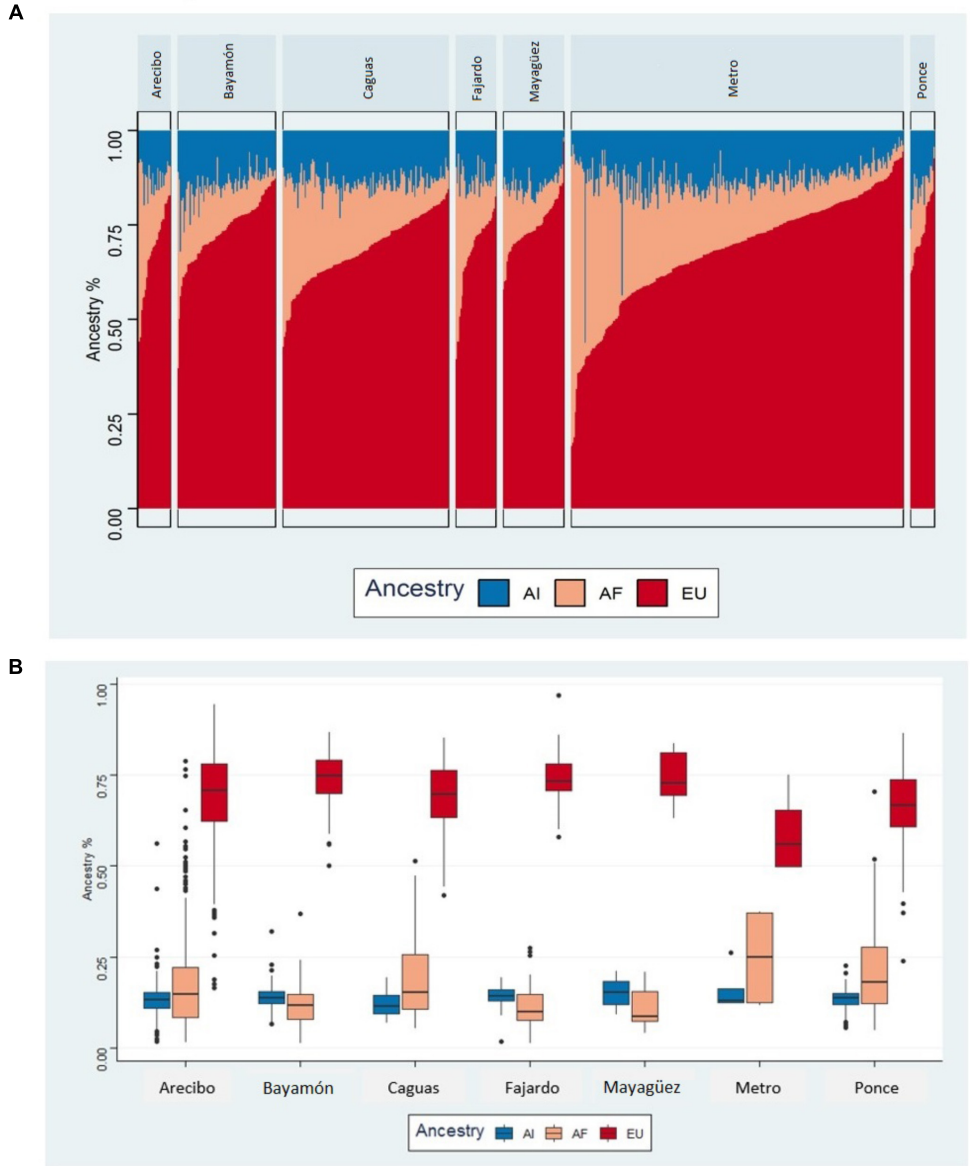

FIGURE 4 | (A) Bar-plot of three way admixed Puerto Rican individuals for each of the health regions estimated using ADMIXTURE software at $K = 3$. **(B)** Box-plot of average ancestries by health region.

homogeneity of variances assumption was violated, as assessed by Levene's Test of Homogeneity of Variance ($p = 0.008$). The one-way Welch ANOVA results show statistically significantly differences in mean 3MS scores between the health regions Welch's $F(5,52.96) = 3.81$, $p = 0.005$. Games-Howell *post hoc* analysis revealed only one statistically significant comparison ($p < 0.01$) between the Metro and Mayagüez regions (23.3+5.8) [mean ± standard error]. For source, the mean values ranged from 63.1 ($SD = 24$) for cases ascertained via the community to 64.2 ($SD = 27.6$) for cases ascertained through AD specialists. Again, Levene's Test of Homogeneity of Variance was significant ($p = 0.03$) indicating that the homogeneity of variances assumption was violated, prompting use of Welch's ANOVA. Results of one way ANOVA found no statistically significant differences in 3MS means Welch's $F(2,130.5) = 0.04$, $p = 0.96$.

DISCUSSION

Using a multisource approach that emphasized community engagement and was tailored to the Puerto Rican population, we were able to enroll eligible participants and their family members across PR. A major feature of our community engagement efforts was the development of partnerships with leaders of health initiatives in municipalities and resources within those municipalities. These included the health department,

TABLE 3 | Mean age at onset (AAO) and 3MS values for cases (*N* = 397) by region and source (note sample sizes reflect missing values).

	Age at onset		3MS	
REGION	***N***	**Mean (*SD*)**	***N***	**Mean (*SD*)**
Arecibo	27	71.6 (8.1)	17	56.7 (27.4)
Bayamón	69	73.1 (9.7)	36	62.9 (19.8)
Caguas	69	74.5 (9.9)	35	63.3 (28.8)
Fajardo	15	75.9 (9.6)	3*	70.0 (10.1)*
Mayagüez	22	70.3 (7.4)	19	46.3 (28.0)
Metro	172	73.1 (9.0)	101	69.5 (19.9)
Ponce	18	73.4 (9.7)	14	51.5 (24.6)
TOTAL	392	73.2 (9.2)	225	63.4 (68.4)
SOURCE	***N***	**Mean (*SD*)**	***N***	**Mean (*SD*)**
AD specialist	176	70.5 (9.0)	79	64.2 (27.6)
Adult care center	102	76.5 (9.0)	48	63.3 (18.9)
Community	107	74.6 (7.9)	92	63.1 (23.3)
TOTAL	385	73.2 (9.3)	219	63.5 (23.0)

**Removed from analysis of 3MS score by region.*

governmental organizations, community-based organizations, religious groups, and various healthcare providers. Establishing strong community partnerships allowed us to develop strategies with input from different parts of the community to achieve an ascertainment approach that was sensitive to the local culture.

Our multisource approach emphasizes community engagement beginning with the identification of and establishment of relationships with key stakeholder groups and organizations. This allowed us to develop mutually agreed upon ways to implement research activities and create memorandums of understanding to formalize implementation. Working with these stakeholders and organizations enabled us to conduct outreach and ascertainment activities in the respective municipalities. Concurrent with the outreach activities and recruiting events (and as a way of giving back to the communities), we provided information and educational opportunities to healthcare providers and the public. This community engagement approach, developed for PRADI by AD clinicians and researchers in Puerto Rico and Miami, is a platform for our ongoing ascertainment efforts.

Using this approach, we have enrolled 674 individuals from PR over the age of 50 for our PRADI study. These individuals were recruited fairly evenly from the three ascertainment sources are and concentrated in the three health regions with the largest numbers of individuals – Metro, Bayamón, and Caguas. We also observed that the main ascertainment sources varied by the health regions, reflecting different resources in the respective regions. Further, while the percentage of individuals ascertained in select regions paralleled the percentage of the total population for the region, the Metro and Ponce regions were disparate as 44% of our participants were ascertained in the Metro region which constitutes 22% of the population vs. 3% of our participants were ascertained in the Ponce region which constitutes 14% of the population. These ascertainment figures have already begun to inform our subsequent recruitment efforts, as we emphasized the need to engage other sectors of PR (e.g., Ponce).

The importance of recruiting in regions such as Ponce and Mayagüez is also reflected in the results of our admixture analysis showing differences in the proportion of European and African ancestry among individuals from these regions. The failure to ascertain participants from regions with different ancestral backgrounds could potentially limit the applicability of important findings to these groups. The significance of this for the PRADI study is reinforced by work showing that different ancestral backgrounds may play a significant role in modifying the effect of APOE on risk for AD (Rajabli et al., 2018). These results are preliminary and will need further investigation, in particular to specify area of origin for participants vs. current area.

In addition to potential ancestral differences across the different regions, we observed clinical differences in our cohort in relation to ascertainment region and sources. For instance, participants' mean 3MS scores varied by ascertainment region although the only significant difference was between the Mayaguez and Metro regions. This may reflect differences in the sources of these participants as most of the individuals from Mayaguez were ascertained in the community. While there were no significant differences in AAO among participants from these different regions, we observed that AAO varied according to ascertainment source. Specifically, individuals who had been seen by AD specialists were more likely to have been identified as having cognitive/memory problems at younger ages. Aside from differences in sample size, the observed differences in AAO and 3MS values by ascertainment region and source most likely reflect the complex interplay of multiple influences, including access to AD specialists, availability of dementia related resources, and general knowledge and acceptance of AD.

The influence of knowledge and acceptance of AD is an important issue that is intertwined with efforts to recruit and enroll participants for genetic studies of AD in PR. While genetic studies of AD in PR have been undertaken by several groups as part of a larger emphasis on understanding AD in Caribbean Hispanics (Lee et al., 2006; Barral et al., 2015), the ascertainment approach developed for PRADI focuses solely on the island and intends to create a program that enhances knowledge of AD in PR.

Efforts to increase knowledge of AD in PR have grown recent years and the multisource approach to recruitment and enrollment is aligned with programs such as the *Un Café por el Alzheimer* program in PR, which provides education about AD at coffee shops and through social media (Friedman et al., 2016). The educational component that we include as part of our larger ascertainment approach is crucial for providing information about AD to healthcare providers and the public across the various communities and will potentially impact participation in biomedical research, including genetic studies (Karlawish et al., 2011).

The goal of the PRADI study is to investigate the genetics of AD in Puerto Ricans. AD is a complex disease with substantial burden on the population – particularly in PR where there is a large aging population suffering from chronic diseases that may exacerbate existing risk (Perreira et al., 2017). To date, there

have been a scarcity of genetic studies of complex traits (e.g., AD) in Puerto Ricans which could exacerbate existing health disparities. Exceptions to this are the Boston Puerto Rican Health Study (BPRHS), a longitudinal cohort study which examines non-genetic, and genetic influences on multiple health outcomes among mainland Puerto Ricans (Tucker et al., 2010) and the Hispanic Community Health Study (HCHS), a large longitudinal multi-cohort project which studies a variety of health outcomes among different Hispanic-Latino groups in the US, including Puerto Ricans (Lavange et al., 2010) – both of which have extensive phenotypic and genotypic data. Using data from these cohorts, investigators have found links between select genes, obesity and asthma (Guo et al., 2018), lipid profiles (Graff et al., 2017), and blood pressure traits (Sofer et al., 2017). A large amount of research has genetic factors contributing to asthma and other pulmonary traits which are a major health problem in Puerto Ricans. The involvement of Puerto Ricans in this work can lead to greater understanding of genetic contributions to disease in this population and intervention opportunities. Central to the success of this research is ensuring participation in this research (Karlawish et al., 2011).

Our results suggest the importance of engaging multiple stakeholders and communities across municipalities. Including stakeholders in the development of outreach and recruitment was an important part of the PRADI ascertainment approach. Another important aspect of our ascertainment approach was the provision of AD and dementia information to providers, care centers, and the public. While our ascertainment results cannot be directly attributed to our multisource approach we have preliminary data that can guide more systematic evaluation of what works best as the PRADI study moves forward. Ultimately, this study and others like it are intended to inform and improve health outcomes and reduce health disparities for Puerto Ricans and other Hispanic Latino populations who have been consistently underserved.

