GENETIC ARCHITECTURE AND EVOLUTION OF COMPLEX TRAITS AND DISEASES IN DIVERSE HUMAN POPULATIONS

EDITED BY: Mashaal Sohail, Jeremy Berg and Diego Ortega-Del Vecchyo
PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88974-871-6
DOI 10.3389/978-2-88974-871-6

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: frontiersin.org/about/contact

GENETIC ARCHITECTURE AND EVOLUTION OF COMPLEX TRAITS AND DISEASES IN DIVERSE HUMAN POPULATIONS

Topic Editors:
Mashaal Sohail, National Autonomous University of Mexico, Mexico
Jeremy Berg, The University of Chicago, United States
Diego Ortega-Del Vecchyo, National Autonomous University of Mexico, Mexico

Citation: Sohail, M., Berg, J., Ortega-Del Vecchyo, D., eds. (2022). Genetic Architecture and Evolution of Complex Traits and Diseases in Diverse Human Populations. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-871-6

Table of Contents

frontiers in Genetics

EDITORIAL
published: 10 March 2022
doi: 10.3389/fgene.2022.869056

Editorial: Genetic Architecture and Evolution of Complex Traits and Diseases in Diverse Human Populations

Diego Ortega-Del Vecchyo[1], *Jeremy Berg*[2] *and Mashaal Sohail*[3]*

[1]Laboratorio Internacional de Investigación sobre el Genoma Humano (LIIGH), Universidad Nacional Autónoma de México (UNAM), Juriquilla, México, [2]Department of Human Genetics, University of Chicago, Chicago, IL, United States, [3]Centro de Ciencias Genómicas (CCG), Universidad Nacional Autónoma de México (UNAM), Cuernavaca, México

Keywords: evolution, complex traits, complex disease, diversity, genetic architecture

Editorial on the Research Topic

Genetic Architecture and Evolution of Complex Traits and Diseases in Diverse Human Populations

The research topic "*Genetic architecture and evolution of complex traits and diseases in diverse human populations*" presents six review articles and three research studies. Our Research Topic reflects data generation and research performed in Germany, Brazil, United States, Denmark, Estonia, Mexico, United Kingdom and Singapore, spanning countries across the global income spectrum. We present reviews on methodologies to aid complex trait studies in diverse populations, reviews on complex trait studies in locally and globally understudied groups, reviews on studying the evolution and maintenance of genetic variation influencing complex traits using theory and ancient DNA, and new research studies providing resources or insights to understand the genetic architecture of complex traits in Latin America.

With respect to methodologies that aid complex trait studies in diverse populations, Simonin-Wilmer et al. present an overview of strategies for detecting genotype-phenotype associations across diverse populations. With increased interest and movement towards multi-ethnic association studies that promise to improve detection power and prediction accuracy, this review provides a primer for researchers on assessment and control of confounders related to genetic ancestry to help identify true genotype-phenotype associations. Sticca et al. review the recent methodological developments in detecting identity-by-descent (IBD) segments in the genome. These developments enable IBD detection in large biobank-scale datasets and the authors argue for the need to incorporate IBD-based analyses into genetic studies to improve imputation accuracy, the power to detect rare causal variants, and to gain insights into the demographic history of causal variants.

Two articles in our Research Topic reflect on progress conducting complex trait genetics research in groups that have been historically understudied. In his perspective, Charleston Chiang discusses his work on Native Hawaiians as a motivating example to review the parameters of ethical community engagement, along with the challenges and opportunities of conducting genetic studies in minority populations. He reviews how the complex peopling of the Hawaiian islands over several millennia has shaped patterns of genetic variation there, and the hypothesis that the high rates of metabolic disease observed among Native Hawaiians and related Polynesian populations may be linked to ancient genetic adaptations. In their article, Sun et al. review studies of obesity in samples with ancestry from East Asia. Due to several 100 generations of partially independent

Edited by:
Sander W. van der Laan, University Medical Center Utrecht, Netherlands

Reviewed by:
Wouter Van Rheenen, University Medical Center Utrecht, Netherlands

***Correspondence:**
Mashaal Sohail
mashaal@ccg.unam.mx

Received: *03 February 2022*
Accepted: *22 February 2022*
Published: *10 March 2022*

Citation:
Ortega-Del Vecchyo D, Berg J and Sohail M (2022) Editorial: Genetic Architecture and Evolution of Complex Traits and Diseases in Diverse Human Populations. Front. Genet. 13:869056. doi: 10.3389/fgene.2022.869056

evolution, genetic studies in east Asian samples are well positioned to discover variants that are at low frequencies or are poorly tagged in European samples. Additionally, inconsistencies between GWAS of obesity in European and east Asian samples point to genetic architectures that only partly overlap, suggesting that the environment may be interacting with genetic effects in different ways in different samples.

Koch and Sunyaev review classical and modern population genetic theory on the maintenance, evolution and distribution of genetic variation for complex traits. There is now strong evidence that trait-associated genetic variation is both highly pleiotropic and shaped by natural selection. While population geneticists have been keenly interested in both of these phenomena for decades, the existing theory is not well suited to the rich but complex data that the field now has access to. Koch and Sunyaev highlight more recently developed theory aimed at bridging this gap, as well as the shortcomings of existing statistical tools and challenges for theoreticians going forward. Further, Irving-Pease et al. discuss what particular insights on the evolution of complex traits can be extracted from the analysis of ancient DNA data. The authors analyze how ancient DNA has been used to study the evolution of traits such as height or skin pigmentation, and also assess the potential to predict phenotypes in archaic hominins. The authors review the prospects of using ancient DNA to detect events of polygenic adaptation, how the degraded ancient DNA data can hinder phenotype studies in ancient individuals and point to strategies to improve complex trait studies using ancient DNA.

This Research Topic also contains three papers that provide new insights on complex traits in Latin American populations. First, Secolin et al. analyze a region on chromosome 8 associated with obesity. A study from Brazilian individuals of the BIPMed cohort, sampled in Campinas, showed that this region is under positive selection and has a high proportion of Native American ancestry compared to the rest of the genome. The authors analyze this region in individuals collected in the cities of Barretos, Ribeirao Preto and Belo Horizonte and find the same overrepresentation of Native American ancestry. Second, Kaibara et al. find that 48 SNPs associated with genetic generalized epilepsies (GGE) in non-Brazilian cohorts do not retain an association with GGE in Brazilian patients using genotype data from 87 Brazilian patients with GGE and 340 Brazilian controls. However, they find that nine SNPs in the imputed flanking 1 Mb region surrounding the 48 SNPs retain an association suggesting that there are some shared variants that impact the risk for GGE in Brazilians and individuals from other cohorts around the world. Finally, Jimenez-Kaufmann et al. show that the inclusion of more genomes from Mexican individuals with a high proportion of Native American ancestry in reference imputation panels improves genotype imputation accuracy for rare variants in admixed Mexican individuals. This result, along with other observations from this study, suggest that improvements in genotype imputation accuracy in Latin American individuals from particular regions will require local sequencing efforts to include more individuals with a high proportion of Native American ancestry in imputation reference panels.

The lofty goal of this research topic was to identify and present state-of-the-art research themes in the genetic architecture and evolution of complex traits and diseases in diverse human populations. The major conceptual and practical research directions and challenges that this Research Topic has helped identify that need to be addressed are 1) Theoretical and conceptual advances to understand trait evolution with clear expectations and assumptions regarding shared and variable genetic architecture across human diversity. 2) The clarification and contextualization of the use of ancestry and populations as sampling and analysis variables in research studies. 3) Methodological advances with transparent assumptions to appropriately analyze complex traits using genetic and environmental data across diverse groups. 4) The construction of local scientific capacity globally to allow for fair and inclusive strategies of data collection, analysis and dissemination of results. 5) Large enough sample sizes across human diversity to allow for meaningful inference. 6) Frameworks incorporating equitable reciprocity for communication with study participants especially those that have been historically marginalized and discriminated against across the globe. In this Research Topic, we present considerations to work towards meeting these challenges. By presenting voices and research done across countries and ethnicities, we show the possibilities of a globally well distributed scientific practice, which we believe is an important precedent for widely representative studies of complex traits and disease.

AUTHOR CONTRIBUTIONS

DO, JB, and MS drafted and edited the manuscript. All authors approved the final version.

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

REVIEW
published: 20 October 2020
doi: 10.3389/fgene.2020.575049

Genetics of Obesity in East Asians

Chang Sun, Peter Kovacs and Esther Guiu-Jurado*

Medical Department III – Endocrinology, Nephrology, Rheumatology, University of Leipzig Medical Center, Leipzig, Germany

Obesity has become a public health problem worldwide. Compared with Europe, people in Asia tend to suffer from type 2 diabetes with a lower body mass index (BMI). Genome-wide association studies (GWASs) have identified over 750 loci associated with obesity. Although the majority of GWAS results were conducted in individuals of European ancestry, a recent GWAS in individuals of Asian ancestry has made a significant contribution to the identification of obesity susceptibility loci. Indeed, owing to the multifactorial character of obesity with a strong environmental component, the revealed loci may have distinct contributions in different ancestral genetic backgrounds and in different environments as presented through diet and exercise among other factors. Uncovering novel, yet unrevealed genes in non-European ancestries may further contribute to explaining the missing heritability for BMI. In this review, we aimed to summarize recent advances in obesity genetics in individuals of Asian ancestry. We therefore compared proposed mechanisms underlying susceptibility loci for obesity associated with individuals of European and Asian ancestries and discussed whether known genetic variants might explain ethnic differences in obesity risk. We further acknowledged that GWAS implemented in individuals of Asian ancestries have not only validated the potential role of previously specified obesity susceptibility loci but also exposed novel ones, which have been missed in the initial genetic studies in individuals of European ancestries. Thus, multi-ethnic studies have a great potential not only to contribute to a better understanding of the complex etiology of human obesity but also potentially of ethnic differences in the prevalence of obesity, which may ultimately pave new avenues in more targeted and personalized obesity treatments.

Keywords: genetics, GWAS, obesity, BMI, East Asians

OPEN ACCESS

Edited by:
Mashaal Sohail,
University of Chicago, United States

Reviewed by:
Ayush Giri,
Vanderbilt University Medical Center, United States
Samantha Laber,
University of Oxford, United Kingdom

****Correspondence:***
Peter Kovacs
peter.kovacs@medizin.uni-leipzig.de
Esther Guiu-Jurado
Esther.GuiuJurado@medizin.uni-leipzig.de

Specialty section:
This article was submitted to Human Genomics, a section of the journal Frontiers in Genetics

Received: *22 June 2020*
Accepted: *17 September 2020*
Published: *20 October 2020*

Citation:
Sun C, Kovacs P and Guiu-Jurado E (2020) Genetics of Obesity in East Asians. Front. Genet. 11:575049. doi: 10.3389/fgene.2020.575049

INTRODUCTION

Obesity has become a public health problem throughout the world, whether in developing or developed countries (Ng et al., 2014), and is well recognized as a risk factor for a wide variety of health problems such as diabetes, dyslipidemia, hypertension, and cardiovascular diseases (Van Gaal et al., 2006). Along with the generally acknowledged role of environmental factors such as sedentary lifestyle combined with the intake of energy-dense nutrition and insufficient energy expenditure, development of obesity likely also has a genetic component as demonstrated by both monogenic and common polygenic forms of obesity. Early data from white male twins (Stunkard, 1986), Quebec inhabitants (Bouchard et al., 1990), French family (Pérusse et al., 1996), and Danish adoption studies (Stunkard et al., 1986) showed that obesity and fat distribution have a strong genetic susceptibility, with heritability estimates ranging from 40 to 70% for obesity risk and from

36 to 61% for waist-to-hip ratio (WHR) (Stunkard, 1986; Bouchard et al., 1990). It is important to note that the measured heritability depends on the environmental variance in the study population, which if low, can overestimate heritability. Recent research efforts in ethnically diverse individuals highlighted the genetic contribution even for changes in body weight after interventions such as metabolic surgeries (Fesinmeyer et al., 2013; Parikh et al., 2013). Although body mass index (BMI) is a standard measure of obesity, WHR reflecting central body fat distribution is the main predictor of obesity-related metabolic sequelae such as type 2 diabetes (T2D) or cardiovascular diseases (Manolopoulos et al., 2010). Being aware of the significant health burden associated with obesity, a better understanding of its complex pathophysiology including the genetic component remains a major challenge of the current obesity research.