ETHICS STATEMENT

This study was carried out in accordance to the recommendations of the National Institute of Health Guiding Principles for Ethical Research Pursuing Potential Research Participants Protection and the 2016 National Institute of Health Single Review Board (sIRB) Policy. This study received ethical approval from University of Miami Institutional Review Board (approved protocol #20070307) and Universidad Central del Caribe Institutional Review Board (approved protocol # 2016-26). The Universidad Central del Caribe is relying on the designated UM-IRB by an Institutional Review Board Authorization Agreement (Protocol: Genetic Studies in Dementia). All subjects (participant or proxy) gave written informed consent. This study was carried out in accordance with the Declaration of Helsinki and amendments.

AUTHOR CONTRIBUTIONS

MC helped with study design, assisted with clinical adjudication of patient and control data, and wrote and proofread the manuscript. BF-A and KC assisted with study design, ascertainment, and clinical adjudication of patient and control data, and wrote and proofread the manuscript. JR and FR performed statistical analyses and helped to writing the manuscript. LA and JV helped with study design, ascertainment, and clinical adjudication of patient and control data. PB, PRM, AR, and VR helped with ascertainment and clinical adjudication of patient and control data. CS, PM, AG, MP, and JM helped with ascertainment of patient and control data. KH-N compiled data for the publication and ran clinical queries. NF helped with ascertainment of patient and control data, and proofread the manuscript. AC and HA helped with ascertainment of patient and control data, diagnosis, and adjudication. GB conceived of and implemented the study design. MP-V conceived of and implemented the study design, assisted with ascertainment and clinical adjudication of patient and control data, and helped to writing the manuscript.

FUNDING

Financial support for the research, authorship, and publication of this article was provided by the grant "Genomic Characterization of Alzheimer's Disease Risk in the Puerto Rican Population" (1RF1AG054074-01) from the National Institute of Health (NIH) and National Institute on Aging (NIA).

ACKNOWLEDGMENTS

We wish to acknowledge the community, faith, and government organizations, as well as the healthcare professionals, and individuals who participated and collaborated in this research project. We are grateful to the families and staff who participated in this study. We also gratefully acknowledge the resources provided by the John P. Hussman Institute for Human Genomics and the Universidad Central del Caribe.

REFERENCES

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., et al. (2011). The diagnosis of mild cognitive impairment due to alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimers Dement* 3, 270–279. doi: 10.1016/j.jalz.2011.03.008

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 9, 1655–1664. doi: 10.1101/gr.094052.109

Amezcua, L., Beecham, A. H., Delgado, S. R., Chinea, A., Burnett, M., Manrique, C. P., et al. (2018). Native ancestry is associated with optic neuritis and age of onset in hispanics with multiple sclerosis. *Ann. Clin. Transl. Neurol.* 11, 1362–1371. doi: 10.1002/acn3.646

Armstrong, K., Micco, E., Carney, A., Stopfer, J., and Putt, M. (2005). Racial differences in the use of BRCA1/2 testing among women with a family history of breast or ovarian cancer. *JAMA* 14, 1729–1736.

Armstrong, K., Putt, M., Halbert, C. H., Grande, D., Schwartz, J. S., Liao, K., et al. (2012). The influence of health care policies and health care system distrust on willingness to undergo genetic testing. *Med. Care* 5, 381–387. doi: 10.1097/MLR.0b013e31824d748b

Barral, S., Cheng, R., Reitz, C., Vardarajan, B., Lee, J., Kunkle, B., et al. (2015). Linkage analyses in Caribbean Hispanic families identify novel loci associated with familial late-onset Alzheimer's disease. *Alzheimers Dement* 11, 1397–1406. doi: 10.1016/j.jalz.2015.07.487

Beecham, G., Bis, J., Martin, E., Choi, S., DeStefano, A., and van Duijn, C. (2017). The Alzheimer's disease sequencing project: study design and sample selection. *Neurol. Genet.* 3:e194.

Calderon, J. L., Baker, R. S., Fabrega, H., Conde, J. G., Hays, R. D., Fleming, E., et al. (2006). An ethno-medical perspective on research participation: a qualitative pilot study. *MedGenMed* 2:23.

Claudio-Campos, K., Orengo-Mercado, C., Renta, J. Y., Peguero, M., Garcia, R., Hernandez, G., et al. (2015). Pharmacogenetics of healthy volunteers in Puerto Rico. *Drug Metab. Pers. Ther.* 4, 239–249. doi: 10.1515/dmpt-2015-0021

Diaz-Zabala, H. J., Ortiz, A. P., Garland, L., Jones, K., Perez, C. M., Mora, E., et al. (2018). A recurrent BRCA2 mutation explains the majority of hereditary breast and ovarian cancer syndrome cases in puerto rico. *Cancers* 11:E419. doi: 10.3390/cancers10110419

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). Mini-mental state. A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 3, 189–198.

Friedman, D. B., Gibson, A., Torres, W., Irizarry, J., Rodriguez, J., Tang, W., et al. (2016). Increasing community awareness about Alzheimer's Disease in Puerto Rico through coffee shop education and social media. *J. Commun. Health* 5, 1006–1012. doi: 10.1007/s10900-016-0183-9

Gignoux, C. R., Torgerson, D. G., Pino-Yanes, M., Uricchio, L. H., Galanter, J., and Roth, L. A. (2019). An admixture mapping meta-analysis implicates genetic variation at 18q21 with asthma susceptibility in Latinos. *J. Allergy Clin. Immunol.* 3, 957–969. doi: 10.1016/j.jaci.2016.08.057

Graff, M., Emery, L. S., Justice, A. E., Parra, E., Below, J. E., Palmer, N. D., et al. (2017). Genetic architecture of lipid traits in the Hispanic community health study/study of Latinos. *Lipids Health Dis.* 16:200. doi: 10.1186/s12944-017-0591-6

Guo, Y., Moon, J. Y., Laurie, C. C., North, K. E., Sanchez-Johnsen, L. A. P., Davis, S., et al. (2018). Genetic predisposition to obesity is associated with asthma in US Hispanics/Latinos: results from the hispanic community health study/study of latinos. *Allergy* 7, 1547–1550. doi: 10.1111/all.13450

Hampel, H., O'Bryant, S. E., Durrleman, S., Younesi, E., Rojkova, K., Escott-Price, V., et al. (2017). A precision medicine initiative for alzheimer's disease: the road ahead to biomarker-guided integrative disease modeling. *Climacteric* 2, 107–118. doi: 10.1080/13697137.2017.1287866

Karlawish, J., Barg, F. K., Augsburger, D., Beaver, J., Ferguson, A., and Nunez, J. (2011). What Latino Puerto Ricans and non-Latinos say when they talk about Alzheimer's disease. *Alzheimers Dement* 2, 161–170. doi: 10.1016/j.jalz.2010.03.015

Lambert, J. C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., et al. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* 12, 1452–1458. doi: 10.1038/ng.2802

Lavange, L. M., Kalsbeek, W. D., Sorlie, P. D., Aviles-Santa, L. M., Kaplan, R. C., Barnhart, J., et al. (2010). Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. *Ann. Epidemiol.* 8, 642–649. doi: 10.1016/j.annepidem.2010.05.006

Lee, J. H., Cheng, R., Santana, V., Williamson, J., Lantigua, R., Medrano, M., et al. (2006). Expanded genomewide scan implicates a novel locus at 3q28 among Caribbean hispanics with familial Alzheimer disease. *Arch. Neurol.* 11, 1591–1598.

McKhann, G., Drachman, D., and Folstein, M. (1984). Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA work group under the auspices of the department of health and human services task force on Alzheimer's Disease. *Neurology* 34, 939–944.

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R. Jr., Kawas, C. H., et al. (2011). The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* 3, 263–269. doi: 10.1016/j.jalz.2011.03.005

Morales-Borges, R. H. (2017). Need for pharmacogenetic studies on the prevalence of MTHFR mutations in Puerto Ricans and Hispanics. *Drug Metab. Pers. Ther.* 3, 169–171. doi: 10.1515/dmpt-2017-0010

Morris, J. C., Weintraub, S., Chui, H. C., Cummings, J., Decarli, C., Ferris, S., et al. (2006). The uniform data set (uds): clinical and cognitive variables and descriptive data from alzheimer disease centers. *Alzheimer Dis. Assoc. Disord.* 4, 210–216. doi: 10.1097/01.wad.0000213865.09806.92

Perreira, K. M., Lallemond, N., Napoles, A., and Zuckerman, S. (2017). *Puerto Rico's Health Care Infrastructure Assessment: Site Visit Report*. Washington, DC: Urban Institute.

Puerto Rico Department of Health (2015). *Resumen general de la salud en Puerto Rico*. San Juan: Puerto Rico Department of Health.

Rajabli, F., Feliciano, B. E., Celis, K., Hamilton-Nelson, K. L., Whitehead, P. L., Adams, L. D., et al. (2018). Ancestral origin of ApoE epsilon4 Alzheimer disease risk in Puerto Rican and African American populations. *PLoS Genet.* 12:e1007791. doi: 10.1371/journal.pgen.1007791

Ramirez, A. G., Miller, A. R., Gallion, K., San Miguel de Majors, S., Chalela, P., and Garcia Aramburo, S. (2008). Testing three different cancer genetics registry recruitment methods with Hispanic cancer patients and their family members previously registered in local cancer registries in Texas. *Commun. Genet.* 4, 215–223. doi: 10.1159/000116882

Ricker, C., Lagos, V., Feldman, N., Hiyama, S., Fuentes, S., Kumar, V., et al. (2006). If we build it. will they come?–establishing a cancer genetics services clinic for an underserved predominantly Latina cohort. *J. Genet. Counsel.* 6, 505–514. doi: 10.1007/s10897-006-9052-5

Salgado-Montilla, J. L., Rodriguez-Caban, J. L., Sanchez-Garcia, J., Sanchez-Ortiz, R., and Irizarry-Ramirez, M. (2017). Impact of FTO SNPs rs9930506 and rs9939609 in prostate cancer severity in a cohort of Puerto Rican Men. *Arch. Cancer. Res* 5:148. doi: 10.21767/2254-6081.1000148

SAS Institute Inc. (2011). *SAS 9.3 System Options: Reference*. Cary, NC: SAS Institute Inc.

Shavers, V. L., Lynch, C. F., and Burmeister, L. F. (2002). Racial differences in factors that influence the willingness to participate in medical research studies. *Ann. Epidemiol.* 4, 248–256. doi: 10.1016/s1047-2797(01)00265-4

Sheppard, V. B., Cox, L. S., Kanamori, M. J., Canar, J., Rodriguez, Y., Goodman, M., et al. (2005). Brief report: if you build it, they will come: methods for recruiting Latinos into cancer research. *J. Gen. Intern. Med.* 5, 444–447. doi: 10.1111/j.1525-1497.2005.0083.x

Sofer, T., Wong, Q., Hartwig, F. P., Taylor, K., Warren, H. R., Evangelou, E., et al. (2017). Genome-Wide association study of blood pressure traits by hispanic/latino background: the hispanic community health study/study of Latinos. *Sci. Rep.* 7:10348. doi: 10.1038/s41598-017-09019-1

SPSS (2013). *IBM SPSS Statistics for Windows, Version 22.0*. Armonk, NY: IBM Corp.

Tang, M. X., Cross, P., Andrews, H., Jacobs, D. M., Small, S., Bell, K., et al. (2001). Incidence of AD in African-Americans, Caribbean Hispanics, and Caucasians in northern Manhattan. *Neurology* 56, 49–56. doi: 10.1212/wnl.56.1.49

Tang, M. X., Stern, Y., Marder, K., Bell, K., Gurland, B., Lantigua, R., et al. (1998). The APOE-epsilon4 allele and the risk of Alzheimer disease among African Americans, whites, and Hispanics. *JAMA* 10, 751–755.

Teng, E. L., and Chui, H. C. (1987). The modified Mini-Mental State (3MS) examination. *J. Clin. Psychiatry* 48, 314–318.

Tucker, K. L., Mattei, J., Noel, S. E., Collado, B. M., Mendez, J., Nelson, J., et al. (2010). The Boston Puerto Rican Health Study, a longitudinal cohort study on health disparities in Puerto Rican adults: challenges and opportunities. *BMC Public Health* 10:107. doi: 10.1186/1471-2458-10-107

Vardarajan, B. N., Schaid, D. J., Reitz, C., Lantigua, R., Medrano, M., Jimenez-Velazquez, I. Z., et al. (2014). Inbreeding among Caribbean Hispanics from the dominican republic and its effects on risk of Alzheimer disease. *Genet. Med.* 8, 639–643. doi: 10.1038/gim.2014.161

Via, M., Gignoux, C. R., Roth, L. A., Fejerman, L., Galanter, J., Choudhry, S., et al. (2011). History shaped the geographic distribution of genomic admixture on the island of Puerto Rico. *PLoS One* 1:e16513. doi: 10.1371/journal.pone.0016513

Weintraub, S., Salmon, D., Mercaldo, N., Ferris, S., Graff-Radford, N. R., Chui, H., et al. (2009). The Alzheimer's Disease Centers' Uniform Data Set (UDS): the neuropsychologic test battery. *Alzheimer Dis. Assoc. Disord.* 2, 91–101. doi: 10.1097/WAD.0b013e318191c7dd

Wilkins, C. H. (2018). *Precision Medicine for Everyone. NEJM: Catalyst*. Available at: https://catalyst.nejm.org/precisionmedicine-initiative-everyone/ (accessed January 28, 2019).

Yesavage, J. A. (1988). Geriatric depression scale. *Psychopharmacol. Bull.* 24, 709–711.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

BRIEF RESEARCH REPORT
published: 16 July 2019
doi: 10.3389/fgene.2019.00658

Motivations for Participation in Parkinson Disease Genetic Research Among Hispanics versus Non-Hispanics

Karen Nuytemans[1,2], Clara P. Manrique[1], Aaron Uhlenberg[1], William K. Scott[1,2], Michael L. Cuccaro[1,2], Corneliu C. Luca[3], Carlos Singer[3] and Jeffery M. Vance[1,2]*

[1] John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, FL, United States, [2] Dr. John T. Macdonald Foundation Department of Human Genetics, University of Miami Miller School of Medicine, Miami, FL, United States, [3] Department of Neurology, University of Miami Miller School of Medicine, Miami, FL, United States

OPEN ACCESS

Edited by:
Jessica Nicole Cooke Bailey, Case Western Reserve University, United States

Reviewed by:
Kenneth M. Weiss, Pennsylvania State University, United States
Mark Z. Kos, University of Texas Rio Grande Valley Edinburg, United States

***Correspondence:**
Karen Nuytemans
knuytemans@med.miami.edu

Specialty section:
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics

Received: *29 November 2018*
Accepted: *21 June 2019*
Published: *16 July 2019*

Citation:
Nuytemans K, Manrique CP, Uhlenberg A, Scott WK, Cuccaro ML, Luca CC, Singer C and Vance JM (2019) Motivations for Participation in Parkinson Disease Genetic Research Among Hispanics versus Non-Hispanics. Front. Genet. 10:658. doi: 10.3389/fgene.2019.00658

Involvement of participants from different racial and ethnic groups in genomic research is vital to reducing health disparities in the precision medicine era. Racial and ethnically diverse populations are underrepresented in current genomic research, creating bias in result interpretation. Limited information is available to support motivations or barriers of these groups to participate in genomic research for late-onset, neurodegenerative disorders. To evaluate willingness for research participation, we compared motivations for participation in genetic studies among 113 Parkinson disease (PD) patients and 49 caregivers visiting the Movement Disorders clinic at the University of Miami. Hispanics and non-Hispanics were equally motivated to participate in genetic research for PD. However, Hispanic patients were less likely to be influenced by the promise of scientific advancements ($N = 0.01$). This lack of scientific interest, but not other motivations, was found to be likely confounded by lower levels of obtained education ($N = 0.001$). Overall, these results suggest that underrepresentation of Hispanics in genetic research may be partly due to reduced invitations to these studies.

Keywords: participation, genetics, research, diversity, Parkinson disease

INTRODUCTION

Disproportionate participation in genomic studies across different racial and ethnic groups has significant long-term implications for translational benefits associated with this research. Research that focuses on a limited pool of racial and ethnic diverse populations versus a broader array of populations may lead to potential biases in genomic research findings and restrict benefit to the limited population group (Bustamante et al., 2011). Often the same groups that are underserved in health care are underrepresented in genomic research, thus increasing the potential for future overall health disparities. With the development of a precision care model for health services, more genomic information will be integrated into health services (Biesecker and Green, 2014). This development emphasizes the importance of racial or ethnic specific genetic information with respect to disease risk. Our understanding of population-specific genomic information relies on the inclusion of all racial and ethnic groups in genomic research. The absence of such information will render the implementation of precision medicine in the understudied groups less effective at

best. Recent summary studies on genomic reports confirm that approximately 80% of individuals included in genome-wide association studies are European or of European descent, with only 1% Hispanic representation (Bustamante et al., 2011; Popejoy and Fullerton, 2016; Sirugo et al., 2019). Even applications of the more recent next-generation sequencing technology still include over 60% European (descent) individuals (Bustamante et al., 2011; Popejoy and Fullerton, 2016; Sirugo et al., 2019). A common belief is that underrepresentation of racial and ethnic groups in biomedical research is the result of reduced willingness to participate because of mistrust and stigma (Shavers et al., 2002; George et al., 2014; Erves et al., 2017). An alternative position is that underrepresentation can also be ascribed to limited access to research opportunities and reduced invitations to participate (Wendler et al., 2006; Ceballos et al., 2014). Previous studies focusing on participation in research in general have found that among non-white populations, willingness to participate is closely linked to and motivated by concern for personal or overall family or community health (Sanderson et al., 2013; Ulrich et al., 2013; Ceballos et al., 2014; George et al., 2014). Unwillingness, in contrast, is often driven by negative perception of research, lack of personal benefit, and/or fear of results (Sanderson et al., 2013; George et al., 2014; Erves et al., 2017).

Little information is known on the perceived barriers and motivations of patients with late-onset neurodegenerative diseases to participate in research. Despite the higher prevalence of Parkinson disease (PD) and Alzheimer disease (AD) in Hispanics versus white non-Hispanics, reports discussing race or ethnicity in AD or PD health care provide evidence of disparities in prescribing medications (Hemming et al., 2011; Thorpe et al., 2016), referrals to clinical trials (Schneider et al., 2009), availability of resources (Graham-Phillips et al., 2016), and cost (Gilligan et al., 2013) (benefitting non-Hispanic whites more than any other group). Specifically for PD, disparities relating to referrals to deep brain stimulation surgery have also already observed (Chan et al., 2014). This disparity for these disorders extends to genomic research as there is little data on genetic variation (variants, frequency and/or effect size) contributing to PD (or AD) in non-whites. Interestingly, variants unique to a specific racial or ethnic background are reported for PD (e.g., *PINK1* in Asians; Nuytemans et al., 2010) as well as AD (e.g., *ABCA7* frameshift deletion in African Americans; Farrer et al., 1997; Collins, 1999; Calderon et al., 2006; Reitz et al., 2013; Cukier et al., 2016; Feliciano et al., 2016), indicating a clear need to increase research in non-white populations.

Hispanics can harbor variable levels of admixture of European, African, and Native American ancestry in their genetic background (Mao et al., 2007; Price et al., 2007; Bryc et al., 2010). Detailed analyses in the Hispanic population and other admixed populations can thus inform on genetic contributions of disease in the others. Therefore, these groups can be highly instructive in our understanding of genetic disease across race and ethnicity. For example, through analyses of local ancestry, Dr. Rajabli et al. found different risk effects associated with APOEε4 in Hispanic AD depending on the ancestral origin of the region the ε4 was located on (European OR = 10 versus African OR = 3; Rajabli et al., 2018), consistent with previously observed lower APOEε4 risk for AD in an admixed population of African Americans (Farrer et al., 1997). Additionally, after identifying a strong risk effect in African Americans for *ABCA7* (Reitz et al., 2013) (similar to *APOE* in WNH), we recently identified a pathogenic 44-bp deletion in *ABCA7* specific to the African American population and Caribbean Hispanics with an African ancestral background in the *ABCA7* region (Cukier et al., 2016).

To date, despite theirs being the largest minority group in the US (U.S. Census Bureau American Community Survey, 2017), only a handful of genomic studies studying the major PD genes (*LRRK2*, *PARK2*, *PARK7*, *PINK1*, and *SNCA*) have focused on PD patients of Hispanic ancestry (Deng et al., 2006; Alcalay et al., 2010; Marder et al., 2010; Saunders-Pullman et al., 2011; Gatto et al., 2013; Duque et al., 2015; Cornejo-Olivas et al., 2017). These studies often present data in a small sample size of Hispanic patients and summarize across all Hispanic PD patients, regardless of ancestry. Given the high variability of admixture in these populations, caution is warranted for the interpretation and extrapolation of these results. The only study of a large cohort of Hispanic patients ($N = 1{,}150$) originating from southern South America reports highly variable contribution of LRRK2 p.G2019S (originally observed in European patients) to PD in different Latin American countries (Mata et al., 2011). Additionally, Mata et al. observed an enrichment of an LRRK2 variant p.Q1111H in Peruvian and Chilean, but not Uruguayan or Argentinian PD patients (Mata et al., 2011), suggesting that this variant originated from the Native American genetic background in these patients. Follow-up analyses showed that this variant is common on Native American background and not contributing to disease (Cornejo-Olivas et al., 2017). Alternatively, when screening GBA, a population-specific variant (p.K198E) contributing to disease was only found in the Colombian population (Velez-Pardo et al., 2019). Taken together, the data presented above underscore the need to include admixed and non-European populations in biomedical research of PD and other neurodegenerative disorders to further our understanding of genetic contribution to PD in these populations with complex genetic architectures as well as across all populations (i.e., transethnic).

Here, we wished to evaluate the willingness of patients affected by a late-onset, complex disease (PD) and their caregivers to participate in genomic research, and the main drivers of this willingness across race and ethnicity to potentially identify issues to address and adjust current enrollment protocols to improve participation across all populations.

MATERIAL AND METHODS

Human Subject Research Compliance

The presented study was approved by the Institutional Review Board at the University of Miami, and informed consent for the survey was obtained from all participants.