Recent advances in high-throughput genotyping technologies allowed the development of powerful analytical tools like genome-wide association studies (GWASs) to explore novel genes and loci contributing to the genetic susceptibility of complex diseases. In the past decade, large-scale GWASs uncovered hundreds of genetic risk loci for BMI and WHR in European populations, making remarkable progress in our understanding of the genetics of these complex polygenic traits. A large meta-analysis of GWAS for BMI in ∼700,000 European individuals revealed over 750 BMI-associated single-nucleotide polymorphisms (SNPs), although only explaining 6.0% of the BMI variance (Yengo et al., 2018). The vast majority of BMI and obesity loci map to the non-coding genome; therefore, GWASs for the vast majority of cases do not usually identify specific genes but only susceptibility regions in the genome. Converting genetic risk loci into effector transcripts and function has been slow and has become a field of research in itself. Also, imperfect correlation between effect sizes measured from a population-based approach such as GWAS and effect sizes measured from a sibling approach suggests potentially a significant amount of indirect genetic effects (genetic nurture) captured in GWAS "direct" effects for a trait like BMI (Young et al., 2020). It is also noteworthy that currently established methods such as Genome-wide Polygenic Score (GPS), which estimates heritability of human complex traits in unrelated individuals using whole-genome sequencing data, further evidenced a missing heritability in BMI compared with assumptions based on earlier studies (Khera et al., 2019). Based on the fact that most of the previously reported GWAS were done in individuals of European ancestry, non-European ancestries may provide an attractive and very promising source for upcoming genetic studies aimed at identification of novel proposed mechanisms underlying the associations of genetic loci with obesity. Although these studies may contribute to a better understanding of the genetics of obesity and to decreasing the proportion of the missing heritability, there are several specific features which have to be considered in multi-ethnic genetic analyses. Due to the fact that the multifactorial character of obesity includes both genetics and a strong environmental component, the specified loci may have distinct contributions in different ancestral genetic backgrounds and different environments as presented through diet and exercise among other factors. An epidemiological survey found that people in Asia have lower obesity rates compared with people in Europe and the United States, but type 2 diabetes is more prevalent in Asia even with lower BMI (Yoon et al., 2006).

In the case of the United States, obesity rates also vary by ethnic groups or social classifications of race. Only 4.8% of Asian Americans (including Chinese, Japanese, Korean, Asian Indian, Vietnamese, Filipino, and others) over the age of 30 had obesity between 2001 and 2002. The prevalence was lower than in other ethnic groups (21.8% of European Americans, 34.8% of African Americans) living in the United States (Wang and Beydoun, 2007). The National Center for Health Statistics reported that the prevalence of obesity was lowest among non-Hispanic Asian adults (11.7%) and youth (8.6%), followed by non-Hispanic white (34.5%, 14.7%), Hispanic (42.5%, 21.9%), and non-Hispanic black (48.1%, 19.5%) in the United States between 2011 and 2014 (Ogden et al., 2015). A recent study by Commodore-Mensah et al. (2018), also confirmed the lowest prevalence of overweight/obesity in Asians, even after adjusting for WHO-recommended Asian-specific BMI cutoffs (overweight: 23–27.4 kg/m^2; obesity: $\geq$27.5 kg/m^2) in the United States between 2010 and 2016. The reason for these disparities are multifactorial including lifestyle, health care, income status, experience of discrimination, and changes in diet after migration to the United States; however, these data may also implicate the role of genetics and interactions between genetics and the environment in creating variation in obesity rates. Even though highly challenging, investigations of the genetic and environmental factors underlying variation in the pathophysiology of obesity are highly desirable as they could ultimately lead to improved knowledge of the causal mechanistic chains underlying the pathophysiology of this disease and its related metabolic sequelae.

Currently, there is increasing evidence indicating the potential role of genetic ancestry in variable predisposition to obesity in different environments. This review provides a comprehensive overview of recent advances in obesity genetics in individuals of Asian ancestry. In particular, we compared obesity susceptibility loci discovered in individuals of European and Asian ancestries and addressed the potential role of genetic variants in variation in obesity risk.

It has to be acknowledged that the information on ancestry variables in GWASs is commonly based on self-reported questionnaires. This practical way to adjust for ancestry in genetic association studies has been previously certificated to be sufficiently accurate for assessing population stratification in genetic association studies (Rosenberg et al., 2002). However, it may be misleading in comparisons of complex traits across populations and may overestimate polygenic adaptation due to residual population stratification (Sohail et al., 2019). Despite strong associations reaching *p*-values with genome-wide statistical significance, these analyses may be all subtly affected by population structure, leading to partly incorrect effect estimations (Sohail et al., 2019). Differences in genetic structure among populations are mostly due to genetic drift, natural selection, *de novo* mutations, and admixture.

Although this review focuses on East Asian populations, no ancestry (geographic) region can be considered in isolation

in terms of human population history because migrations between Asians and Europeans have had a substantial impact on current genetic structure. For instance, the ancient DNA studies showed that most present-day Europeans derive from at least three highly differentiated populations: west European hunter-gatherers, ancient north Eurasians from the Steppe, and early farmers from Anatolia, and that there are varying proportions of these different ancestries across Europe. In early Bronze Age pastoralists, West Eurasian ancestry and East Asian ancestry have already undergone genomic mixture through the Eurasian steppes (Mathieson et al., 2015; Damgaard et al., 2018). There is still a controversy about the genetic gradients in present-day Asians, where the ancestry variables may be more complicated and diverse. There are at least three genetic gradients in the South Asian region: Anatolian/Iranian farmer-related ancestry, Ancestral North Indians, and Ancestral South Indians who were mixed with northwestern and southeastern groups with Steppe ancestry (Narasimhan et al., 2019); at least four ancient populations in Southeast Asia: mainland Hoabinhians, Andamanese onge, Malaysian jehai, and ancient Japanese Ikawazu Jomon (Mccoll et al., 2018); and at least three ancient populations in East Asia (e.g., Japanese): Hondo, Ryukyu, and Ainu (Takeuchi et al., 2017).

We have to admit that although these aspects are not addressed in our review, the readers should be aware of them. Also, we do not address the diversity among South Asians, Southeast Asians, and East Asians, which is based on the following two points: (1) in 13 BMI-related genetic Asian studies (**Table 1**), only two studies included South Asians and South East Asians. In addition, the sample size was strongly limited as compared with East Asians (totally: East Asian: 483,795; South Asian and South East Asian: 12,033), which did not allow a valid comparison and drawing competent and robust conclusions; (2) although South Asians may appear closer to Europeans than East Asians from the genetic point of view [e.g., there are no observed systematic differences in risk allele frequencies of WHR-related loci between South Asians and Europeans (Scott et al., 2016)], the BMI and the degree of abdominal obesity and the risk of diabetes in South Asians are comparable with East Asians (Nanditha et al., 2016).

GENETIC STUDIES OF OBESITY BEFORE THE GWAS ERA

In the last two decades of the past century, physiologic (candidate) gene association studies and genome-wide linkage studies represented the major analytical tools employed in the identification of genetic determinants of complex polygenic traits. The success of these strategies was mostly limited by poor statistical power due to the small sample sizes of the studied cohorts. Whereas linkage studies turned out to be powerful in the identification of genes responsible for monogenic Mendelian traits and diseases, their impact on polygenic traits was rather moderate. In 1999, the 825 T polymorphism in the G Protein Subunit Beta 3 gene (*GNB3*) was found as one of the first BMI-related variants in Asian ancestry individuals and replicated afterward in cohorts of European and African ancestry (Siffert et al., 1999) (**Table 1**). Although *GNB3* polymorphisms were not associated with BMI in a Japanese cohort (Ohshiro et al., 2001), a recent large-scale multi-population meta-analysis disclosed associations of genetic variants in *GNB3* with being overweight/obese (Li et al., 2016). In 2005, a meta-analysis containing five genome-wide linkage scan studies provided significant evidence for the association of genetic variation in the lipoprotein lipase (*LPL*) and adrenoceptor beta 3 (*ADRB3*) with BMI (Johnson et al., 2005). Subsequently, in 2007, a larger well-powered genome scan meta-analysis summarized previous genome-wide linkage scans in individuals of European ancestry (Saunders et al., 2007). Although it has not explicitly shown specific loci associated with BMI or obesity, one of the strongest candidates was the FTO alpha-ketoglutarate dependent dioxygenase (*FTO*) locus along with uncoupling protein 1 (*UCP1*), leptin (*LEP*), insulin-like growth factor 1 (*IGF-I*), scavenger receptor class B member 1 (*SCARB1*), and insulin receptor substrate 2 (*IRS2*). It should be mentioned that associations of genetic variants in *UCP1* and *LEP* with obesity have further been replicated in cohorts of Asian ancestry (Nakano et al., 2006; Wang et al., 2006).

GWAS FOR BMI IN ASIAN POPULATIONS

Within the last decade, GWAS has emerged as a powerful tool to identify loci associated with complex polygenic diseases such as obesity. As yet, GWAS contributed to the identification of more than 750 loci reaching associations with BMI at the genome-wide significance level ($p < 10^{-8}$) (Yengo et al., 2018). Whereas most of the GWASs have been performed in cohorts of European ancestry (Loos et al., 2008; Thorleifsson et al., 2009; Speliotes et al., 2010; Pei et al., 2014; Locke et al., 2015; Winkler et al., 2015; Wood et al., 2016; Graff et al., 2017; Hoffmann et al., 2018; Riveros-Mckay et al., 2019), similar studies in cohorts of Asian ancestry were rather scarce. **Table 2** summarizes current obesity susceptibility loci exclusively associated with cohorts of Asian ancestry.

In 2009, Cho et al. (2009) reported the first large-scale two-stage GWAS for quantitative traits such as BMI and height in cohorts of East Asian ancestry. The study showed that the *FTO* gene locus, which has been well acknowledged as the major contributor to polygenic obesity in European populations (Frayling et al., 2007), also provided the most prominent association signal in East Asian cohorts. Further support came from Hotta et al. (2008) who found *FTO* variant rs1558902 significantly associated with obesity in a Japanese cohort as well. It should be pointed out that *FTO* variants have not been related to obesity and being overweight only in European and Asian but also in African (Monda et al., 2013), Hispanic (Villalobos-Comparán et al., 2008; Dong et al., 2011), and Native American populations (Rong et al., 2009), in both adults and children (Dina et al., 2007; Frayling et al., 2007), implicating the global impact of *FTO* polymorphisms on obesity. *FTO* is encoding FTO alpha-ketoglutarate-dependent dioxygenase and is widely expressed in multiple tissues throughout the body,

TABLE 1 | Studies conducted in cohorts of Asian ancestry.

Study type	Publication year	Sample size	Male/female	Cohort age Mean (SD)	Criteria for discovery[c]	Number of variants[d] (discovery stage)	Criteria for replication[e]	Number of variants[f] (replication stage)	Number of variants replicated from[h]	Number of variants successfully replicated from[h]	References
Candidate gene association study	1999	2056 East Asian	2056/0	24 (6)	NA	NA	$p < 5.0E{-}2$	NA	1	1	Siffert et al. (1999)
Candidate gene association study	2001	208 East Asian	118/90	50.2 (1.2)	NA	NA	$p < 5.0E{-}2$	NA	1	1	Ohshiro et al. (2001)
Candidate gene association study	2006	251 East Asian	251/0	25.5 (3.5)	NA	NA	$p < 5.0E{-}2$	NA	1	1	Nakano et al. (2006)
Candidate gene association study	2006	408 East Asian	135/273	59.4 (13.2)	NA	NA	$p < 5.0E{-}2$	NA	2	1	Wang et al. (2006)
GWAS	2009	16,703 (Dis[a]: 8842 East Asian; Rep[b]: 7861 East Asian)	7397/9306	54.4 (8.4)	$p < 1.0E{-}5$	2	$p < 5.0E{-}2$	1	NA	NA	Cho et al. (2009)
Fine mapping *FTO* study	2008	2427 East Asian	1077/1350	48.7 (15.4)	MAF > 10%	90	$p < 1.7E{-}4$	15	NA	NA	Hotta et al. (2008)
Replication study	2009	2865 East Asian	1420/1445	49 (14.7)	NA	NA	$p < 5.0E{-}2$	NA	27	11	Hotta et al. (2009)
Replication study	2010	7705 East Asian	3511/4194	49 (11.9)	NA	NA	$p < 5.0E{-}2$	NA	14	4	Ng et al. (2010)
Meta-analysis-GWAS	2012	83,048 (Dis: 22,762 East Asian, 4953 South and East Asian; Rep: 2118 South East Asian, 53,215 East Asian)	34,906/48,142	55.2 (9.9)	$p < 1.0E{-}4$	848	$p < 5.0E{-}7$	7	NA	NA	Wen et al. (2012)
GWAS	2012	10,391 (2431 South East Asian, 5429 East Asian, 2531 South Asian); 1006 (1006 Chinese)	5185/5297; 512/492	57.3 (10.6); 9	$p < 5.0E{-}2$	31	$p < 5.0E{-}2$	13	NA	NA	Dorajoo et al. (2012)
GWAS	2012	62,245 (Dis[a]: 26,620 East Asian Rep[b]: 35,625 East Asian)	35,870/26,375	58 (12.1)	$p < 5.0E{-}5$	36	$p < 5.0E{-}8$	7	NA	NA	Okada et al. (2012)
Meta-analysis-GWAS	2014	134,091 (Dis[a]: 86,739 East Asian, 4301 South East Asian; Rep[b]: 47,352 East Asian)	60,628/73,463	55.4 (9.8)	$p < 7.59E{-}6$	8	$p < 5.0E{-}2$	4	55	26[g]	Wen et al. (2014)
GWAS	2017	173,430 (Dis[a]: 158,284 East Asian; Rep[b]: 15,146 East Asian, 322,154 European)	90,992/82,438	59.1 (11)	$p < 5.0E{-}8$; $p < 1.0E{-}6$	72; 134	$p < 5.0E{-}8$	85 (51 novel)	163	66	Akiyama et al. (2017)

[a]*Discovery stage sample.* [b]*Replication stage sample.* [c]*Criteria for discovery stage.* [d]*Number of variants found in the discovery stage by* [c] *criteria.* [e]*Criteria for replication in Asian study (also as a criteria for other type studies).* [f] *Number of variants found in the replication stage by* [e] *criteria.* [g]*26 variants replicated by p < 1E–3 from 55 variants which were identified in previous European GWASs.* [h]*European GWAS (Frayling et al., 2007; Loos et al., 2008; Thorleifsson et al., 2009; Willer et al., 2009; Speliotes et al., 2010; Wen et al., 2012; Berndt et al., 2003; Guo et al., 2013; Locke et al., 2015; Shungin et al., 2015; Winkler et al., 2015). NA, not applicable.*

TABLE 2 | Obesity susceptibility loci identified ($p < 5.0E{-}8$) in cohorts of Asian ancestry.