Participants and Enrollment

All patients were seen by physicians specializing in movement disorders (CS, CCL) at the University of Miami (UM) Health System's Division of Parkinson's Disease and Movement Disorders clinic. This division serves as the premier referral center for movement disorders patients from abroad with a particular connection to Latin America and the Caribbean. Both Dr. Singer and Dr. Luca speak Spanish and can address the patient in their preferred language. Summary data from the UM Health System suggest that approximately 35% of PD patients identify as Hispanic. Individuals were eligible for this study if they a) had a clinical diagnosis of PD or b) were caregivers of a person with PD and c) were 18 years of age or older. All eligible individuals were referred to the study by their physicians. Patients who agreed to contribute to this survey were approached about a proposed, hypothetical PD genetic research study closely resembling the one ongoing at the John P. Hussman Institute for Human Genomics at UM. All interviewees received the same information including description of the study purpose and requirements (e.g., a single blood draw, collection of personal and family medical history, and no return of personal results). They were then asked whether they were willing to participate in such a study and to complete a brief survey to indicate the reasons for their decision (i.e., participate vs. not participate). All interactions with participants were conducted in the preferred language of the participant.

Survey Items

Adapting from a prior in-house study (Cuccaro et al., 2014), we constructed a multi-item survey to assess influences on willingness to participate in genomic/genetic research. This survey asked participants to select reasons that influenced their decision to (refuse to) participate in the proposed genetic study of PD. Individuals who agreed to participate were asked to select from six predefined reasons that could have influenced their decision (e.g., "I want to help find a cure for PD" or "I want to help improve science and knowledge on PD") (**Table 1**, **Supplementary Table**). Individuals who declined participation were asked to select from 10 reasons for this decision (e.g., "I don't like having my blood drawn," "I am concerned my insurance company will find out my results," or "I don't trust what will happen with my sample") (**Supplementary Table**). In addition, we collected socio-demographic information including age, sex, race/ethnicity, and education level. To assess race/ethnicity, we asked participants to indicate what race they identify with, as well as to describe themselves as Latino (indicating geographical ancestral origin in Latin America), Hispanic (referring to Spanish-speaking populations in Latin America with ancestral origin in the Iberian Peninsula), neither, or unknown, and indicate country/region of ancestral origin if known (questions available in **Supplementary Data**).

Data Analysis

As described below, we restricted our analyses to individuals who agreed to participate in the proposed, hypothetical genetic study to PD. We tested whether the frequency of endorsement for each of the six reasons for participation differed based on ethnicity using Fisher's exact tests. Given that only 12 individuals indicated that they would not participate in the proposed genetic study, we did not include these data in the statistical analyses.

RESULTS

Participant Description

Over the course of 27 clinic days (1 day a week from November to July), we interviewed 162 individuals, of which 113 were PD patients. The remaining 49 individuals presented themselves as caregivers for the patient (35 spouses/partners, 11 children, 3 other). The majority of patients and caregivers identified as white, Hispanic (WH; ~63% and ~59%, respectively) or white, non-Hispanic (WNH; ~30% and ~37%). Most of the Hispanic participants were of Cuban ancestry (~60%), followed by Colombian (~10%) and Puerto Rican (~9%) ancestry. These figures correspond to the demographic figures for the larger Miami area. Among remaining participants, 3% of individuals

TABLE 1 | Comparison of endorsement rates per reason in Hispanics versus non-Hispanics in patient and caregiver groups.

	Patients			Caregiver		
	WH	**WNH**	***p*-value***	**WH**	**WNH**	***p*-value***
N	72	34		29	18	
Mean age	66.7	67.1		59.3	59.9	
Higher education**, *N* (%)	44 (61.1)	31 (91.2)	0.01	17 (58.6)	16 (88.9)	0.05
Willing to participate, *N* (%)	70 (97.2)	33 (97.1)	1	25 (86.2)	15 (83.3)	1
Motivations selected by those willing to participate, *N* (%)***						
I suffer from PD/have a relative who suffers from PD	59 (84.3)	27 (81.8)	0.78	23 (92.0)	12 (80.0)	0.34
To help future generations with PD	34 (48.6)	20 (62.5)	0.29	11 (44.0)	13 (86.7)	0.01
To help find a cure for PD	66 (94.2)	30 (93.8)	0.68	25 (100.0)	12 (80.0)	0.05
To find new/better treatments for PD	43 (61.4)	28 (87.5)	0.02	17 (68.0)	12 (80.0)	0.49
To improve science and knowledge about PD	32 (45.7)	24 (75.0)	0.01	15 (60.0)	11 (73.3)	0.50
I'm encouraged by family/friend to participate	7 (10.0)	10 (30.3)	0.04	3 (12.0)	4 (26.7)	0.39

**Fisher's exact test; **Higher education defined as education received after high school; ***Other (encompassing any other reasons participants indicated as motivation, not in the predetermined list) was not analyzed due to low number of endorsements.*

identified as Black/African American, 1.2% identified as Arab, and less than 1% identified as Asian. Given these small numbers, we restricted our analyses to WH ($N = 101$) and WNH ($N = 52$) participants.

Mean age at interview as a function of ethnicity did not differ within the patient (WH 66.7y/WNH 67.1y) or caregiver (WH 59.3y/WNH 59.9y) groups. However, a significant difference in reported education level was observed, with >88% of WNH holding a degree of higher education than high school versus ~60% in the WH participant group (p~0.0001 across combined patient/caregiver group, **Table 1**).

Survey Results

Overall, ~91% of WH and WNH participants reported they would participate in the proposed PD genomic study. Specifically, 97% of the 106 patients would agree to participate, with no observed difference between the two ethnic groups ($p = 1.00$). Among patients, "*To help find a cure for PD*" and "*I suffer from PD*" were endorsed at similarly high frequencies between ethnic groups (~91%, $p = 0.68$ and ~83%, $p = 0.78$, respectively), making them the most frequently endorsed reasons for participating in the proposed study (**Table 1**). In contrast, nominally significant higher frequencies of WNH versus WH patients endorsed reasons driven by scientific discovery; "*To find new/better treatments for PD*" (87.5% vs 61.4%, respectively; $p = 0.02$), and "*To improve science and knowledge about PD*" (75% vs 46%, respectively; $p = 0.01$). Additionally, we observed a nominally significant difference for "*I'm encouraged by family/friend to participate*" (WNH 30% vs WH 10%; $p = 0.04$).

In the caregivers group, no difference in willingness to participate in genetic studies was observed, though overall percentage was lower than in patients (86% in WH versus 83% in WNH). Interestingly, while a similar pattern of results was observed among WNH and WH caregivers for the same statements as for the patients, one exception was noted as 86.7% of WNH caregivers endorsed "*To help future generations with PD*" as a reason for participating in the proposed study versus only 44% of WH caregivers ($p = 0.01$).

Given the observed differences in level of education between the two groups, we also analyzed the data based on education status regardless of ethnicity to assess confounding effects (**Table 2**). We observed a nominally significant difference for motivation by "*To improve science and knowledge about PD*" for higher educated versus non-higher educated participants in the patient group ($p = 0.001$; 62.5% vs 26.7%) as well as overall ($p = 0.002$; 64.6% vs 34.9%). No difference in motivation was observed for the caregiver group based on education level.

DISCUSSION

Given the growing impact of genomic information on clinical care for increasing numbers of conditions, it is of utmost importance to recognize the genetic differences among racial and ethnic groups. Available research findings for PD or other neurodegenerative disorders on mostly WNH or Asian population groups are not necessarily generalizable to all individuals (Bustamante et al., 2011). Very recently, genetic research for the more common neurodegenerative disease AD in diverse populations of African Americans and Hispanics has shown the power of these analyses across race and ethnicity to identify variants contributing to disease and improve the field's understanding of disease mechanisms [e.g., GBA (p.K198E) in Hispanics (Velez-Pardo et al., 2019), ABCA7 in Africans and African Americans (Reitz et al., 2013; Cukier et al., 2016), differential risk of APOEe4 on different background (Rajabli et al., 2018)]. These data support the importance to extend genomic research to diverse populations for neurodegenerative disease to fully understand genetic risk factors contributing to disease.

More recently, the number of studies evaluating recruitment issues and methods in different racial and ethnic groups versus the traditional European research population has grown with the rise of precision medicine initiatives, though there are very few for complex, late-onset diseases (Zhou et al., 2016; Hughes et al., 2017). Our results indicate that WH individuals affected by or caring for someone with PD seen at the UM Movement Disorders clinic would

TABLE 2 | Comparison of endorsement rates per reason in higher educated versus non-higher educated participants.

	Patients			Caregiver		Overall	
	Yes	No	*p*-value*	Yes	No	*p*-value*	*p*-value*
Higher education**	74	32		33	14		
Willing to participate, *N* (%)	72 (97.3)	30 (93.7)	0.58	27 (81.8)	13 (92.9)	0.66	1
Motivations selected by those willing to participate, *N* (%)***							
I suffer from PD/have a relative who suffers from PD	59 (81.9)	25 (83.3)	1	23 (85.2)	12 (92.3)	1	0.81
To help future generations with PD	39 (54.2)	14 (46.7)	0.52	17 (63.0)	7 (53.8)	0.73	0.46
To help find a cure for PD	69 (95.8)	28 (93.3)	1	25 (92.6)	12 (92.3)	1	1
To find new/better treatments for PD	51 (70.8)	19 (63.3)	0.49	22 (81.5)	7 (53.8)	0.13	0.16
To improve science and knowledge about PD	45 (62.5)	8 (26.7)	0.001	19 (70.4)	7 (53.8)	0.48	0.002
I'm encouraged by family/friend to participate	11 (15.2)	5 (16.7)	1	7 (25.9)	0 (0.0)	0.07	0.46

**Fisher's exact test; **Higher education defined as education received after high school; ***Other (encompassing any other reasons participants indicated as motivation, not in the predetermined list) was not analyzed due to low number of endorsements.*

be equally willing to participate in genomic research for late-onset disease PD as WNH individuals, given current enrollment protocols. One could argue that the high rate of willingness reflects an increase in interest in research in those individuals seeking treatment at an academic medical center. However, we have observed a significant difference between WNH and WH participants driven by research progress to participate, indicating interest in research has at the very least less priority than other, more personal reasons for WH participants. Additional analyses showed that the lack of motivation of scientific improvement is likely correlated with lower educational level. This divergence could potentially be explained by an underlying lower level of knowledge of or familiarity with basic science and medical research in the WH participant group. Interestingly, the few participants who provided an open answer as reason to participate ("*other*") indicated they are more willing to participate to help their doctor with whom they have a good relationship. Though these were limited numbers, these data might suggest a higher level of trust between physicians/researchers and participants through a more personal relationship and being helped in the participant's language of choice.

Taking together the high willingness seen here but current underrepresentation in medical research of WH participants, we offer that the underrepresentation of WH individuals in PD research is in part due to a reduced invitation to participate. It is therefore important for the medical and scientific fields to make a concerted effort to reach out to the different communities and truly establish a relationship as well as inform on and extend participation in (PD) studies to all races and ethnicities. This investment in community outreach will lead to a more equal representation in research and ultimately to a reduction in health disparities.

ETHICS STATEMENT

The presented study was approved by the Institutional Review Board at the University of Miami and informed consent for the survey was obtained from all participants.

AUTHOR CONTRIBUTIONS

KN, MC, WS, CS, and CL contributed conception and design of the study; KN, CM, and AU managed and organized the project; KN performed the statistical analysis; KN and MC wrote the first draft of the manuscript; WS, JV, CS, and CL critically reviewed the manuscript. All authors contributed to manuscript revision and read and approved the submitted version.

FUNDING

This research was funded by a National Parkinson Foundation Moving Day® grant (PI Nuytemans).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00658/full#supplementary-material

REFERENCES

Alcalay, R. N., Caccappolo, E., Mejia-Santana, H., Tang, M. X., Rosado, L., Ross, B. M., et al. (2010). Frequency of known mutations in early-onset Parkinson disease: implication for genetic counseling: the consortium on risk for early onset Parkinson disease study. *Arch. Neurol.* 67, 1116–1122. doi: 10.1001/archneurol.2010.194

Biesecker, L. G., and Green, R. C. (2014). Diagnostic clinical genome and exome sequencing. *N. Engl. J. Med.* 371, 1170. doi: 10.1056/NEJMra1312543

Bryc, K., Auton, A., Nelson, M. R., Oksenberg, J. R., Hauser, S. L., Williams, S., et al. (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc. Natl. Acad. Sci. U.S.A.* 107, 786–791. doi: 10.1073/pnas.0909559107

Bustamante, C. D., Burchard, E. G., and De la Vega, F. M. (2011). Genomics for the world. *Nature* 475, 163–165. doi: 10.1038/475163a

Calderon, J. L., Baker, R. S., Fabrega, H., Conde, J. G., Hays, R. D., Fleming, E., et al. (2006). An ethno-medical perspective on research participation: a qualitative pilot study. *MedGenMed* 8, 23.

Ceballos, R. M., Knerr, S., Scott, M. A., Hohl, S. D., Malen, R. C., Vilchis, H., et al. (2014). Latino beliefs about biomedical research participation: a qualitative study on the U.S.-Mexico border. *J. Empir. Res. Hum. Res. Ethics* 9, 10–21. doi: 10.1177/1556264614544454

Chan, A. K., McGovern, R. A., Brown, L. T., Sheehy, J. P., Zacharia, B. E., Mikell, C. B., et al. (2014). 2nd Disparities in access to deep brain stimulation surgery for Parkinson disease: interaction between African American race and Medicaid use. *JAMA Neurol.* 71, 291–299. doi: 10.1001/jamaneurol.2013.5798

Collins, F. S. (1999). Shattuck lecture—medical and societal consequences of the human genome project. *N. Engl. J. Med.* 341, 28–37. doi: 10.1056/NEJM199907013410106

Cornejo-Olivas, M., Torres, L., Velit-Salazar, M. R., Inca-Martinez, M., Mazzetti, P., Cosentino, C., et al. (2017). Variable frequency of LRRK2 variants in the Latin American research consortium on the genetics of Parkinson's disease (LARGE-PD), a case of ancestry. *NPJ Parkinsons Dis.* 3, 19–017-0020-6. doi: 10.1038/s41531-017-0020-6

Cuccaro, M. L., Manrique, C. P., Quintero, M., and McCauley, J. L., (2014). Motivations for participation in genomic research in Hispanics vs non-Hispanics. Genetics Awareness Project: "Why we can't wait: conference to eliminate health disparities in genomic medicine" (oral presentation).

Cukier, H. N., Kunkle, B. W., Vardarajan, B. N., Rolati, S., Hamilton-Nelson, K. L., Kohli, M. A., et al. (2016). Alzheimer's Disease Genetics Consortium ABCA7 frameshift deletion associated with Alzheimer disease in African Americans. *Neurol. Genet.* 2, e79. doi: 10.1212/NXG.0000000000000079

Deng, H., Le, W., Guo, Y., Hunter, C. B., Xie, W., Huang, M., et al. (2006). Genetic analysis of LRRK2 mutations in patients with Parkinson disease. *J. Neurol. Sci.* 251, 102–106. doi: 10.1016/j.jns.2006.09.017

Duque, A. F., Lopez, J. C., Benitez, B., Hernandez, H., Yunis, J. J., Fernandez, W., et al. (2015). Analysis of the LRRK2 p.G2019S mutation in Colombian Parkinson's disease patients. *Colomb. Med. (Cali.)* 46, 117–121.

Erves, J. C., Mayo-Gamble, T. L., Malin-Fair, A., Boyer, A., Joosten, Y., Vaughn, Y. C., et al. (2017). Needs, priorities, and recommendations for engaging underrepresented populations in clinical research: a community perspective. *J. Community Health* 42, 472–480. doi: 10.1007/s10900-016-0279-2

Farrer, L. A., Cupples, L. A., Haines, J. L., Hyman, B., Kukull, W. A., Mayeux, R., et al. (1997). Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *JAMA* 278, 1349–1356. doi: 10.1001/jama.1997.03550160069041

Feliciano, B., Hamilton-Nelson, K. L., Adams, L. D., Whitehead, P. L., Hofmann, N. K., Cukier, H. N., et al. (2016). Apolipoprotein E and Alzheimer disease risk in a Puerto Rican Alzheimer disease data set. *The 66th Annual Meeting of the American Society of Human Genetics (ASHG)*, Oct. 18–22, 2016, Vancouver, BC, Canada (Poster presentation).

Gatto, E. M., Parisi, V., Converso, D. P., Poderoso, J. J., Carreras, M. C., Marti-Masso, J. F., et al. (2013). The LRRK2 G2019S mutation in a series of Argentinean patients with Parkinson's disease: clinical and demographic characteristics. *Neurosci. Lett.* 537, 1–5. doi: 10.1016/j.neulet.2013.01.011

George, S., Duran, N., and Norris, K. (2014). A systematic review of barriers and facilitators to minority research participation among African Americans, Latinos, Asian Americans, and Pacific Islanders. *Am. J. Public Health* 104, e16–31. doi: 10.2105/AJPH.2013.301706

Gilligan, A. M., Malone, D. C., Warholak, T. L., and Armstrong, E. P. (2013). Health disparities in cost of care in patients with Alzheimer's disease: an analysis across 4 state Medicaid populations. *Am. J. Alzheimers Dis. Other Demen.* 28, 84–92. doi: 10.1177/1533317512467679

Graham-Phillips, A., Roth, D. L., Huang, J., Dilworth-Anderson, P., and Gitlin, L. N. (2016). Racial and ethnic differences in the delivery of the resources for enhancing Alzheimer's Caregiver Health II Intervention. *J. Am. Geriatr. Soc.* 64, 1662–1667. doi: 10.1111/jgs.14204

Hemming, J. P., Gruber-Baldini, A. L., Anderson, K. E., Fishman, P. S., Reich, S. G., Weiner, W. J., et al. (2011). Racial and socioeconomic disparities in parkinsonism. *Arch. Neurol.* 68, 498–503. doi: 10.1001/archneurol.2010.326

Hughes, T. B., Varma, V. R., Pettigrew, C., and Albert, M. S. (2017). African Americans and clinical research: evidence concerning barriers and facilitators to participation and recruitment recommendations. *Gerontologist* 57, 348–358. doi: 10.1093/geront/gnv118

Mao, X., Bigham, A. W., Mei, R., Gutierrez, G., Weiss, K. M., Brutsaert, T. D., et al. (2007). A genomewide admixture mapping panel for Hispanic/Latino populations. *Am. J. Hum. Genet.* 80, 1171–1178. doi: 10.1086/518564

Marder, K. S., Tang, M. X., Mejia-Santana, H., Rosado, L., Louis, E. D., Comella, C. L., et al. (2010). Predictors of parkin mutations in early-onset Parkinson disease: the consortium on risk for early-onset Parkinson disease study. *Arch. Neurol.* 67, 731–738. doi: 10.1001/archneurol.2010.95

Mata, I. F., Wilhoite, G. J., Yearout, D., Bacon, J. A., Cornejo-Olivas, M., Mazzetti, P., et al. (2011). Lrrk2 p.Q1111H substitution and Parkinson's disease in Latin America. *Parkinsonism Relat. Disord.* 17, 629–631. doi: 10.1016/j.parkreldis.2011.05.003

Nuytemans, K., Theuns, J., Cruts, M., and Van Broeckhoven, C. (2010). Genetic etiology of Parkinson disease associated with mutations in the SNCA, PARK2, PINK1, PARK7, and LRRK2 genes: a mutation update. *Hum. Mutat.* 31, 763–780. doi: 10.1002/humu.21277

Popejoy, A. B., and Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature* 538, 161–164. doi: 10.1038/538161a

Price, A. L., Patterson, N., Yu, F., Cox, D. R., Waliszewska, A., McDonald, G. J., et al. (2007). A genomewide admixture map for Latino populations. *Am. J. Hum. Genet.* 80, 1024–1036. doi: 10.1086/518313

Rajabli, F., Feliciano, B. E., Celis, K., Hamilton-Nelson, K. L., Whitehead, P. L., Adams, L. D., et al. (2018). Ancestral origin of ApoE epsilon4 Alzheimer disease risk in Puerto Rican and African American populations. *PLoS Genet.* 14, e1007791. doi: 10.1371/journal.pgen.1007791

Reitz, C., Jun, G., Naj, A., Rajbhandary, R., Vardarajan, B. N., Wang, L. S., et al. (2013). Alzheimer Disease Genetics Consortium Variants in the ATP-binding cassette transporter (ABCA7), apolipoprotein E 4, and the risk of late-onset Alzheimer disease in African Americans. *JAMA* 309, 1483–1492. doi: 10.1001/jama.2013.2973

Sanderson, S. C., Diefenbach, M. A., Zinberg, R., Horowitz, C. R., Smirnoff, M., Zweig, M., et al. (2013). Willingness to participate in genomics research and desire for personal results among underrepresented minority patients: a structured interview study. *J. Community Genet.* 4, 469–482. doi: 10.1007/s12687-013-0154-0

Saunders-Pullman, R., Cabassa, J., San Luciano, M., Stanley, K., Raymond, D., Ozelius, L. J., et al. (2011). LRRK2 G2019S mutations may be increased in Puerto Ricans. *Mov. Disord.* 26, 1772–1773. doi: 10.1002/mds.23632

Schneider, M. G., Swearingen, C. J., Shulman, L. M., Ye, J., Baumgarten, M., and Tilley, B. C. (2009). Minority enrollment in Parkinson's disease clinical trials. *Parkinsonism Relat. Disord.* 15, 258–262. doi: 10.1016/j.parkreldis.2008.06.005

Shavers, V. L., Lynch, C. F., and Burmeister, L. F. (2002). Racial differences in factors that influence the willingness to participate in medical research studies. *Ann. Epidemiol.* 12, 248–256. doi: 10.1016/S1047-2797(01)00265-4

Sirugo, G., Williams, S. M., and Tishkoff, S. A. (2019). The missing diversity in human genetic studies. *Cell* 177, 1080. doi: 10.1016/j.cell.2019.04.032

Thorpe, C. T., Fowler, N. R., Harrigan, K., Zhao, X., Kang, Y., Hanlon, J. T., et al. (2016). Racial and ethnic differences in initiation and discontinuation of antidementia drugs by Medicare beneficiaries. *J. Am. Geriatr. Soc.* 64, 1806–1814. doi: 10.1111/jgs.14403

U.S. Census Bureau American Community Survey 2017. https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=CF.