SNP	Candidate gene(s)[a]	Chr.[b]	Allele[c] ALT/REF[d]	RAF[e,f]	Beta-estimates[c] (SE)[e]	*p*-value	References	Explained variance (%)[f]
rs2237892	*KCNQ1*	11	T/C	0.355	0.0298 (0.0042)	9.29E−13	Wen et al. (2014)	0.041
rs671	*ALDH2*	12	A/G	0.267	−0.0378 (0.0057)	3.40E−13	Wen et al. (2014)	0.056
rs12229654	*MYL2*	12	G/T	0.224	−0.0341 (0.0058)	4.56E−09	Wen et al. (2014)	0.040
rs2076463	*FGR,IFI6*	1	G/A	0.343	−0.023 (0.004)	1.68E−08	Akiyama et al. (2017)	0.024
rs77489951	*LOC101929596,HNRNPLL*	2	T/C	0.06	0.044 (0.008)	9.39E−09	Akiyama et al. (2017)	0.022
rs8192473	*CCK*	3	T/C	0.109	−0.035 (0.006)	3.58E−09	Akiyama et al. (2017)	0.024
rs4308481	*PRDM6,CEP120*	5	C/T	0.398	0.021 (0.008)	1.00E−18	Akiyama et al. (2017)	0.021
rs183975233	*HLA-DRA,HLA-DRB5*	6	T/A	0.689	−0.031 (0.004)	7.51E−16	Akiyama et al. (2017)	0.041
rs148546399	*EYS*	6	A/G	0.01	0.050 (0.008)	1.13E−09	Akiyama et al. (2017)	0.005
rs143665886	*LINC01392,TFEC*	7	C/T	0.4	0.022 (0.004)	9.46E−09	Akiyama et al. (2017)	0.023
rs10868215	*SLC28A3,NTRK2*	9	C/T	0.299	−0.021 (0.004)	1.34E−08	Akiyama et al. (2017)	0.018
rs10795945	*CDC123,CAMK1D*	10	C/T	0.461	0.021 (0.003)	1.10E−09	Akiyama et al. (2017)	0.022
rs80117551	*HERC4*	10	C/T	0.679	−0.022 (0.004)	1.57E−08	Akiyama et al. (2017)	0.021
rs12569457	*FRAT2,RRP12*	10	T/C	0.122	0.025 (0.004)	6.67E−09	Akiyama et al. (2017)	0.013
rs1907240	*MIR5694,FGFR2*	10	G/A	0.393	−0.024 (0.004)	3.47E−11	Akiyama et al. (2017)	0.027
rs80234489	*FAM60A*	12	A/C	0.812	−0.031 (0.005)	1.05E−11	Akiyama et al. (2017)	0.029
rs75766425	*NID2*	14	C/G	0.105	0.034 (0.005)	1.28E−10	Akiyama et al. (2017)	0.022
rs4788694	*ZFHX3*	16	C/G	0.179	−0.021 (0.004)	2.54E−08	Akiyama et al. (2017)	0.013
rs180950758	*SUZ12P1*	17	T/A	0.11	0.027 (0.005)	2.63E−08	Akiyama et al. (2017)	0.014
rs1379871	*DMD*	X	C/G	0.68	0.018 (0.003)	1.05E−08	Akiyama et al. (2017)	0.014
rs6529684	*HSD17B10,HUWE1*	X	G/A	0.54	0.016 (0.003)	2.78E−08	Akiyama et al. (2017)	0.013
rs3121672	*IL13RA1*	X	C/T	0.43	0.024 (0.003)	2.90E−17	Akiyama et al. (2017)	0.028
rs1190736	*GPR101*	X	C/A	0.65	−0.017 (0.003)	1.31E−08	Akiyama et al. (2017)	0.013
rs5945324	*FAM58A,DUSP9*	X	C/G	0.4	0.022 (0.003)	1.33E−11	Akiyama et al. (2017)	0.023
rs2206271	*TFAP2B*	6	A/T	0.365	0.031 (0.008)	3.00E−18	Akiyama et al. (2017)	0.045
rs2495707	*HIF1AN,PAX2*	10	A/G	0.549	0.025 (0.008)	1.00E−09	Akiyama et al. (2017)	0.031
rs60808706	*KCNQ1*	11	A/G	0.369	0.046 (0.004)	1.24E−38	Akiyama et al. (2017)	0.099
rs3205718	*FAIM2*	12	T/C	0.231	0.023 (0.008)	4.62E−10	Akiyama et al. (2017)	0.019
rs7305242	*ALDH2,MAPKAPK5-AS1*	12	C/T	0.576	−0.021 (0.004)	2.21E−08	Akiyama et al. (2017)	0.022
rs2540034	*ADCY9*	16	T/C	0.312	0.028 (0.004)	2.97E−12	Akiyama et al. (2017)	0.034
rs35560038	*GIPR,QPCTL*	21	A/T	0.532	0.054 (0.008)	3.00E−52	Akiyama et al. (2017)	0.145

[a]*Predicted gene reported in reference studies.* [b]*Chromosomes based on NCBI Build154 (GRCh38).* [c]*Alternative alleles were treated as effective allele.* [d]*The allele frequency based on genome Aggregation Database (gnomAD).* [e]*"Standard error" according to reference studies reported.* [f]*"Explained variance" is the variance explained by each reported variant using the formula which uses the allele frequency (f) estimated in GWAS and estimates of the additive effect (β) in meta-analysis: explained variance* β^2 $(1 - f)$ $2f$ *(Akiyama et al., 2017). To estimate the additive explained variance of 31 newly identified BMI loci in Asian population, the explained variance of each individual variant were summed up and resulted in a total of 0.926%.*

in particular, in the thalamic arcuate nucleus with the central role in body weight regulation (Gerken et al., 2007). It should be recognized that the mechanistic basis for the *FTO*-related association with obesity has been finally explained in 2015 by Claussnitzer et al. (2015). The authors showed that the functional *FTO* variant disrupted an evolutionarily conserved motif of AT-Rich Interaction Domain 5B (ARID5B) repressor, which leads to the loss of binding, releases of a potent preadipocyte super-enhancer, and activation of downstream targets Iroquois Homeobox 3 and 5 (IRX3 and IRX5) (Claussnitzer et al., 2015). This results in alterations of mechanisms controlling the shift from white adipocyte browning to lipid-storage gene expression programs, repression of basal mitochondrial respiration, decrease in thermogenesis in response to stimulus, and increase in adipocyte size, which ultimately results in human obesity (Claussnitzer et al., 2015).

Hotta et al. (2009) reported the first Japanese study aimed to replicate the association signals from BMI GWAS in individuals of European descent. The study indicated that SEC16 homolog B (*SEC16B*), transmembrane protein 18 (*TMEM18*), glucosamine-6-phosphate deaminase 2 (*GNPDA2*), brain-derived neurotrophic factor (*BDNF*), fas apoptotic inhibitory molecule 2 (*FAIM2*), and melanocortin 4 receptor (*MC4R*) loci are not only associated with BMI in European ancestry individuals but also with obesity in Japanese ancestry individuals. On the other hand, 16 obesity-related SNPs could not be

replicated in this study, supporting the heterogeneity of genetic susceptibility to obesity among various genetic ancestries. For instance, in contrast to the European cohorts, genes such as phosphotriesterase related (*PTER*) and secretogranin III (*SCG3*) were monomorphic for the respective variants in the studied Asian cohort. One of the genes whose polymorphisms were replicated in this study was *SEC16B*. *SEC16B* encodes long (Sec16L) and short (Sec16S) proteins required for mammalian cells to deliver intracellular substances from the endoplasmic reticulum to the Golgi apparatus (Watson et al., 2006; Bhattacharyya et al., 2007). Although Schmid et al. (2012) showed that *Sec16b* has the highest expression in subcutaneous adipose tissue and the lowest expression in the hypothalamus, Hotta et al. (2009) proposed that *Sec16b* expressed in the hypothalamus might be affecting energy regulation. *SEC16B* is not only an obesity susceptibility locus in individuals of Asian and European ancestry but also is related to BMI in individuals of African ancestry (Sahibdeen et al., 2018). Also, the polymorphism of *Tmem18*, which is highly expressed in the hypothalamus (Schmid et al., 2012), is one of the BMI-related loci being robustly replicated in individuals of Asian ancestry. A study focusing on *Tmem18* expression in the hypothalamic nucleus showed that *Tmem18*-deficient mice gain body weight compared with a control mouse, especially in males under a strict high-fat diet (Larder et al., 2017). Overexpression of *Tmem18* in hypothalamic paraganglia may affect food intake, increase energy expenditure, and reduce systemic fat and body weight.

Another gene highly expressed in the hypothalamus is *GNPDA2*. Genetic variants in or near *GNPDA2* have been shown to be associated with obesity in Asians (Hong et al., 2013), Pima Indians (Muller et al., 2019), and Europeans (Willer et al., 2009). In contrast, there are controversial data in childhood obesity; it has been shown that the *GNPDA2* locus is associated with BMI in a cohort from Mexico (León-Mimila et al., 2013), but not in an Asian cohort (Wang et al., 2012). *MC4R* is also a centrally acting gene known to be the most common cause of monogenic obesity in extreme childhood obesity. It is well recognized that hypothalamic pro-opiomelanocortin neurons regulate feeding behavior through the production of melanocortins and beta-endorphin from these neurons. MC4R is a major melanocortin receptor involved in regulating food intake and energy expenditure (Nogueiras et al., 2007). The *MC4R* has been reported as a risk gene associated with extreme obesity in adolescence and adulthood (Chambers et al., 2008; Hotta et al., 2009; Tenesa et al., 2009; Thorleifsson et al., 2009). Short-term administration of an MC4R agonist RM-493 increased individual resting energy expenditure and limited fat oxidation in obese individuals (Chen et al., 2015). However, two other clinical studies using MC4R agonists did not show any effects of regulating body weight (Krishna et al., 2009; Royalty et al., 2014). Further studies are needed in other to clarify the controversial findings reported.

In 2010, Ng et al. (2010) carried out a replication study of 12 BMI-associated loci from a European ancestry GWAS in a Chinese cohort. Five loci located at or near *GNPDA2*, *BCDIN3* domain containing RNA methyltransferase (*BCDIN3D*), SH2B adaptor protein 1 (*SH2B1*), *FTO*, and potassium channel tetramerization domain containing 15 (*KCTD15*) seem to be related to BMI in Chinese individuals. Two of them, *SH2B1* and *KCTD15* polymorphism (rs7498665 and rs29941), were replicated for the first time in an Asian cohort. *SH2B1* encodes SH2B adaptor protein 1, a member of the SH2-domain containing mediators family. It is expressed in both central and peripheral tissues (Ren et al., 2007). A study showed that central *Sh2b1* controls glucose homeostasis and insulin sensitivity (Duan et al., 2004) as well as hypothalamic leptin sensitivity (Ren et al., 2007). Peripheral *Sh2b1* regulates insulin sensitivity and glucose metabolism (Ren et al., 2007) whereas hepatic *Sh2b1* regulates lipid metabolism, particularly triacylglycerol and very-low-density lipoprotein content in the liver (Sheng et al., 2013).

The function of the *KCTD15* is still unknown. However, it has been shown that *Kctd15* deficiency resulted in a slow-growth/small size phenotype in zebrafish (Heffer et al., 2017). Particularly, the Kctd15 likely acts through interaction with adipocyte protein 2 (AP-2) (Liu et al., 2013), which is a critical regulator in adipogenesis (Shan et al., 2013), suggesting a possible molecular basis for the observed associations of *KCTD15* variants with obesity.