Ulrich, A., Thompson, B., Livaudais, J. C., Espinoza, N., Cordova, A., and Coronado, G. D. (2013). Issues in biomedical research: what do Hispanics think? *Am. J. Health Behav.* 37, 80–85. doi: 10.5993/AJHB.37.1.9

Velez-Pardo, C., Lorenzo-Betancor, O., Jimenez-Del-Rio, M., Moreno, S., Lopera, F., Cornejo-Olivas, M., et al. (2019). The distribution and risk effect of GBA variants in a large cohort of PD patients from Colombia and Peru. *Parkinsonism Relat. Disord.* doi: 10.1016/j.parkreldis.2019.01.030 (Forthcoming)

Wendler, D., Kington, R., Madans, J., Van Wye, G., Christ-Schmidt, H., Pratt, L. A., et al. (2006). Are racial and ethnic minorities less willing to participate in health research? *PLoS Med.* 3, e19. doi: 10.1371/journal.pmed.0030019

Zhou, Y., Elashoff, D., Kremen, S., Teng, E., Karlawish, J., and Grill, J. D. (2016). African Americans are less likely to enroll in preclinical Alzheimer's disease clinical trials. *Alzheimers Dement. (N. Y.)* 3, 57–64. doi: 10.1016/j.trci.2016.09.004

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor declared a past co-authorship with one of the authors WS.

frontiers in Genetics

ORIGINAL RESEARCH
published: 13 March 2020
doi: 10.3389/fgene.2020.00120

Understanding Participation in Genetic Research Among Patients With Multiple Sclerosis: The Influences of Ethnicity, Gender, Education, and Age

Michael L. Cuccaro[1,2], *Clara P. Manrique*[1], *Maria A. Quintero*[1], *Ricardo Martinez*[1] *and Jacob L. McCauley*[1,2*]

[1] John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, FL, United States, [2] Dr. John T. Macdonald Foundation, Department of Human Genetics and Genomics, University of Miami Miller School of Medicine, Miami, FL, United States

OPEN ACCESS

Edited by:
Jessica Nicole Cooke Bailey, Case Western Reserve University, United States

Reviewed by:
Satyanarayana M. R. Rao, Jawaharlal Nehru Centre for Advanced Scientific Research, India
Marsha Michie, Case Western Reserve University, United States

***Correspondence:**
Jacob L. McCauley
jmccauley@med.miami.edu

Specialty section:
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics

Received: 12 October 2018
Accepted: 31 January 2020
Published: 13 March 2020

Citation:
Cuccaro ML, Manrique CP, Quintero MA, Martinez R and McCauley JL (2020) Understanding Participation in Genetic Research Among Patients With Multiple Sclerosis: The Influences of Ethnicity, Gender, Education, and Age. Front. Genet. 11:120. doi: 10.3389/fgene.2020.00120

This study examined reasons for participation in a genetic study of risk for multiple sclerosis (MS). Our sample consisted of 101 patients diagnosed with MS who were approached about enrolling in the Multiple Sclerosis Genetic Susceptibility Study. Participants were predominantly Hispanic (80%), female (80%), and well educated (71%), having at least some level of college education. Of these 101 individuals who were approached, 95 agreed to participate and are the focus of this report. Among enrollees, the most frequently cited reasons for participation were to find a cure for MS (56%), having MS (46%), and helping future generations (37%). Regression models comparing ethnic groups, Hispanics endorsed having MS as a reason to participate significantly more frequently than non-Hispanics (HI 52%, non-HI 19%, $p = 0.015$) while non-Hispanics endorsed finding new and better treatments significantly more frequently than Hispanics (Hispanic 17%, non-Hispanic 50%, $p = 0.003$). Among our three age groups, younger individuals endorsed finding a cure for MS significantly more frequently (74% of 18–35-year olds vs. 56% of 36–55 year olds vs. 39% of >55 year olds). Our results suggest that motivations for participation in genetic research vary by ethnicity, and that these influences need to be considered in developing more inclusive programs of disease-related genetic research. Future efforts should focus on development of standard methods for understanding participation in genetic and genomic research, especially among underrepresented groups as a catalyst for engaging all populations.

Keywords: participation, genetics, research, minorities, motivation, multiple sclerosis

INTRODUCTION

It is widely believed that underrepresented groups are less willing to participate in biomedical research due to barriers such as mistrust, stigma, and competing demands, leading to under-representation (Shavers et al., 2002; George et al., 2014). However, under-representation in biomedical research is also a by-product of limited access to research opportunities and reduced invitations to participate (Wendler et al., 2006; Katz et al., 2007), which persists to this day (Jones et al., 2017). Thus, even in situations where willingness to participate in biomedical research among underrepresented populations is indistinguishable from other groups, levels of participation may differ for other reasons (Katz et al., 2009; Fisher and Kalbaugh, 2011). Importantly, it is not clear that underrepresented groups' attitudes about participation in biomedical research extend to participation in genetic research. Reduced willingness to participate in genetic research has generally been attributed to unfavorable attitudes about this type of research (Matsui et al., 2005). Clearly, there is much to be learned about why individuals from underrepresented populations participate in genetic research.

Among underrepresented populations, consistent themes for participation include altruism, benefit to family members, self-benefit, and personal curiosity (Sanderson et al., 2013; Walker et al., 2014). Similarly, concerns about individual and family health as well as helping the common good were primary motivations for participation in genetic research among African Americans enrolled in the Jackson Heart study (Walker et al., 2014). Respondents in this study also reported being motivated by the opportunity to get involved in something that would help African Americans across the country; most expressed a high confidence and trust in the study leaders and staff. Sanderson and colleagues conducted structured interviews to assess willingness to participate in genomics research on complex diseases among a diverse group of participants from an inner-city hospital, which included black, Hispanic, and non-Hispanic white individuals (Sanderson et al., 2013). Results showed that willingness to participate was motivated by altruism, benefit to family members, personal health benefit, personal curiosity and improving understanding. In contrast, unwillingness to participate was motivated by negative perceptions of research, lack of perceived personal relevance, negative feelings about procedures (e.g., blood draws), practical barriers, and fear of results (Sanderson et al., 2013).

The importance of participation in genetic research has implications for translational benefits associated with such research. For various groups that may already be under-served, an under-representation in genetic research can amplify future health disparities. For instance, Bustamante and colleagues report that failure to investigate a "broader ensemble of populations" will bias findings from genomic research and benefit only the privileged segment of the population who participate (Bustamante et al., 2011). While this situation has improved somewhat, there is still an underrepresentation of non-European populations in genetic research, which is crucial to ensuring that the benefits of research are available for all (Popejoy and Fullerton, 2016). The importance of genetics for health services has been anticipated for some time (Sterling et al., 2006). More than 10 years after Sterling and colleagues described the importance of genetics for health services (Sterling et al., 2006), the integration of genetics in health services has arrived as whole exome and whole genome sequencing technologies are increasingly present in clinical settings (Biesecker and Green, 2014; Krier et al., 2016). However, as noted by Landry and colleagues, a lack of equitable representation in this new era of precision medicine research will inhibit translational benefits for groups not represented (Landry et al., 2018).

Efforts to include underrepresented groups in genetic and genomic research have increased, albeit slowly. One line of study has examined influences on willingness to participate, including motivations. To date, findings from studies of motivation to participate in genomic research among underrepresented populations have been mixed, and some of the observed differences in outcomes may be attributable to study design. For example, some studies assess motivations to participate among individuals who enroll or decline participation in a genetic risk study (i.e., actual participation) (Parikh et al., 2017) while others survey intentions to participate (Halbert et al., 2016; Cooke Bailey et al., 2018). Similarly, some studies enroll patients who are from the general population of patients in both hospital and non-hospital setting (Sanderson et al., 2013; Walker et al., 2014; Jones et al., 2017), while others assess factors associated with participation among patients with specific diseases (Parikh et al., 2017). This is an important distinction as motivational factors vary considerably depending on the type of study and population (e.g., clinical trial vs. observational study, disease group vs. healthy population) (Goodman et al., 2018; Goodman et al., 2019). Further, the set of reasons that motivate healthy individuals to participate is likely very different from reasons that motivate individuals with specific diseases. To date, there have been limited studies using methods which directly ask individuals with specific diseases about reasons for participating in genetic research for those diseases. Acknowledging the concerns raised by Goodman and colleagues around conflating disease and healthy population studies and methods, we believe that asking patients who enroll in genetic studies about their reasons for enrollment is the most informative approach. This belief is supported by the work of the Clinical Sequencing Exploratory Research (CSER) consortium, which has investigated multiple facets of participation in genomic research, including why patients decline to participate (Amendola et al., 2018).

For this study, we asked patients with multiple sclerosis (MS) who were participating in a genetic risk study for MS to identify the primary reasons or motivations for participation using questions based on information from prior qualitative studies. We examined the frequencies of responses in relation to ethnicity, age, and gender. To date, incorporating genetics into precision medicine for MS is a work in progress (Giovannoni, 2017; Hansen and Okuda, 2018), but there has been considerable progress over the past several years (Matthews, 2015). As these genetic discoveries slowly accrue and become clinically useful, it is equally important that they are applicable across populations (Hindorff et al., 2018; Bonham et al., 2018). However, as noted above, the utility of

genomic information in clinical settings rests on a foundation of established findings from prior studies and the absence of such information affects interpretation of clinical findings. Thus, a lack of diversity in research has the potential to exacerbate existing inequalities in health care (Popejoy and Fullerton, 2016). Given the under inclusion of non-European ancestry groups in genetic and genomic research, a necessary first step is to understand the factors that influence participation and then use this information to create more inclusive ascertainment.

METHODS

Human Subjects Research Compliance

All procedures followed were in accordance with the ethical standards of the Institutional Review Board at the University of Miami Miller School of Medicine, and with the Helsinki Declaration of 1975, as revised in 1999 (Human, 1999). Informed consent was obtained from all participants included in the study.

Participants and Enrollment

Participants for this study consisted of 101 patients with a diagnosis MS who were ascertained through the University of Miami Health System's MS Center of Excellence, as well as the local community. Patients were eligible for this study if they had a clinical diagnosis of MS and were 18 years of age or older.

Potential enrollees in the genetic risk for MS study were recruited in the clinic setting or at a community outreach events, at which time they were invited to participate. Most of our participants were enrolled in the clinic setting, indicative of the volume of patients available at that site. Once they indicated their decision, the clinical coordinator would ask individuals to select a reason(s) for their decision (i.e., to participate in the genetic research study or not) from a list of possible reasons (which were presented to the participant) and record their answers. Participants also provided socio-demographic information at that time. All materials were presented in the preferred language of the participant.

Measures

Sociodemographic information

Participants were asked their gender, race-ethnicity, and religious affiliation. In addition, they were asked to indicate their age group and education level.

Reasons for participation

We identified 11 possible reasons for participation (two of which were "other" and "not sure") in a genetic research study (see list of reasons in **Supplementary Information**). The reasons were derived from multiple studies of reasons for participating in biomedical research (e.g., clinical trials and observational studies) as well as biobank and genetic studies (Streicher et al., 2011; Lang et al., 2013; Sanderson et al., 2013; Walker et al., 2014) that were primarily conducted among convenience samples of individuals with no known disease or illness. Given the paucity of published methods for evaluating willingness to participate in clinical populations we created questions that reflected the primary themes from other types of qualitative research (e.g., structured interviews and focus groups) that assessed willingness to participate in genetic research for reasons such as altruism (e.g.,*To help future generations*), personal benefit (e.g., *I suffer from MS*), and advancing research (e.g., *To help improve science and knowledge about MS*). The questions were drafted by one of the investigators (clinical psychologist) and subsequently reviewed by other team members including the director of patient and family ascertainment and senior clinical coordinators, both who have extensive experience in participant recruitment. Following revisions, the survey was administered to various staff to evaluate wording, item order, and item complexity.

Data Analysis

Our primary questions of interest involved whether endorsement of reasons for participating in the genetic risk for MS study differed by ethnicity, gender, education, and age. To answer these questions, we conducted separate logistic regression analyses using ethnicity, gender, and education as binary outcomes (i.e., Hispanic vs. non-Hispanic, male vs. female, any college vs. no college), and our survey items as predictor variables. For age, we conducted multinomial logistic regression with three levels of our outcome variable (young = 18-35 years, middle = 36-55 years, and older = > 55 years). We tested each of the models for significance and report on those items which are significant contributors to the respective models (i.e., which items predict the outcomes of interests (e.g., Hispanics vs non-Hispanics), thereby reducing the number of significance tests to those associated with the four overall tests (corrected significance level $p = 0.0125$). Odds ratios and confidence intervals are available for each model. All statistical analyses were performed using SPSS version 24 software (SPSS, 2013) and were restricted to individuals who agreed to participate (n = 95).

RESULTS

Among the 101 individuals approached about participating in the genetic risk for MS study, 95 (94%) agreed to participate. All results are based on this group of 95 individuals. As seen in **Table 1**, most of our participants were Hispanic (N = 79; 83%) and female (N = 78; 82%). We tested whether our Hispanic and non-Hispanic participants differed with respect to gender and found no differences in the proportions of males and females by ethnicity (Fisher's Exact Test, $p = 0.15$). Similarly, while a large percentage of the sample was college educated (71%), we found that our Hispanic and non-Hispanic participants did not differ in education ($p = 0.58$). Finally, there were no differences in age by ethnic group ($p = 0.47$)

Examination of overall endorsement patterns (**Figure 1**) showed that finding a cure, endorsed by 56% of participants,

TABLE 1 | Cohort description (N=95).

Ethnicity	
Hispanic	N=79 (83%)
Non-Hispanic	N=16 (17%)
Sex	
Female	N=78 (82%)
Male	N=17 (18%)
Age	
18–35 years	N=23 (24%)
36–55 years	N=45 (48%)
> 55 years	N=26 (28%)
Education	
College	N=66 (70%)
Non-College	N=29 (30%)
Recruitment Site	
Clinic	N=78 (82%)
Home	N=12 (13%)
Other	N=5 (5%)

was the most frequently cited reason for participating in the study. In addition, having MS and helping future generations, were endorsed by a majority of participants as reasons to enroll in the MS study.

Table 2 summarizes the endorsement patterns for the respective items by ethnicity, gender, education, and age. At the descriptive level, inspection of the frequencies of endorsements shows that both Hispanic and non-Hispanic participants cited finding a cure equally (56% per group). This was the most common reason for the respective groups. However, compared to non-Hispanics, Hispanic participants endorsed having a disease as a reason to participate in the genetic risk for MS study more frequently than non-Hispanics (HI 52%, NH 19%). Conversely, non-Hispanic participants cited finding new/better treatments more frequently than Hispanics (NHI 50%, HI 17%).

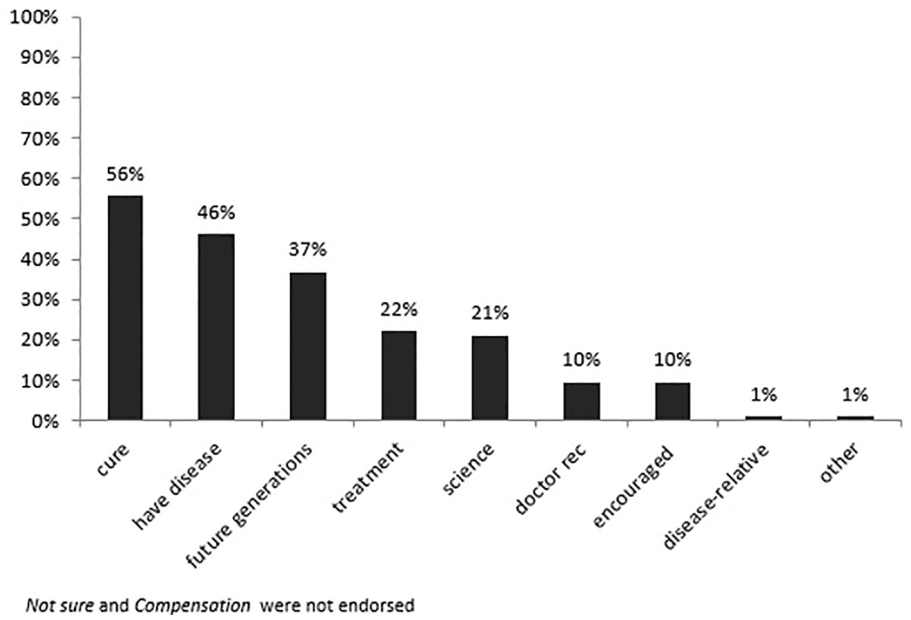

FIGURE 1 | Percentage of endorsements per the respective reasons for participation in the overall sample.

TABLE 2 | Percentage of endorsements for reasons to participate by ethnicity, sex, education, and age (N and % values).

	Ethnicity		Sex		Education		Age (years)		
	HI N=79	NH N=16	M N=17	F N=78	College N=66	~College N=29	young N=23	middle N=45	old N=26
Cure for MS	44 56%	9 56%	6 35%	47 60%	39 59%	14 48%	17 74%	25 56%	10 39%
Suffer from MS	41 52%	3 19%	7 41%	37 47%	32 49%	12 41%	10 44%	22 49%	11 42%
Help future generations	27 34%	8 50%	5 29%	30 39%	25 38%	10 34%	7 30%	17 38%	10 39%
Better treatments for MS	13 17%	8 50%	5 29%	16 21%	17 26%	4 14%	7 30%	10 22%	3 12%
Improve science	15 19%	5 31%	3 18%	17 22%	15 23%	5 17%	6 26%	10 22%	3 12%
Recommended by Doctor	9 11%	0 –	2 12%	7 9%	6 9%	3 10%	1 4%	4 9%	4 15%
Encouraged by others	6 8%*	3 19%*	3 18%	6 8%	3 5%	6 21%	3 13%	3 7%	3 12%

Young=18–35 years.
Middle=36–55 years.
Old= > 55 years.

Endorsement patterns by sex, age, and education were similar to those identified in our ethnic groups as finding a cure and having multiple sclerosis were endorsed consistently as reasons for participating in the MS study.

To test for differences in reasons for participating in genetic research we conducted separate logistic regressions to ascertain the effects of the respective survey items (i.e., reasons for participating) on different binary (ethnicity, sex, and education groups) and multinomial (age groups) outcomes. For each of the respective analyses, we restricted our predictors to the following survey items: *I want to help find a cure for MS*; *To help improve science and knowledge about MS*; *To find new/better treatments for MS*; *I suffer from MS*; *To help future generations*; *The doctor asked/recommended that I participate*; *and, Encouragement from a family member or friend*. The remaining items were not cited as reasons for participating by more than one individual.

Ethnic Group

Our logistic regression model evaluating the ability of survey items to predict ethnic group (Hispanic vs. non-Hispanic) was statistically significant, $\chi^2(6) = 20.61$, $p = 0.002$. Of the six predictor variables (i.e., survey items that were reasons for participating in the study), three contributed significantly to the model: *I suffer from MS, To find new/better treatments for MS*, and *Encouragement from a family member or friend*. These items differed between our Hispanic and non-Hispanic participants. Among the three items, the largest OR (7.34; CI 1.52, 35.68) was found for the item, *I suffer from MS*, indicating that endorsing this item as a reason was more likely among Hispanics vs. non-Hispanics. Conversely, *To find new/better treatments for MS* (OR = 0.15), and *Encouragement from a family member or friend* (OR = 0.13), were associated with a reduced likelihood of endorsement by Hispanics vs. non-Hispanics. **Table 3** has the odds ratios and confidence intervals for these results.

Sex

The logistic regression model evaluating the ability of survey items to predict sex was not significant, $\chi^2(7) = 6.54$, $p = 0.478$, as none of the items differed between males and females. The odds ratios and confidence intervals for the respective items are available in Supplementary material (**Supplementary Table 1**).

Education

Similar to the logistic regression model for sex, the model which evaluated the ability of survey items to predict educational group (college vs. no college) was not significant, $\chi^2(7) = 7.33$, $p = 0.396$). The odds ratios and confidence intervals for the respective items are also available in Supplementary material (**Supplementary Table 2**).

Age

As seen in **Table 2**, we collapsed the various age groups into three categories (18–35 years of age, 36–55 years of age, and >55 years of age). Assessment of how well the model fits using likelihood ratio tests was not significant $\chi^2(14) = 13.23$, $p = 0.508$. For one of the predictors, we observed a trend in comparison of the older and younger groups ($p = 0.021$) although given that the omnibus test was not significant, this finding did not survive correction for multiple tests. However, the odds for selecting this as a reason to participate among younger vs. older participants was 4.896, 95% CI 1.28, 18.79) suggesting that this item is more likely among younger vs. older participants. These results along with the additional parameter estimates are available in supplementary material (**Supplementary Table 3**).

DISCUSSION

Overall, our logistic regression analyses yielded only one significant model which showed that there were different reasons for participating in genetic research between Hispanics and non-Hispanics. Among the reasons for participating, personal experience with MS (i.e., *I suffer from MS*), was strongly associated with Hispanics vs. non-Hispanics with an odds ratio of 7.36. In contrast, non-Hispanics were significantly more likely to endorse helping to discover new treatments (OR = 0.15) as a reason to participate. While personal experience with MS and discovery of new treatments are generally aligned with a theme of deriving personal benefit, the differences may hint at subtle distinctions between Hispanics and non-Hispanics or how the items were interpreted. Certainly, our findings regarding Hispanics being motivated by having a disease (i.e., MS) are in line with prior research showing that Hispanics are more likely to participate in biomedical research if it is relevant to them

TABLE 3 | Summary of logistic regression model for ethnic group (Hispanic vs non-Hispanic) using reasons for participation as predictors (predicted outcome=Hispanic).

	B	S.E.	Wald	*df*	*p*	OR	95% CI for OR	
							Lower	Upper
Cure for MS	.195	.681	.082	1	0.775	1.215	.320	4.619
Improve science	-.709	.844	.704	1	0.401	.492	.094	2.576
Better Treatments for MS*	-1.871	.723	6.708	1	0.010	.154	.037	.634
Suffer from MS*	1.995	.806	6.135	1	0.013	7.356	1.517	35.677
Help future generations	-1.021	.692	2.179	1	0.140	.360	.093	1.397
Encouraged by others*	-2.058	.959	4.607	1	0.032	.128	.019	.836

*OR, odds ratio *significant coefficients.*
MS, multiple sclerosis.

(Ulrich et al., 2013). Note that one additional item, encouragement from others (OR = 0.13), was less likely to be endorsed by Hispanics as a reason to participate in genetic research—again possibly reflecting personal motivation. The second item, finding new and better treatments, was endorsed by 50% of non-Hispanics vs. only 17% of Hispanics, and has elements of personal benefit as well as altruism. Further, while not significant, 50% of non-Hispanics endorsed helping future generations as a reason for motivation compared to 34% of Hispanics. Even though this difference was not significant, when coupled with the results regarding the item finding new and better treatments, there is a suggestion that Hispanics and non-Hispanics with MS may have different perspectives on what they see as priorities for participation.

Importantly, while interpretations of the above response patterns are reasonable and fit with previously published findings regarding personal meaningfulness and benefit to society (Goodman et al., 2019), we would encourage caution in interpretation of the results. In particular, given that we only asked participants to indicate if a particular reason motivated them to participate, endorsements could be interpreted in multiple ways. For instance, endorsement of *I suffer from MS* as a reason to participate could simply be acknowledging that their participation is important for research vs. a desire to derive personal benefit. Ultimately, in the absence of open-ended responses that could explain participant reasoning, multiple inferences about the meaningfulness of the data are possible.

Interestingly, while not significant, 50% of non-Hispanics endorsed helping future generations as a reason for motivation compared to 34% of Hispanics. Even though this difference was not significant, when coupled with the results regarding the item finding new and better treatments, there is a suggestion that Hispanics and non-Hispanics with MS differ in altruism. One additional item, encouragement from others (OR = 0.13), was less likely to be endorsed by Hispanics as a reason to participate in genetic research—again reflecting personal motivation.