In 2012, Dorajoo et al. (2012) performed a BMI GWAS meta-analysis in Asian ancestry individuals (Singaporean, Malay, and Asian-Indian) and, among others, confirmed the relevance of the *FTO* locus. The authors replicated 13 loci which have been previously reported in European ancestry cohorts and found three novel variants (rs2287019, rs2241423, rs516175) associated with BMI in their Asian cohort. Interestingly, 16 loci previously found in the European ancestry GWAS were not associated with BMI in this study, possibly due to the genetic heterogeneity between present-day Asian and European ancestries. Rs2287019 variant maps in the vicinity of the *GIPR*, the gene encoding a G protein-coupled receptor for a gastric inhibitory polypeptide, which is strongly expressed in pancreatic beta cells (Saxena et al., 2010). It is involved in the incretin effect and in early pathophysiologic pathways that could lead to impaired glucose tolerance and T2D in humans. *Gipr*-deficient mice are more resistant to obesity after a high-fat diet (Miyawaki et al., 2002), which is likely due to the interplay of enhanced insulin sensitivity and inhibition of GIP signaling pathways in adipose tissue (Joo et al., 2017). *MAP2K5*, mitogen-activated protein kinase kinase 5, which is the closest gene to the BMI-related loci (rs2241423), plays a crucial role in the MAPK signaling pathway. Chen et al. (2014) showed that MAP2K5 is regulated by mir-143 and affects lipogenesis. Methionine sulfoxide reductase A, *MSRA*, located near the previously mentioned variant rs516175, regulates glucose metabolism and insulin response in mitochondria and has protective effects on insulin sensitivity in obese mice (Hunnicut et al., 2015). It is also a target of miR-193b which stimulates reactive oxygen species signal transduction and regulates lip sarcoma cell survival and adipose tissue–derived stromal/stem cells cell differentiation (Mazzu et al., 2017).

In 2012, a two-stage GWAS (Okada et al., 2012) in an East Asian cohort discovered two novel loci, nearby CDK5 regulatory subunit associated protein 1 like 1 (*CDKAL*) and kruppel like factor 9 (*KLF9*), which were associated with BMI.

The study clearly implicated ancestry-specific effects of the *KLF9* locus, which was not found in previous analyses in European individuals (Speliotes et al., 2010), despite the sufficient statistical power to detect the locus based on the assumption of the same effect size and allele frequencies (MAF_{Eur} = 0.5, MAF_{Asi} = 0.4). A three-stage meta-analysis of eight BMI GWAS was performed, with the second phase being computer replication and the third phase being a *de novo* replication study (Wen et al., 2012). The analysis resulted in 10 loci reaching associations at genome-wide significance ($p < 10^{-8}$). Seven of the ten loci are at the *FTO*, *SEC16B*, *MC4R*, *GIPR*/glutaminyl-peptide cyclotransferase like (*QPCTL*), adenylate cyclase 3 (*ADCY3*), *BNDF*, and *MAP2K5*, which have been previously shown to be associated with BMI in European ancestry individuals. Three novel loci in or near cyclin-dependent kinase 5 (*CDKAL1*), proprotein convertase subtilisin/kexin type 1 (*PCSK1*), and glycoprotein 2 (*GP2*) associated with BMI in an East Asian cohort. Kim et al. (2013) identified the prospero homeobox 1 (*PROX1*) locus in a GWAS for BMI in a cohort from Mongolia and replicated it in a cohort from Korea. However, the associations only reached suggestive significance with $p < 10^{-7}$. The limited statistical power was likely attributed to the relatively small sample size (n = 1301). Albeit not statistically significant at the genome-wide level, the study also suggested protein tyrosine phosphatase receptor type D (*PPTRD*) and reelin (*RELN*) to be potential candidate genes that may have a role in the development of obesity.

In 2014, a two-stage GWAS (Wen et al., 2014) including 82,438 East Asian and 4301 South-East Asian individuals in the discovery and 47,352 East Asian individuals in the replication stage indicated four novel BMI-related loci reaching a significant level of genome-wide association: these loci in or near potassium voltage-gated channel subfamily Q member 1 (*KCNQ1*), aldehyde dehydrogenase 2 family member (*ALDH2*), inter-alpha-trypsin inhibitor heavy chain 4 (*ITIH4*), and 5′-nucleotidase cytosolic II *(NT5C2*).

KCNQ1 variant (rs2237892) was initially reported in GWAS of T2D in Asian cohorts (Unoki et al., 2008; Yasuda et al., 2008), followed by replication reports in European cohorts (Unoki et al., 2008; Voight et al., 2010). Moreover, *KCNQ1* locus has been shown to be associated with waist circumference (WC adjusted for BMI) in an Asian cohort (Graff et al., 2017). KCNQ1 is expressed in islet cells and has been implicated in the regulation of insulin secretion (Ullrich et al., 2005).

ALDH2 polymorphism rs671 is not only related to obesity but also to multiple complex traits such as drinking behavior (Jorgenson et al., 2017), triglycerides (Tan et al., 2012), and blood pressure (Feitosa et al., 2018). As suggested by Akiyama et al. (2017), the *ADH-ALDH* gene family may have a greater significant impact on BMI in East Asian individuals. Recent studies (Yu et al., 2016) suggested that ALDH2 is a positive regulator of adipocyte differentiation through the interaction with its upstream regulatory factor protein kinase C mediated by peroxisome proliferator-activated receptor gamma transcriptional activity. *ITIH4* is widely distributed in the blood and liver (Cai et al., 1998). The gene locus has been associated with schizophrenia (Goes et al., 2015) and blood serum protein levels in several GWAS (Emilsson et al., 2018). Early studies (Fujita et al., 2004) suggested that *ITIH4* locus is also associated with hypercholesterolemia in a Japanese cohort. *NT5C2* encodes a downstream cytosolic hydrolase that plays a considerable role in cellular purine metabolism by acting primarily on inosine 5′-monophosphate and other purine nucleotides (Novarino et al., 2014).

In 2017, Akiyama et al. (2017) implemented, so far, the largest imputation-based GWAS in 158,284 East Asians. They reported 112 BMI loci, 61 of which were novel, and pointed out that BMI-related loci are most likely shared among different ancestries; however, the effects of particular loci on BMI may vary among genetic ancestries.

COMPARISON OF BMI SUSCEPTIBILITY LOCI BETWEEN EUROPEAN AND ASIAN ANCESTRIES

Most BMI-associated loci initially uncovered in studies with individuals of European ancestry have been widely replicated in Asian individuals (Frayling et al., 2007; Hotta et al., 2008; Cho et al., 2009; Yajnik et al., 2009; Dong et al., 2011; Moore et al., 2012; Vasan et al., 2012; Monda et al., 2013).

By reviewing all current BMI-related studies in Asian cohorts (**Table 1**), we found 92 loci (**Supplementary Table S1**) and compared them with GWAS in European cohorts. Forty-two of 92 BMI-related loci have been previously reported in European cohorts with $p < 5 \times 10^{-8}$ and had a consistent direction of effect on BMI. For the remaining 50 BMI-related loci, we observed no compelling evidence of replication (**Supplementary Table S1**). According to our defined criteria ($p < 5 \times 10^{-8}$ for GWAS, $p < 0.05$ for replication), the replications failed in the following cases: (1) 6 of 50 SNPs reached the genome significant p-value ($p < 5 \times 10^{-8}$) in Asian cohorts but not in European cohorts. Because it is unlikely that limited statistical power due to small sample size and minor allele frequencies would be a crucial factor (see **Supplementary Table S1**) explaining the failed replications, other reasons such as genetic heterogeneity or distinct phenotypic expression in different genetic ancestries may be considered. Exemplarily, East Asians showed a lower mean BMI (22.7 ± 3.59 kg/m^2) (Okada et al., 2012) than European cohorts (27.24 ± 3.9 kg/m^2) (Speliotes et al., 2010). (2) Twenty-three of 50 SNPs had a genome-wide significant p-value ($p < 5 \times 10^{-8}$) in Europeans, but no significant associations ($p > 0.05$) in Asian populations. Eighteen of these 23 variants were only directionally consistent but not significantly associated with BMI, and the remaining five SNPs were neither directionally consistent nor statistically significantly associated with BMI ($p > 0.05$) in Asian individuals. Limited statistical power could be a likely explanation for this observation. Compared with some large-scale studies in European cohorts with sample sizes ranging from 100,000 to 700,000, these Asian studies that failed to replicate the 23 variants were relatively small (1000 to 10,000). Two of the 23 loci non-replicated variants may have been due to marked differences in MAF between European and Asian individuals, as the MAF of rs17381664 and rs925946 were 0.002 and 0.06, respectively, in Asian individuals, and 0.37 and 0.29

in Europeans. Another possibility to be taken into account could be different causal variants between Asian and European individuals, resulting in a weak LD pattern in the Asian cohort, consequently leading to a weaker correlation between causal variants and marker SNPs. (3) For the remaining 21 SNPs, there was no convincing evidence for association with BMI in any of the two populations.

In summary, the majority of BMI-associated loci overlap between studies in Asian and European cohorts with regard to the respective risk alleles, although their frequencies may slightly vary. We found 82 BMI susceptibility loci (results not shown) by screening associations with a p-value $< 10^{-8}$ in individuals from Asia, after pruning by checking linkage disequilibrium (LD) through LD proxy module in a public LD online database from the National Institutes of Health (Division of Cancer Epidemiology & Genetics, 2020). We finally found 31 BMI loci that have only been associated with Asian cohorts (**Figure 1** and **Table 2**). Of these 31 loci, two of them were monomorphic in European subjects (eyes shut homolog (*EYS*)—rs148546399 and nidogen 2 (*NID2*)—rs75766425) and eight were rare mutations [FGR proto-oncogene (*FGR*)—rs2076463, heterogeneous nuclear ribonucleoprotein L like (*HNRNPLL*)—rs77489951, cholecystokinin (*CCK*)—rs8192473, transcription factor EC (*TFEC*)—rs143665886, *LOC102724612*—rs77636220, FRAT regulator of WNT signaling pathway 2 (*FRAT2*)/ribosomal RNA processing 12 homolog (*RRP12*)—rs12569457, fibroblast growth factor receptor 2 (*FGFR2*)—rs1907240, and *ALDH2*—rs7305242]. *NID2*, *FGFR2*, and *ALDH2* had already been reported in the latest T2D GWAS in an East Asian cohort (Spracklen et al., 2020). On the other hand, five loci were monomorphic or rare in Asian cohorts [*FTO*—rs9930333, LDL receptor related protein 1B (*LRP1B*)—rs2890652, cell adhesion molecule 2 (*CADM2*)—rs13078807, solute carrier family 39 member 8 (*SLC39A8*)—rs1310735, and protein kinase D1 (*PRKD1*)—rs11847697]. These variants of the 31 loci associated with GWAS in Asian cohorts explained only 0.926% of the phenotypic variance (**Table 2**).

We also integrated four GWAS works (Kang et al., 2010; Ng et al., 2012; Salinas et al., 2016; Chen et al., 2017) (**Supplementary Table S2**) conducted in African ancestry individuals (including African American and Afro-Caribbean and sub-Saharan African). Only one novel SNP (rs80068415) reached a GWAS significant threshold and five novel loci showed suggestive association with BMI at $p < 1 \times 10^{-5}$ in the studied African cohorts (**Supplementary Table S3**). The variant (rs80068415) identified in this GWAS only explained 0.065% of the variance (**Supplementary Table S3**). Because of LD patterns, six loci contained two SNPs in high LD with previously identified index SNP related to BMI with $p < 5 \times 10^{-8}$ in a European cohort. The variant (rs2033195) was in high LD with rs10055843 ($r^2 = 0.96$) which are associated with BMI at $p = 7.2 \times 10^{-18}$ in European individuals. The variants rs815611 and rs1346482 were in high LD ($r^2 = 0.92$), and the latter was associated with BMI ($p = 2 \times 10^{-19}$) in European cohorts. The underlying susceptibility locus of potentially African-specific rs80068415 is located in the region of semaphoring-4D (*SEMA-4D*). The proposed mechanism behind *SEMA-4D* on obesity is likely complex and may be mediated through regulatory multiple biological processes. Obesity usually follows a chronic inflammatory condition and T-cell accumulation has a positive correlation with adiposity. SEMA-4D seems to be a key player in the activation and differentiation of T cells and SEMA-4A could promote T helper 1 (Th1) cell differentiation (Worzfeld and Offermanns, 2014). This novel variant (rs80068415) seems to be highly specific to Africans as it is monomorphic in other populations.

The *FTO* locus manifests the strongest association signal with obesity in both Asian and European populations. Although the effect direction is consistent, the number of genetic variants varies between the populations. Nineteen *FTO* variants reached a genome-wide significance level for association with BMI in Europeans, whereas only four variants were associated with Asians. For instance, the top BMI-associated *FTO* signal found in GWAS in European individuals was rs1558902 ($p = 4.8 \times 10^{-120}$), whereas rs11642015 ($p = 2.04 \times 10^{-81}$) was the prominent hit in the Asian population. It is evident that differences in genetic architecture (e.g., rs9930333 with $p = 10^{-103}$ is the only polymorphism in Europeans) and evolutionary selective pressure (Liu et al., 2015) may at least partially explain the observed differences in associations at the variants level. However, it is worth mentioning that GWAS in Asian cohorts have emerged recently and genetic association studies have mostly focused on replication of previously reported signals from other GWAS. Furthermore, the reported studies in Asian cohorts are limited by a relatively small sample size compared with studies in European cohorts. Nevertheless, there is an enormous potential for large-scale genome-wide studies in cohorts of Asian ancestry, which may lead to the identification of novel players in the genetic architecture of human obesity.

Although most of the BMI associated loci showed consistent effect directions between Asians and Europeans (**Figure 2**), the sample effect sizes differ substantially. The frequency of risk alleles varies from 1 to 40% between the genetic ancestries. For instance, the allele frequency of the *FTO*-rs12149832 obesity risk alleles differs by 40%, whereas the effect size on BMI is comparable. In contrast, the frequency of variants in *MC4R* (rs571312) or *ADCY3/DNAJC27* (rs713586) in the European populations is similar to that in the East Asian populations based on genome Aggregation Database (gnomAD), but the difference in effect size on BMI accounted for about 20% (**Figure 3** and **Supplementary Table S1**).

COPY NUMBER VARIATIONS IN OBESITY IN ASIAN COHORTS

Along with the SNPs, copy number variations (CNVs), which are not only abundant in the human genome (Tuzun et al., 2005) but also have a vital influence on gene expression (Stranger et al., 2007), have emerged as another critical genetic label continuously attracting researchers' attention in the field of complex polygenic traits. Their "gene dosage" effect mediates the risk or protection against human diseases such as obesity (Jacquemont et al., 2011). One of the important CNVs related to BMI in the Asian population was reported on 10q11.22 (Sha

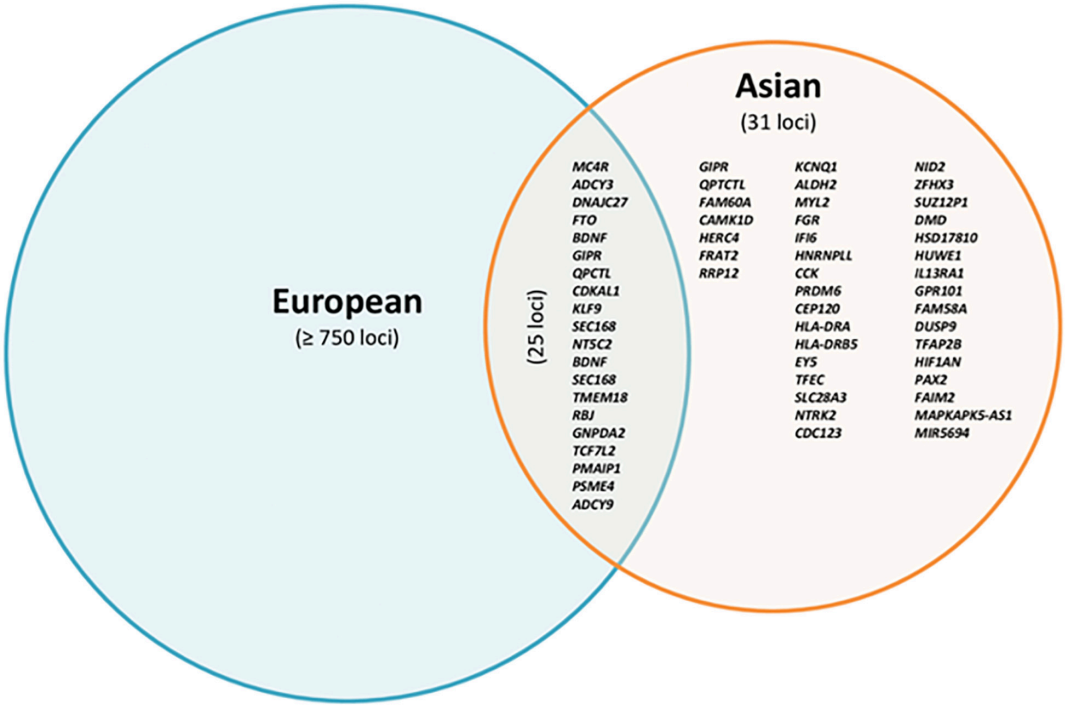

FIGURE 1 | Overlap of reported loci associated with BMI in European and Asian cohorts based on p-value $< 10^{-8}$.

et al., 2009). Pancreatic polypeptide Y receptor Y4 (*PPYR1*) located under this CNV area appears to be a plausible gene eventually related to obesity, and Shebanits et al. (2018) have also found similar findings in a Swedish cohort which suggested an association of *PPYR1 (NPY4R)* with WC in women. PPYR1 is one of the receptors of pancreatic polypeptide (PP), and several studies demonstrated that PP regulates the food intake via PPYR1 (Batterham et al., 2003). Yang et al. (2013) confirmed the association of a CNV on 16p12.3 with obesity-related phenotypes in a European but not in Asian cohort and suggested G protein-coupled receptor class C group 5 member B (*GPRC5B*) as a candidate obesity gene mapping within this chromosomal region. The authors emphasized the necessity of considering various ancestries in genetic association studies, particularly CNVs, which are characteristic for their considerable variation across genetic ancestries. Sun et al. (2013) tested eight CNVs (2p11.2, 10q11.22, 11q13.4, 16p11.2, 5p15.33, 15q11.2, 8q24.3) in young Chinese subjects, but proved only CNV 8q24.3 being associated with obesity, whereas no significant association was found for the other seven CNV candidates. *BAI1*, brain-specific angiogenesis inhibitor 1, which is located within the CNV 8q24.3, is postulated to be a member of the secretin receptor family and is the only member in its family transcriptionally regulated by p53 (Van Meir et al., 1994). Zhang et al. (2015) replicated three obesity-related loci 10q11.22, 4q25, and 11q11 in Han Chinese children and noted the strong cumulative effect of these loci on the risk of obesity. Furthermore, they also pointed out a significant interplay between CNVs (10q11.22) and dietary behaviors (meat-based). On the other hand, the salivary amylase gene (*AMY1*), whose copy number has been positively correlated with salivary amylase protein level (Perry et al., 2007), has delivered rather inconsistent findings concerning obesity. Perry et al. (2007) suggested that more *AMY1* copies exist in populations with high-starch diets than in those with traditionally low-starch diets. This points to a restricted selection of *AMY1* copies through a dietary shift early during human evolutionary history, especially in some ethnic groups such as East Asians known to prefer high-starch diets. However, a recent study failed to support the association of *AMY1* and *AMY2A* CNVs with obesity in two East Asian cohorts (Yong et al., 2016). A similar conclusion was drawn by Usher et al. (2015) who did not find any association of *AMY1* CNVs with obesity or BMI in a study including three European cohorts. Nevertheless, despite lacking evidence of an association between *AMY* CNVs and obesity, these studies inspired and promoted a novel perspective for future genetic association studies for obesity whereby variation in diet or environment exposures and their interaction with our genomes need to be considered.

FUTURE PERSPECTIVES

Gene × Environment Interaction

Obesity is a complex disease affected by both environment and genes. Because of increasing globalization, urbanization, and improved economic status, human diet structure and life habits have changed in Asia. Precisely, it has been observed that increased availability of food, better transport facilities, better healthcare facilities, reduced physical activity by mechanization,

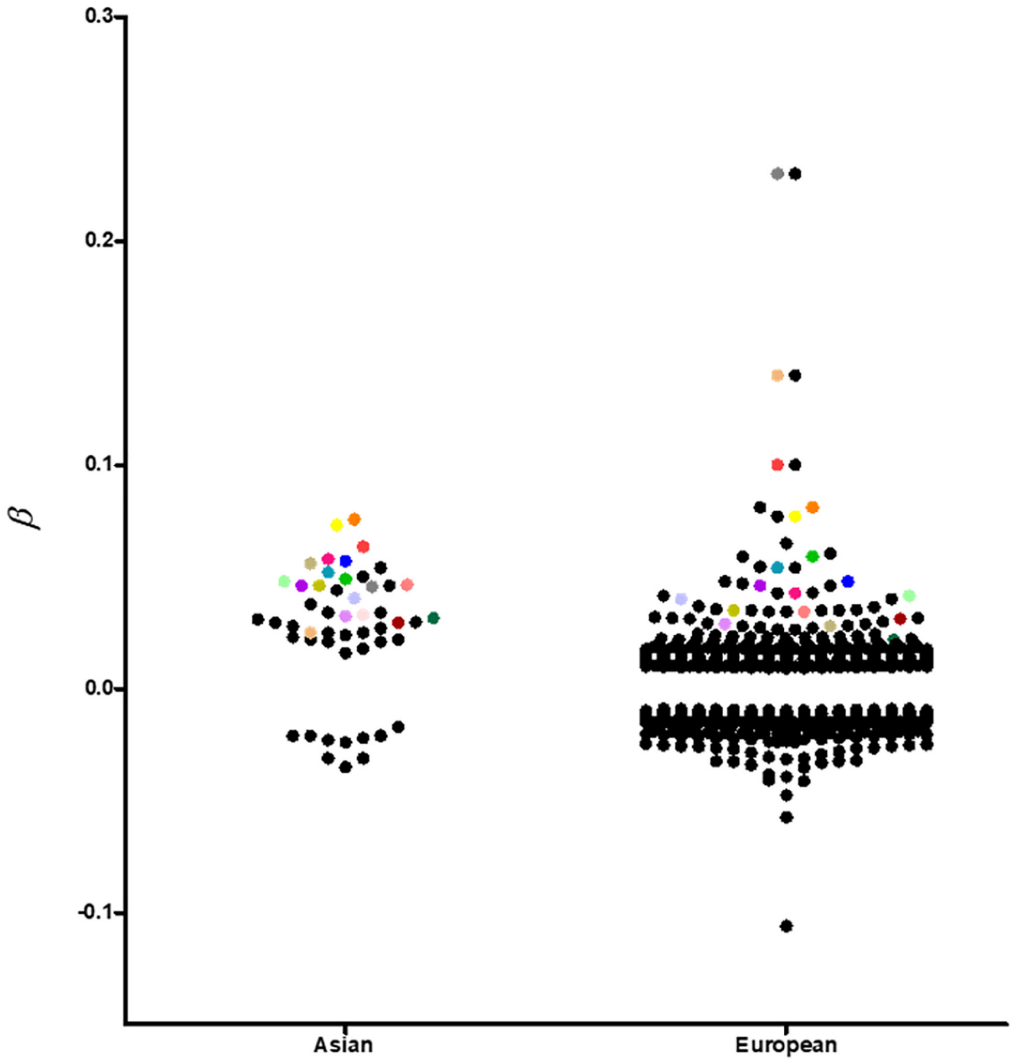

FIGURE 2 | A visual representation of all replicated and non-replicated loci at *p*-value $< 10^{-8}$ in GWAS conducted in individuals of European and Asian ancestries. Replicated loci are coded with the same color.

preference of viewing TV and videos (sedentary style), and increased use of automobiles and these changes in their life habits are associated with increased obesity prevalence in urban and rural populations, particularly in developing countries. Moreover, it is important to emphasize that there are also changes in their diet structure, such as a tendency to eat more finely processed carbohydrates (such as rice) and fat-rich items (Ramachandran et al., 2012). While the societal scale environment could cause the obesity epidemic, it is also known that genetic differences underlie the variation in BMI between individuals and that gene × environment interactions may be important in this context. A recent study concluded that nutrition has the strongest environmental effect on obesity risk at the *FTO* locus. Using genetic, anthropometric, and lifestyle variables collected as part of the UK Biobank, they assessed gene-by-environment interactions and how they modify the effect of *FTO* variants on BMI. The authors reported significant interactions between rs1421085 and a number of lifestyles and environmental factors, including alcohol, consumption, and mean sleep duration, with overall diet having the strongest effect on modifying *FTO* risk (Young et al., 2016). There is no doubt that gene–environment interactions are necessary to be understood to explain the underlying pathophysiology of obesity as a complex disease across the genetic diversity present in contemporary individuals across the globe.

Rare Genetic Variants

The loci associated with obesity remain to be further investigated, as the currently known loci only explain a small fraction of the variation in obesity and its measures such as BMI. Whereas common polymorphisms have been the main target of the majority of large-scale genetic studies so far, rare genetic (low frequency) variants with significant effects may substantially contribute to our understanding of the genetic heterogeneity

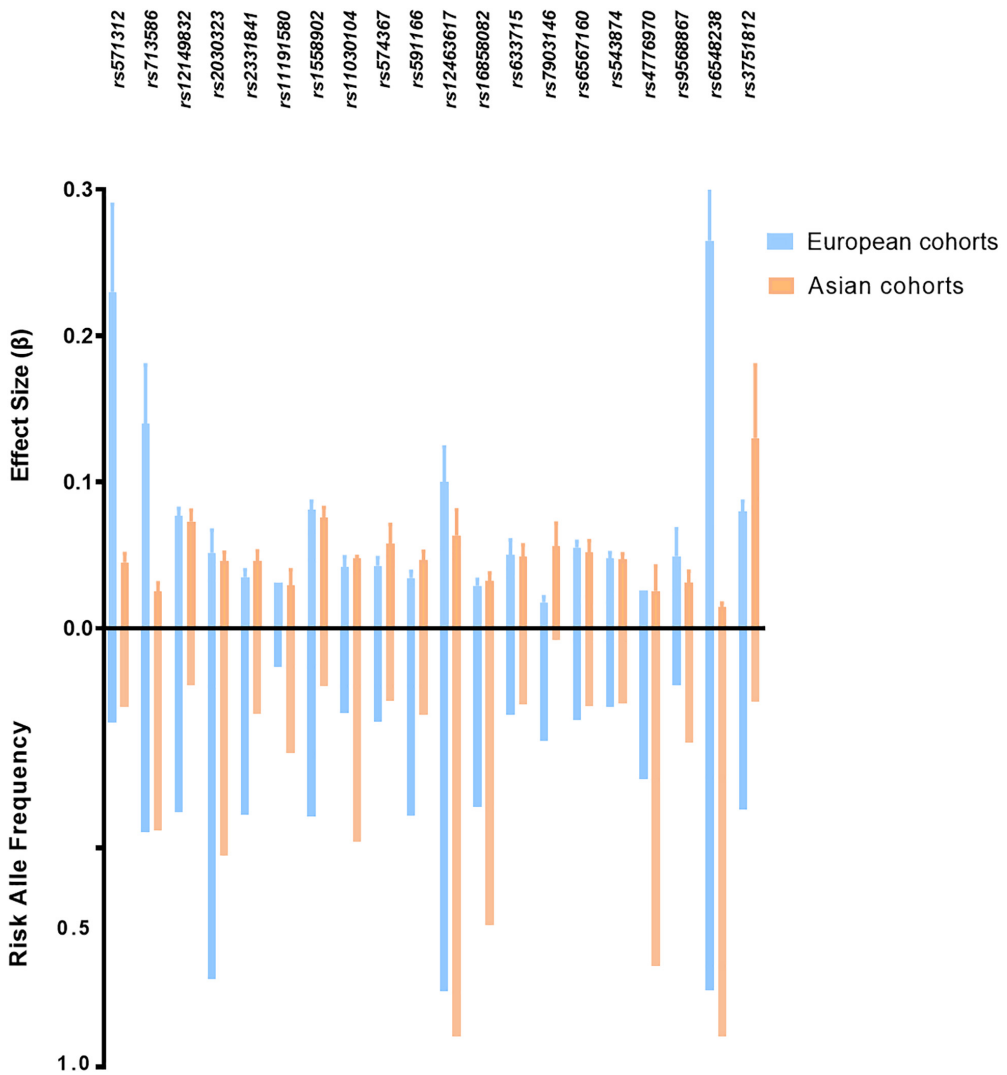

FIGURE 3 | Risk allele frequency and effect size of top ranked obesity susceptibility loci which reached genome association significant *p*-value in European and Asian cohorts. BMI-related loci shown in this figure with $p < 5 \times 10^{-8}$ in both cohorts. *p*-values and effect size are according to reference studies reported (**Supplementary Table S1**). The allele frequency is based on genome Aggregation Database (gnomAD).

of obesity and fat distribution. In this regard, further intensive research is inevitable in the cohorts of Asian ancestry to identify novel obesity loci either specific in Asian ancestry or common for various ancestries and thus provide new insights into the mechanisms underlying obesity.

Understanding the Functional Consequences of Obesity-Associated Variants

The function of most of the genes within obesity-associated loci remains to be clarified. Although numerous polymorphisms associated with obesity have been revealed so far in studies including various ethnicities, identification of the respective target genes of these variants remains challenging. This is mostly attributed to the variety of regulatory mechanisms SNP may be involved in, which makes it difficult to predict the most likely target gene. While in most cases, these genes map in close vicinity of their functional variant, they may also be positioned hundreds of kilobases upstream or downstream of the genes.

In line with this, it has to be noted that most of the genes reported in this review are based on the "closest" gene approach, which admittedly is not a highly accurate approach. Although it may be true for some obesity loci (e.g., FTO), for most of the currently known obesity susceptibility loci, no target genes of the associated genetic variants have been robustly validated. Instead, the closest or nearby genes are being reported and proposed as potential candidate genes explaining the observed associations.

Measures of Obesity

The classical and mostly applied measure of obesity is BMI. However, because of differences in phenotypes and body

composition in Asian and European populations, BMI may not be the most appropriate measure to assess the degree of obesity globally. This phenomenon may cause GWAS to miss important genetic variants in specific populations or subgroups. At the same time, inaccuracy in the measured phenotypes may result in false-positive association signals. Establishing new tools/measures including whole-body MRI scan and body composition techniques to easily and quickly assess obesity will be inevitable to refine and make the search for obesity-related genes more efficient.

Fine Mapping in Multi-Ethnic/Trans-Ethnic Studies

A growing number of multi-ethnic/trans-ethnic studies have been completed in populations of non-European ancestry in addition to replication studies in recent years. The potential ability to use trans-ethnic studies is identifying common genetic variants shared across different ancestries, as well as ancestry-specific disease predisposing variants, and interactions between genetic variants and the environment that can be shared or ancestry specific as well. Moreover, the diversity of LD patterns across various genetic ancestries can be leveraged to indicate causal variants. Moving beyond GWAS, also other approaches such as fine mapping studies are a valuable attempt to apply to multi-ethnic cohorts to get a better understanding of the role of novel loci implicated in obesity.

Fine-mapping strategies typically follow the GWAS findings aiming at prioritization of variants within susceptibility regions in the genome. Although the original GWAS can suggest a region that is likely to include a causal variant, additional strategies (fine mapping, whole-exome, and whole-genome sequencing) are necessary to distinguish most likely functional variants from only correlated causal variants. A major challenge in identifying underlying causal SNPs are the presence of LD, which can lead to highly correlated association results and multiple significant SNPs at a locus of interest. Most of the GWAS performance so far assume association analyses in relatively homogenous populations with consistent patterns of LD; this is straightforward for discovering associated variants. However, it can be challenging in multi-ethnic studies, where distinguishing multiple nearly equivalent variants may need hundreds of thousands of individual samples. Fine mapping in different ancestries is a method of lessening the barrier of LD and aids this process by selecting and prioritizing variants most likely responsible for complex traits. In addition, trans-ethnic fine mapping is a powerful approach for both narrowing the underlying causal variants in known loci as well as in discovering novel variants for complex traits (Zaitlen et al., 2010). Fine mapping in populations with relatively limited LD patterns like in individuals of African (Guo et al., 2013) or Asian (Hotta et al., 2008) ancestry may be helpful in the dissection of genetic architecture within a population and in pinpointing the causal variant. In the future, more trans-ethnic fine mapping studies will be inevitable in dissecting the genetic architecture of complex traits such as obesity. Considering that many complex traits are driven by large numbers of variants of small effects, which likely interact with the environment in complex ways, detailed mapping of genetic architecture regulatory networks and G × E effects will be an essential task for fully understanding human disease biology (Boyle et al., 2017).

CONCLUSION

In summary, GWAS has exhibited a large number of BMI-associated loci over the past decade, providing an effective way to understand better obesity mechanisms which are essential on our way to improve the treatment of obesity. Although the pioneering large-scale GWAS were mostly conducted on individuals of European ancestry, there has been remarkable progress, which is now closing the gap between our knowledge of obesity genetics in European versus Asian ancestries. It should be noted that GWAS executed in Asian cohorts have not only affirmed the potential role of previously associated obesity loci but also displayed novel ones, which have been missed in the initial genetic studies in individuals of European ancestries. In addition, follow-up GWAS research strategies in multi-ethnic/trans-ethnic studies are worthwhile to conduct. At last, despite a large number of currently known obesity risk loci, the molecular mechanisms underlying this complex disease are not fully explained yet, and neither is the variation across human diversity in terms of obesity.

AUTHOR CONTRIBUTIONS

CS wrote the original draft of the manuscript. CS, EG-J, and PK reviewed and edited the manuscript. EG-J and PK supervised the study. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC 1052, project number 209933838, subprojects B3 to PK, by Deutsches Zentrum für Diabetesforschung (DZD) to EG-J, and by China Scholarship Council to CS, no. 201706170052.

ACKNOWLEDGMENTS

The authors acknowledge support from the German Research Foundation (DFG) and Universität Leipzig within the program of Open Access Publishing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.575049/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

Batterham, R. L., Le Roux, C. W., Cohen, M. A., Park, A. J., Ellis, S. M., Patterson, M., et al. (2003). Pancreatic Polypeptide Reduces Appetite And Food Intake In Humans. *J. Clin. Endocrinol. Metab.* 88, 3989–3992. doi: 10.1210/Jc.2003-030630

Berndt, S. I., Gustafsson, S., Mägi, R., Ganna, A., Wheeler, E., Feitosa, M. F., et al. (2013). Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat. Genet.* 45, 501–512. doi: 10.1038/ng.2606

Bhattacharyya, D., And Glick, B. S. (2007). Two Mammalian Sec16 Homologues Have Nonredundant Functions In Endoplasmic Reticulum (Er) Export And Transitional Er Organization. *Mol. Biol. Cell* 18, 839–849. doi: 10.1091/Mbc.E06-08-0707

Bollepalli, S., Dolan, L. M., Deka, R., and Martin, L. J. (2010). Association of FTO gene variants with adiposity in African-American adolescents. *Obesity* 18, 1959–1963. doi: 10.1038/oby.2010.82

Bouchard, C., Tremblay, A., Després, J. P., Nadeau, A., Lupien, P. J., Thériault, G., et al. (1990). The response to long-term overfeeding in identical twins. *N. Engl. J. Med.* 322, 1477–1482. doi: 10.1056/Nejm199005243222101

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell* 169, 1177–1186. doi: 10.1016/J.Cell.2017.05.038

Cai, T., Yu, P., Monga, S. P., Mishra, B., and Mishra, L. (1998). Identification of mouse Itih-4 encoding a glycoprotein with two Ef-hand motifs from early embryonic liver. *Biochim. Biophys. Acta* 1398, 32–37. doi: 10.1016/S0167-4781(98)00049-9

Chambers, J. C., Elliott, P., Zabaneh, D., Zhang, W., Li, Y., Froguel, P., et al. (2008). Common Genetic Variation Near Mc4r Is Associated With Waist Circumference And Insulin Resistance. *Nat. Genet.* 40, 716–718. doi: 10.1038/Ng.156

Chen, G., Doumatey, A. P., Zhou, J., Lei, L., Bentley, A. R., Tekola-Ayele, F., et al. (2017). Genome-Wide Analysis Identifies An African-Specific Variant In Sema4d Associated With Body Mass Index. *Obesity* 25, 794–800. doi: 10.1002/Oby.21804

Chen, K. Y., Muniyappa, R., Abel, B. S., Mullins, K. P., Staker, P., Brychta, R. J., et al. (2015). Rm-493, a melanocortin-4 receptor (Mc4r) agonist, increases resting energy expenditure in obese individuals. *J. Clin. Endocrinol. Metab.* 100, 1639–1645. doi: 10.1210/Jc.2014-4024

Chen, L., Hou, J., Ye, L., Chen, Y., Cui, J., Tian, W., et al. (2014). Microrna-143 Regulates Adipogenesis By Modulating The Map2k5-Erk5 Signaling. *Sci. Rep.* 4:3819. doi: 10.1038/Srep03819

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

Claussnitzer, M., Dankel, S. N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., et al. (2015). Fto obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* 373, 895–907. doi: 10.1056/Nejmoa1502214

Commodore-Mensah, Y., Selvin, E., Aboagye, J., Turkson-Ocran, R.-A., Li, X., Himmelfarb, C. D., et al. (2018). Hypertension, overweight/obesity, and diabetes among immigrants in the United States: an analysis of the 2010–2016 national health interview survey. *BMC Publ. Health* 18:773. doi: 10.1186/S12889-018-5683-3

Damgaard, P. D. B., Marchi, N., Rasmussen, S., Peyrot, M., Renaud, G., Korneliussen, T., et al. (2018). 137 ancient human genomes from across the eurasian steppes. *Nature* 557, 369–374. doi: 10.1038/S41586-018-0094-2

Dina, C., Meyre, D., Gallina, S., Durand, E., Körner, A., Jacobson, P., et al. (2007). Variation In Fto Contributes To Childhood Obesity And Severe Adult Obesity. *Nat. Genet.* 39, 724–726. doi: 10.1038/Ng2048

Division of Cancer Epidemiology & Genetics (2020). *LDlink*. Available online at: https://ldlink.nci.nih.gov/?tab=home

Dong, C., Beecham, A., Slifer, S., Wang, L., Mcclendon, M. S., Blanton, S. H., et al. (2011). Genome-wide linkage and peak-wide association study of obesity-related quantitative traits in caribbean hispanics. *Hum. Genet.* 129, 209–219. doi: 10.1007/S00439-010-0916-2

Dorajoo, R., Blakemore, A. I. F., Sim, X., Ong, R. T.-H., Ng, D. P. K., Seielstad, M., et al. (2012). Replication Of 13 obesity loci among Singaporean Chinese, Malay and Asian-Indian populations. *Int. J. Obes.* 36, 159–163. doi: 10.1038/Ijo.2011.86

Duan, C., Yang, H., White, M. F., and Rui, L. (2004). Disruption of the Sh2-B gene causes age-dependent insulin resistance and glucose intolerance. *Mol. Cell Biol.* 24, 7435–7443. doi: 10.1128/Mcb.24.17.7435-7443.2004

Emilsson, V., Ilkov, M., Lamb, J. R., Finkel, N., Gudmundsson, E. F., Pitts, R., et al. (2018). Co-regulatory networks of human serum proteins link genetics to disease. *Science* 361, 769–773. doi: 10.1126/Science.Aaq1327

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 570 k individuals across multiple ancestries. *PLoS One* 13:E0198166. doi: 10.1371/Journal.Pone.0198166

Fernández-Rhodes, L., Gong, J., Haessler, J., Franceschini, N., Graff, M., Nishimura, K. K., et al. (2017). Trans-ethnic fine-mapping of genetic loci for body mass index in the diverse ancestral populations of the population architecture using genomics and epidemiology (PAGE) study reveals evidence for multiple signals at established loci. *Hum. Genet.* 136, 771–800. doi: 10.1007/s00439-017-1787-6

Fesinmeyer, M. D., North, K. E., Ritchie, M. D., Lim, U., Franceschini, N., Wilkens, L. R., et al. (2013). Genetic risk factors for bmi and obesity in an ethnically diverse population: results from the population architecture using genomics and epidemiology (Page) study. *Obesity* 21, 835–846. doi: 10.1002/Oby.20268

Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., et al. (2007). A common variant in the Fto gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 316, 889–894. doi: 10.1126/Science.1141634

Fujita, Y., Ezura, Y., Emi, M., Sato, K., Takada, D., Iino, Y., et al. (2004). Hypercholesterolemia associated with splice-junction variation of inter-alpha-trypsin inhibitor heavy chain 4 (Itih4) gene. *J. Hum. Genet.* 49, 24–28. doi: 10.1007/S10038-003-0101-8

Gerken, T., Girard, C. A., Tung, Y.-C. L., Webby, C. J., Saudek, V., Hewitson, K. S., et al. (2007). The obesity-associated Fto gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase. *Science* 318, 1469–1472. doi: 10.1126/Science.1151710

Goes, F. S., Mcgrath, J., Avramopoulos, D., Wolyniec, P., Pirooznia, M., Ruczinski, I., et al. (2015). Genome-wide association study of schizophrenia in ashkenazi jews. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 168, 649–659. doi: 10.1002/Ajmg.B.32349

Gong, J., Schumacher, F., Lim, U., Hindorff, L. A., Haessler, J., Buyske, S., et al. (2013). Fine mapping and identification of BMI loci in African Americans. *Am. J. Hum. Genet.* 93, 661–671. doi: 10.1016/j.ajhg.2013.08.012

Graff, M., Scott, R. A., Justice, A. E., Young, K. L., Feitosa, M. F., Barata, L., et al. (2017). Genome-wide physical activity interactions in adiposity – a meta-analysis of 200,452 adults. *PLoS Genet.* 13:E1006528. doi: 10.1371/Journal.Pgen.1006528

Grant, SFA, Bradfield, JP, Zhang, H. (2009). Investigation of the locus near Mc4r with childhood obesity in Americans Of European And African ancestry. *Obesity* 17:1461–1465.

Grant, SFA, Li, M, Bradfield, JP. (2008). Association analysis of the Fto gene with obesity in children of Caucasian and African ancestry reveals a common tagging Snp. *PLoS One* 2008:E1746.

Guo, Y., Lanktree, M. B., Taylor, K. C., Hakonarson, H., Lange, L. A., and Keating, B. J. (2013). Gene-Centric Meta-Analyses Of 108 912 Individuals Confirm Known Body Mass Index Loci And Reveal Three Novel Signals. *Hum. Mol. Genet.* 22, 184–201. doi: 10.1093/Hmg/Dds396

Heffer, A., Marquart, G. D., Aquilina-Beck, A., Saleem, N., Burgess, H. A., And Dawid, I. B. (2017). Generation And Characterization Of Kctd15 Mutations In Zebrafish. *PLoS One* 12:E0189162. doi: 10.1371/Journal.Pone.0189162

Hester, J. M., Wing, M. R., Li, J., Palmer, N. D., Xu, J., Hicks, P. J., et al. (2012). Implication of European-derived adiposity loci in African Americans. *Int. J. Obes.* 36, 465–473. doi: 10.1038/ijo.2011.131

Hoffmann, T. J., Choquet, H., Yin, J., Banda, Y., Kvale, M. N., Glymour, M., et al. (2018). A Large Multiethnic Genome-Wide Association Study Of Adult Body

Mass Index Identifies Novel Loci. *Genetics* 210, 499–515. doi: 10.1534/Genetics.118.301479

Hong, J., Shi, J., Qi, L., Cui, B., Gu, W., Zhang, Y., et al. (2013). Genetic Susceptibility, Birth Weight And Obesity Risk In Young Chinese. *Int. J. Obes.* 37, 673–677. doi: 10.1038/Ijo.2012.87

Hotta, K., Nakamura, M., Nakamura, T., Matsuo, T., Nakata, Y., Kamohara, S., et al. (2009). Association Between Obesity And Polymorphisms In Sec16b, Tmem18, Gnpda2, Bdnf, Faim2 And Mc4r In A Japanese Population. *J. Hum. Genet.* 54, 727–731. doi: 10.1038/Jhg.2009.106

Hotta, K., Nakata, Y., Matsuo, T., Kamohara, S., Kotani, K., Komatsu, R., et al. (2008). Variations In The Fto Gene Are Associated With Severe Obesity In The Japanese. *J. Hum. Genet.* 53, 546–553. doi: 10.1007/S10038-008-0283-1

Hunnicut, J., Liu, Y., Richardson, A., And Salmon, A. B. (2015). MSRA Overexpression Targeted To The Mitochondria, But Not Cytosol, Preserves Insulin Sensitivity In Diet-Induced Obese Mice. *PLos One* 10:E0139844. doi: 10.1371/Journal.Pone.0139844

Jacquemont, S., Reymond, A., Zufferey, F., Harewood, L., Walters, R. G., Kutalik, Z., et al. (2011). Mirror Extreme Bmi Phenotypes Associated With Gene Dosage At The Chromosome 16p11.2 *Locus*. *Nature* 478, 97–102. doi: 10.1038/Nature10406

Johnson, L., Luke, A., Adeyemo, A., Deng, H. -W., Mitchell, B. D., Comuzzie, A. G., et al. (2005). Meta-Analysis Of Five Genome-Wide Linkage Studies For Body Mass Index Reveals Significant Evidence For Linkage To Chromosome 8p. *Int. J. Obes.* 29, 413–419. doi: 10.1038/Sj.Ijo.0802817

Joo, E., Harada, N., Yamane, S., Fukushima, T., Taura, D., Iwasaki, K., et al. (2017). Inhibition Of Gastric Inhibitory Polypeptide Receptor Signaling In Adipose Tissue Reduces Insulin Resistance And Hepatic Steatosis In High-Fat Diet-Fed Mice. *Diabetes* 66, 868–879. doi: 10.2337/Db16-0758

Jorgenson, E., Thai, K. K., Hoffmann, T. J., Sakoda, L. C., Kvale, M. N., Banda, Y., et al. (2017). Genetic Contributors To Variation In Alcohol Consumption Vary By Race/Ethnicity In A Large Multi-Ethnic Genome-Wide Association Study. *Mol. Psych.* 22, 1359–1367. doi: 10.1038/Mp.2017.101

Kang, S. J., Chiang, C. W. K., Palmer, C. D., Tayo, B. O., Lettre, G., Butler, J. L., et al. (2010). Genome-Wide Association Of Anthropometric Traits In African- And African-Derived Populations. *Hum. Mol. Genet.* 19, 2725–2738. doi: 10.1093/Hmg/Ddq154

Khera, A. V., Chaffin, M., Wade, K. H., Zahid, S., Brancale, J., Xia, R., et al. (2019). Polygenic Prediction Of Weight And Obesity Trajectories From Birth To Adulthood. *Cell* 177, 587–596.E9. doi: 10.1016/J.Cell.2019.03.028

Kim, H.-J., Yoo, Y. J., Ju, Y. S., Lee, S., Cho, S.-I., Sung, J., et al. (2013). Combined Linkage And Association Analyses Identify A Novel Locus For Obesity Near Prox1 In Asians. *Obesity* 21, 2405–2412. doi: 10.1002/Oby.20153

Kichaev, G., Bhatia, G., Loh, P. -R., Gazal, S., Burch, K., Freund, M. K., et al. (2019). Leveraging polygenic functional enrichment to improve GWAS power. *Am. J. Hum. Genet.* 104, 65–75. doi: 10.1016/j.ajhg.2018.11.008

Krishna, R., Gumbiner, B., Stevens, C., Musser, B., Mallick, M., Suryawanshi, S., et al. (2009). Potent And Selective Agonism Of The Melanocortin Receptor 4 With Mk-0493 Does Not Induce Weight Loss In Obese Human Subjects: energy Intake Predicts Lack Of Weight Loss Efficacy. *Clin. Pharmacol. Ther.* 86, 659–666. doi: 10.1038/Clpt.2009.167

Larder, R., Sim, M. F. M., Gulati, P., Antrobus, R., Tung, Y. C. L., Rimmington, D., et al. (2017). Obesity-Associated Gene Tmem18 Has A Role In The Central Control Of Appetite And Body Weight Regulation. *Proc. Natl. Acad. Sci. U S A.* 114, 9421–9426. doi: 10.1073/Pnas.1707310114

León-Mimila, P., Villamil-Ramírez, H., Villalobos-Comparán, M., Villarreal-Molina, T., Romero-Hidalgo, S., López-Contreras, B., et al. (2013). Contribution Of Common Genetic Variants To Obesity And Obesity-Related Traits In Mexican Children And Adults. *PLoS One* 8:E70640. doi: 10.1371/Journal.Pone.0070640

Li, H. -L., Zhang, Y. -J., Chen, X. -P., Luo, J. -Q., Liu, S. -Y., and Zhang, Z. -L. (2016). Association Between Gnb3 C.*825c* T Polymorphism And The Risk Of Overweight And Obesity: a Meta-Analysis. *Meta Gene* 9, 18–25. doi: 10.1016/J.Mgene.2016.03.002

Liu, X., Weidle, K., Schröck, K., Tönjes, A., Schleinitz, D., Breitfeld, J., et al. (2015). Signatures Of Natural Selection At The Fto (Fat Mass And Obesity Associated) Locus In Human Populations. *PLoS One* 10:E0117093. doi: 10.1371/Journal.Pone.0117093

Liu, Z., Xiang, Y., and Sun, G. (2013). The Kctd Family Of Proteins: structure, Function, Disease Relevance. *Cell Biosci.* 3:45. doi: 10.1186/2045-3701-3-45

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

Loos, R. J. F., Lindgren, C. M., Li, S., Wheeler, E., Zhao, J. H., Prokopenko, I., et al. (2008). Common Variants Near Mc4r Are Associated With Fat Mass, Weight And Risk Of Obesity. *Nat. Genet.* 40, 768–775. doi: 10.1038/Ng.140

Manolopoulos, K. N., Karpe, F., and Frayn, K. N. (2010). Gluteofemoral Body Fat As A Determinant Of Metabolic Health. *Int. J. Obes.* 34, 949–959. doi: 10.1038/Ijo.2009.286

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., et al. (2015). Genome-Wide Patterns Of Selection In 230 Ancient Eurasians. *Nature* 528, 499–503. doi: 10.1038/Nature16152

Mazzu, Y. Z., Hu, Y., Soni, R. K., Mojica, K. M., Qin, L.-X., Agius, P., et al. (2017). Mir-193b-Regulated Signaling Networks Serve As Tumor Suppressors In Liposarcoma And Promote Adipogenesis In Adipose-Derived Stem Cells. *Cancer Res.* 77, 5728–5740. doi: 10.1158/0008-5472.Can-16-2253

Mccoll, H., Racimo, F., Vinner, L., Demeter, F., Gakuhari, T., Moreno-Mayar, J. V., et al. (2018). The Prehistoric Peopling Of Southeast Asia. *Science* 361, 88–92. doi: 10.1126/Science.Aat3628

Melka, M. G., Bernard, M., Mahboubi, A., Abrahamowicz, M., Paterson, A. D., Syme, C., et al. (2012). Genome-wide scan for loci of adolescent obesity and their relationship with blood pressure. *J. Clin. Endocrinol. Metab.* 97, E145–E150. doi: 10.1210/jc.2011-1801

Meyre, D., Delplanque, J., Chèvre, J.-C., Lecoeur, C., Lobbens, S., Gallina, S., et al. (2009). Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nat. Genet.* 41, 157–159. doi: 10.1038/ng.301

Miyawaki, K., Yamada, Y., Ban, N., Ihara, Y., Tsukiyama, K., Zhou, H., et al. (2002). Inhibition Of Gastric Inhibitory Polypeptide Signaling Prevents Obesity. *Nat. Med.* 8, 738–742. doi: 10.1038/Nm727

Monda, K. L., Chen, G. K., Taylor, K. C., Palmer, C., Edwards, T. L., Lange, L. A., et al. (2013). A Meta-Analysis Identifies New Loci Associated With Body Mass Index In Individuals Of African Ancestry. *Nat. Genet.* 45, 690–696. doi: 10.1038/Ng.2608

Moore, S. C., Gunter, M. J., Daniel, C. R., Reddy, K. S., George, P. S., Yurgalevitch, S., et al. (2012). Common Genetic Variants And Central Adiposity Among Asian-Indians. *Obesity* 20, 1902–1908. doi: 10.1038/Oby.2011.238

Muller, Y. L., Hanson, R. L., Piaggi, P., Chen, P., Wiessner, G., Okani, C., et al. (2019). Assessing The Role Of 98 Established Loci For Bmi In American Indians. *Obesity* 27, 845–854. doi: 10.1002/Oby.22433

Nakano, T., Shinka, T., Sei, M., Sato, Y., Umeno, M., Sakamoto, K., et al. (2006). A/G Heterozygote Of The A-3826g Polymorphism In The Ucp-1 Gene Has Higher Bmi Than A/A And G/G Homozygote In Young Japanese Males. *J. Med. Invest.* 53, 218–222. doi: 10.2152/Jmi.53.218

Nanditha, A., Ma, R. C. W., Ramachandran, A., Snehalatha, C., Chan, J. C. N., Chia, K. S., et al. (2016). Diabetes In Asia And The Pacific: implications For The Global Epidemic. *Diabetes Care* 39, 472–485. doi: 10.2337/Dc15-1536

Narasimhan, V. M., Patterson, N., Moorjani, P., Rohland, N., Bernardos, R., Mallick, S., et al. (2019). The Formation Of Human Populations In South And Central Asia. *Science* 365:aat7487. doi: 10.1126/Science.aat7487

Ng, M. C. Y., Graff, M., Lu, Y., Justice, A. E., Mudgal, P., Liu, C.-T., et al. (2017). Discovery and fine-mapping of adiposity loci using high density imputation of genome-wide association studies in individuals of African ancestry: African ancestry anthropometry genetics consortium. *PLoS Genet.* 13:e1006719. doi: 10.1371/journal.pgen.1006719

Ng, M. C. Y., Hester, J. M., Wing, M. R., Li, J., Xu, J., Hicks, P. J., et al. (2012). Genome-Wide Association Of Bmi In African Americans. *Obesity* 20, 622–627. doi: 10.1038/Oby.2011.154

Ng, M. C. Y., Tam, C. H. T., So, W. Y., Ho, J. S. K., Chan, A. W., Lee, H. M., et al. (2010). Implication Of Genetic Variants Near Negr1, Sec16b, Tmem18, Etv5/Dgkg, Gnpda2, Lin7c/Bdnf, Mtch2, Bcdin3d/Faim2, Sh2b1, Fto, Mc4r, And Kctd15 With Obesity And Type 2 Diabetes In 7705 Chinese. *J. Clin. Endocrinol. Metab.* 95, 2418–2425. doi: 10.1210/Jc.2009-2077

Ng, M., Fleming, T., Robinson, M., Thomson, B., Graetz, N., Margono, C., et al. (2014). Global, Regional, And National Prevalence Of Overweight And Obesity In Children And Adults During 1980–2013: a Systematic Analysis For The Global Burden Of Disease Study 2013. *Lancet* 384, 766–781. doi: 10.1016/S0140-6736(14)60460-8

Nogueiras, R., Wiedmer, P., Perez-Tilve, D., Veyrat-Durebex, C., Keogh, J. M., Sutton, G. M., et al. (2007). The Central Melanocortin System Directly Controls Peripheral Lipid Metabolism. *J. Clin. Invest* 117, 3475–3488. doi: 10.1172/Jci31743

Novarino, G., Fenstermaker, A. G., Zaki, M. S., Hofree, M., Silhavy, J. L., Heiberg, A. D., et al. (2014). Exome Sequencing Links Corticospinal Motor Neuron Disease To Common Neurodegenerative Disorders. *Science* 343, 506–511. doi: 10.1126/Science.1247363

Ogden, C. L., Carroll, M. D., Fryar, C. D., and Flegal, K. M. (2015). Prevalence Of Obesity Among Adults And Youth: united States, 2011-2014. *Nchs. Data Brief.* 219, 1–8.

Ohshiro, Y., Ueda, K., Wakasaki, H., Takasu, N., and Nanjo, K. (2001). Analysis of 825C/T polymorphism of G proteinbeta3 subunit in obese/diabetic Japanese. *Biochem. Biophys. Res. Commun.* 286, 678–680. doi: 10.1006/bbrc.2001.5450

Okada, Y., Kubo, M., Ohmiya, H., Takahashi, A., Kumasaka, N., Hosono, N., et al. (2012). Common Variants At Cdkal1 And Klf9 Are Associated With Body Mass Index In East Asian Populations. *Nat. Genet.* 44, 302–306. doi: 10.1038/Ng.1086

Parikh, M., Hetherington, J., Sheth, S., Seiler, J., Ostrer, H., Gerhard, G., et al. (2013). Frequencies Of Obesity Susceptibility Alleles Among Ethnically And Racially Diverse Bariatric Patient Populations. *Surg. Obes. Relat. Dis.* 9, 436–441. doi: 10.1016/J.Soard.2012.04.004

Pei, Y. -F., Zhang, L., Liu, Y., Li, J., Shen, H., Liu, Y. -Z., et al. (2014). Meta-Analysis Of Genome-Wide Association Data Identifies Novel Susceptibility Loci For Obesity. *Hum. Mol. Genet.* 23, 820–830. doi: 10.1093/Hmg/Ddt464

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., et al. (2007). Diet and The Evolution Of Human Amylase Gene Copy Number Variation. *Nat. Genet.* 39, 1256–1260. doi: 10.1038/Ng2123

Pérusse, L., Després, J. P., Lemieux, S., Rice, T., Rao, D. C., and Bouchard, C. (1996). Familial Aggregation Of Abdominal Visceral Fat Level: results From The Quebec Family Study. *Metab. Clin. Exp.* 45, 378–382. doi: 10.1016/S0026-0495(96)90294-2

Peters, U., North, K. E., Sethupathy, P., Buyske, S., Haessler, J., Jiao, S., et al. (2013). A systematic mapping approach of 16q12.2/FTO and BMI in more than 20,000 African Americans narrows in on the underlying functional variation: results from the population architecture using genomics and epidemiology (PAGE) study. *PLoS Genet.* 9:e1003171. doi: 10.1371/journal.pgen.1003171

Pillay, V., Crowther, N. J., Ramsay, M., Smith, G. D., Norris, S. A., and Lombard, Z. (2015). Exploring genetic markers of adult obesity risk in black adolescent South Africans-the Birth to twenty cohort. *Nutr. Diabetes* 5:e157. doi: 10.1038/nutd.2015.7

Ramachandran, A., Chamukuttan, S., Shetty, S. A., Arun, N., And Susairaj, P. (2012). Obesity In Asia–Is It Different From Rest Of The World. *Diab. Metab. Res. Rev.* 28 (Suppl 2), 47–51. doi: 10.1002/Dmrr.2353

Ren, D., Zhou, Y., Morris, D., Li, M., Li, Z., and Rui, L. (2007). Neuronal Sh2b1 Is Essential For Controlling Energy And Glucose Homeostasis. *J. Clin. Invest.* 117, 397–406. doi: 10.1172/Jci29417

Riveros-Mckay, F., Mistry, V., Bounds, R., Hendricks, A., Keogh, J. M., Thomas, H., et al. (2019). Genetic Architecture Of Human Thinness Compared To Severe Obesity. *PLoS Genet.* 15:E1007603. doi: 10.1371/Journal.Pgen.1007603

Rong, R., Hanson, R. L., Ortiz, D., Wiedrich, C., Kobes, S., Knowler, W. C., et al. (2009). Association Analysis Of Variation In/Near Fto, Cdkal1, Slc30a8, Hhex, Ext2, Igf2bp2, Loc387761, And Cdkn2b With Type 2 Diabetes And Related Quantitative Traits In Pima Indians. *Diabetes* 58, 478–488. doi: 10.2337/Db08-0877

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., et al. (2002). Genetic Structure Of Human Populations. *Science* 298, 2381–2385. doi: 10.1126/Science.1078311

Royalty, J. E., Konradsen, G., Eskerod, O., Wulff, B. S., And Hansen, B. S. (2014). Investigation Of Safety, Tolerability, Pharmacokinetics, And Pharmacodynamics Of Single And Multiple Doses Of A Long-Acting A-Msh Analog In Healthy Overweight And Obese Subjects. *J. Clin. Pharmacol.* 54, 394–404. doi: 10.1002/Jcph.211

Sahibdeen, V., Crowther, N. J., Soodyall, H., Hendry, L. M., Munthali, R. J., Hazelhurst, S., et al. (2018). Genetic Variants In Sec16b Are Associated With Body Composition In Black South Africans. *Nutr. Diabet.* 8:43. doi: 10.1038/S41387-018-0050-0

Salinas, Y. D., Wang, L., And Dewan, A. T. (2016). Multiethnic Genome-Wide Association Study Identifies Ethnic-Specific Associations With Body Mass Index In Hispanics And African Americans. *BMC Genet.* 17:78. doi: 10.1186/S12863-016-0387-0

Saunders, C. L., Chiodini, B. D., Sham, P., Lewis, C. M., Abkevich, V., Adeyemo, A. A., et al. (2007). Meta-analysis of genome-wide linkage studies in Bmi and obesity. *Obesity* 15, 2263–2275. doi: 10.1038/Oby.2007.269

Saxena, R., Hivert, M.-F., Langenberg, C., Tanaka, T., Pankow, J. S., Vollenweider, P., et al. (2010). Genetic Variation In Gipr Influences The Glucose And Insulin Responses To An Oral Glucose Challenge. *Nat. Genet.* 42, 142–148. doi: 10.1038/Ng.521

Schmid, P. M., Heid, I., Buechler, C., Steege, A., Resch, M., Birner, C., et al. (2012). Expression Of Fourteen Novel Obesity-Related Genes In Zucker Diabetic Fatty Rats. *Cardiovasc. Diabetol.* 11:48. doi: 10.1186/1475-2840-11-48

Scott, W. R., Zhang, W., Loh, M., Tan, S.-T., Lehne, B., Afzal, U., et al. (2016). Investigation of genetic variation underlying central obesity amongst South Asians. *PLoS One* 11:E0155478. doi: 10.1371/Journal.Pone.0155478

Sha, B.-Y., Yang, T.-L., Zhao, L.-J., Chen, X.-D., Guo, Y., Chen, Y., et al. (2009). Genome-Wide Association Study Suggested Copy Number Variation May Be Associated With Body Mass Index In The Chinese Population. *J. Hum. Genet.* 54, 199–202. doi: 10.1038/Jhg.2009.10

Shan, T., Liu, W., And Kuang, S. (2013). Fatty Acid Binding Protein 4 Expression Marks A Population Of Adipocyte Progenitors In White And Brown Adipose Tissues. *Faseb. J.* 27, 277–287. doi: 10.1096/Fj.12-211516

Shebanits, K., Andersson-Assarsson, J. C., Larsson, I., Carlsson, L. M. S., Feuk, L., and Larhammar, D. (2018). Copy Number Of Pancreatic Polypeptide Receptor Gene Npy4r Correlates With Body Mass Index And Waist Circumference. *PLoS One* 13:E0194668. doi: 10.1371/Journal.Pone.0194668

Sheng, L., Liu, Y., Jiang, L., Chen, Z., Zhou, Y., Cho, K. W., et al. (2013). Hepatic Sh2b1 And Sh2b2 Regulate Liver Lipid Metabolism And Vldl Secretion In Mice. *PLoS One* 8:E83269. doi: 10.1371/Journal.Pone.0083269

Shungin, D., Winkler, T. W., Croteau-Chonka, D. C., Ferreira, T., Locke, A. E., Mägi, R., et al. (2015). New genetic loci link adipose and insulin biology to body fat distribution. *Nature* 518, 187–196. doi: 10.1038/nature14132

Siffert, W., Forster, P., Jöckel, K. H