At a descriptive level, our results show that among enrollees in an MS genetic risk study, the most frequently cited reason for participating was finding a cure for MS. While this reason for participation did not differ by ethnicity, sex, or education there was a trend among participants in different age groups. Specifically, for the item, *I want to help find a cure for MS*, a positive response was more likely among younger (i.e., 18-35 year olds) vs older (> 55 years) participants; our middle age group (36-55 years) did not differ from younger or older participants for this item. While it is not surprising that endorsement of finding a cure is high among respondents as a whole, especially given that seeking personal benefit is a powerful motivator for participation in biomedical and genetic research, an age-related effect has not been previously reported. Thus, while many studies adjust for age in their analyses to control for its influence on outcomes, this variable may be of value in terms of understanding the likelihood of participation. For instance, participants in the younger age groups may be more enthusiastic about finding a cure as they are still early in the disease process. At a minimum, investigators seeking to enroll participants for genetic studies should be aware of how age may affect motivations to participate in research when developing recruitment strategies.

The current study offers new information about motivations for participation in MS genetic research as a function of ethnicity and age. While the strengths of the study are its focus on individuals who have a disorder (MS) vs a hypothetical scenario, and the inclusion of Hispanics, the results should be interpreted with caution in light of several factors including small sample size, higher education levels, and a high rate of willingness to participate, raising the possibility of bias related to their being approached during a clinical encounter (i.e., at a neurology appointment). Consequently, our results may not be generalizable to individuals with MS who are receiving services outside of academic medical centers or those who are not receiving care. Moving forward, collecting more information such as duration and severity of illness, acculturation, and trust in the health care system, could reveal subtle influences on reasons for participation in genetic research. Finally, as noted in the *Methods* section, we developed the items (i.e., reasons for participation) based on themes from qualitative research conducted with mainly non-disease populations. Given the preliminary nature of our study, the questions have limited formal validation data. However, given the interesting results, we are expanding our efforts to learn more about participant motivations by providing participants an opportunity to explain their choices and recruiting both healthy individuals and those with diseases to compare response patterns. We believe these efforts will increase our ability to understand the nuances of why individuals participate in genetic studies and if those reasons vary by race and ethnicity.

In summary, this study adds to our understanding of influences on actual participation in research studies about genetic risk. Based on our study, it appears that ethnicity was the only significant factor associated with willingness to participate. Studies like this and others provide valuable information about why individuals ultimately participate in genetic research and can inform the development of recruitment strategies. Inclusive enrollment is critical to translational efforts that can play a major role in improving the health and well-being of all individuals.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

This research was approved by the Institutional Review Board, University of Miami Miller School of Medicine. MC, CM, MQ, RM, and JM declare that they have no conflict of interest. All procedures followed were in accordance with the ethical standards

of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2000 (5). Informed consent was obtained from all patients included in the study.

AUTHOR CONTRIBUTIONS

MC, CM, MQ, RM, and JM contributed to the design and implementation of the research, to the analysis of the results, and to the writing of the manuscript.

FUNDING

The research reported in this publication was supported by the National Institutes of Health (NIH) through the National Institute of Neurological Disorders and Stroke (NINDS) under award number 1R01NS096212, the National Institute on Minority Health and Health Disparities (NIMHD) and the National Human Genome Research Institute (NHGRI) under award number U54MD010722, and the National Multiple Sclerosis Society (NMSS) under award number RG4680A1. All content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the NMSS.

ACKNOWLEDGMENTS

We gratefully acknowledge the resources provided by the John P. Hussman Institute for Human Genomics and the strong support of the South Florida chapter of the NMSS. We also thank the multiple sclerosis genetic study participants and their families for their willingness to participate in our research studies.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00120/full#supplementary-material

REFERENCES

Amendola, L. M., Robinson, J. O., Hart, R., Biswas, S., Lee, K., Bernhardt, B. A., et al. (2018). Why patients decline genomic sequencing studies: experiences from the CSER consortium. *J. Genet. Couns.* 5, 1220–1227. doi: 10.1007/s10897-018-0243-7

Biesecker, L. G., and Green, R. C. (2014). Diagnostic clinical genome and exome sequencing. *N. Engl. J. Med.* 12, 1169–1170. doi: 10.1056/NEJMra1312543

Bonham, V. L., Green, E. D., and Perez-Stable, E. J. (2018). Examining how race, ethnicity, and ancestry data are used in biomedical research. *JAMA* 320 (15), 1533–1534. doi: 10.1001/jama.2018.13609

Bustamante, C. D., Burchard, E. G., and De la Vega, F. M. (2011). Genomics for the world. *Nature* 7355, 163–165. doi: 10.1038/475163a

Cooke Bailey, J. N., Crawford, D. C., Goldenberg, A., Slaven, A., Pencak, J., Schachere, M., et al. (2018). Willingness to participate in a national precision medicine cohort: attitudes of chronic kidney disease patients at a cleveland public hospital. *J. Pers. Med.* 8 (3), 21. doi: 10.3390/jpm8030021

Fisher, J. A., and Kalbaugh, C. A. (2011). Challenging assumptions about minority participation in US clinical research. *Am. J. Public Health* 12, 2217–2222. doi: 10.2105/AJPH.2011.300279

George, S., Duran, N., and Norris, K. (2014). A systematic review of barriers and facilitators to minority research participation among African Americans, Latinos, Asian Americans, and Pacific Islanders. *Am. J. Public Health* 2, e16–e31. doi: 10.2105/AJPH.2013.301706

Giovannoni, G. (2017). Personalized medicine in multiple sclerosis. *Neurodegener. Dis. Manage.* 6s, 13–17. doi: 10.2217/nmt-2017-0035

Goodman, D., Bowen, D., Wenzel, L., Tehrani, P., Fernando, F., Khacheryan, A., et al. (2018). The research participant perspective related to the conduct of genomic cohort studies: a systematic review of the quantitative literature. *Transl. Behav. Med.* 1, 119–129. doi: 10.1093/tbm/ibx056

Goodman, D., Johnson, C. O., Bowen, D., Wenzel, L., and Edwards, K. (2019). Factors that motivate participation in observational genetic cancer research studies. *Open J. Epidemiol.* 2, 1–17. doi: 10.4236/ojepi.2019.92014

Halbert, C. H., McDonald, J., Vadaparampil, S., Rice, L., and Jefferson, M. (2016). Conducting precision medicine research with african americans. *PloS One* 11 (7), e0154850. doi: 10.1371/journal.pone.0154850

Hansen, M. R., and Okuda, D. T. (2018). Precision medicine for multiple sclerosis promotes preventative medicine. *Ann. N. Y. Acad. Sci.* 1, 62–71. doi: 10.1111/nyas.13846

Hindorff, L. A., Bonham, V. L., and Ohno-Machado, L. (2018). Enhancing diversity to reduce health information disparities and build an evidence base for genomic medicine. *Per. Med.* 15 (5), 403–412. doi: 10.2217/pme-2018-0037

Human, D. (1999). Declaration of Helsinki. *Lancet* 916, 1888. doi: 10.1016/S0140-6736(05)75101-1

Jones, B. L., Vyhlidal, C. A., Bradley-Ewing, A., Sherman, A., and Goggin, K. (2017). If we would only ask: how henrietta lacks continues to teach us about perceptions of research and genetic research among african americans today. *J. Racial Ethn. Health Disparities* 4, 735–745. doi: 10.1007/s40615-016-0277-1

Katz, R. V., Green, B. L., Kressin, N. R., Claudio, C., Wang, M. Q., and Russell, S. L. (2007). Willingness of minorities to participate in biomedical studies: confirmatory findings from a follow-up study using the tuskegee legacy project questionnaire. *J. Natl. Med. Assoc.* 9, 1052–1060.

Katz, R. V., Green, B. L., Kressin, N. R., James, S. A., Wang, M. Q., Claudio, C., et al. (2009). Exploring the "legacy" of the tuskegee syphilis study: a follow-up study from the tuskegee legacy project. *J. Natl. Med. Assoc.* 2, 179–183. doi: 10.1016/S0027-9684(15)30833-6

Krier, J. B., Kalia, S. S., and Green, R. C. (2016). Genomic sequencing in clinical practice: applications, challenges, and opportunities. *Dialogues Clin. Neurosci.* 3, 299–312.

Landry, L. G., Ali, N., Williams, D. R., Rehm, H. L., and Bonham, V. L. (2018). Lack of diversity in genomic databases is a barrier to translating precision medicine research into practice. *Health Aff. (Millwood)* 5, 780–785. doi: 10.1377/hlthaff.2017.1595

Lang, R., Kelkar, V. A., Byrd, J. R., Edwards, C. L., Pericak-Vance, M., and Byrd, G.S. (2013). African american participation in health-related research studies: Indicators for effective recruitment. *Journal of Public Health Management and Practice : JPHMP*, 19(2), 110–118.

Matsui, K., Kita, Y., and Ueshima, H. (2005). Informed consent, participation in, and withdrawal from
a population based cohort study involving genetic analysis. *J. Med. Ethics* 7, 385–392. doi: 10.1136/jme.2004.009530

Matthews, P. M. (2015). Decade in review-multiple sclerosis: new drugs and personalized medicine for multiple sclerosis. *Nat. Rev. Neurol.* 11, 614–616. doi: 10.1038/nrneurol.2015.200

Parikh, R., O'Keefe, L., Salowe, R., Mccoskey, M., Pan, W., Sankar, P., et al. (2017). Factors associated with participation by African Americans in a study of the

genetics of glaucoma. *Ethn. Health* 24 (6), 1–11. doi: 10.1080/13557858.2017.1346189

Popejoy, A. B., and Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature* 7624, 161–164. doi: 10.1038/538161a

Sanderson, S. C., Diefenbach, M. A., Zinberg, R., Horowitz, C. R., Smirnoff, M., Zweig, M., et al. (2013). Willingness to participate in genomics research and desire for personal results among underrepresented minority patients: a structured interview study. *J. Community Genet.* 4, 469–482. doi: 10.1007/s12687-013-0154-0

Shavers, V. L., Lynch, C. F., and Burmeister, L. F. (2002). Racial differences in factors that influence the willingness to participate in medical research studies. *Ann. Epidemiol.* 4, 248–256. doi: 10.1016/S1047-2797(01)00265-4

SPSS (2013). *IBM SPSS Statistics for Windows, Version 22.0* (Armonk, NY: IBM Corp).

Streicher, S.A., Sanderson, S.C., Jabs, E.W., Diefenbach, M., Smirnoff, M., Peter, I., Horowitz, C.R., Brenner, B., and Richardson, L.D. (2011). Reasons for participatingand genetic information needs among racially and ethnically diverse biobankparticipants: a focus group study. *J. Community Genet*. 2, 153–163.

Sterling, R., Henderson, G. E., and Corbie-Smith, G. (2006). Public willingness to participate in and public opinions about genetic variation research: a review of the literature. *Am. J. Public Health* 11, 1971–1978. doi: 10.2105/AJPH.2005.069286

Ulrich, A., Thompson, B., Livaudais, J. C., Espinoza, N., Cordova, A., and Coronado, G. D. (2013). Issues in biomedical research: what do hispanics think? *Am. J. Health Behav.* 1, 80–85. doi: 10.5993/AJHB.37.1.9

Walker, E. R., Nelson, C. R., Antoine-LaVigne, D., Thigpen, D. T., Puggal, M. A., Sarpong, D. E., et al. (2014). Research participants' opinions on genetic research and reasons for participation: a jackson heart study focus group analysis. *Ethn. Dis.* 3, 290–297.

Wendler, D., Kington, R., Madans, J., Van Wye, G., Christ-Schmidt, H., Pratt, L. A., et al. (2006). Are racial and ethnic minorities less willing to participate in health research? *PloS Med.* 3 (2), e19. doi: 10.1371/journal.pmed.0030019

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer MM and handling Editor declared their shared affiliation.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership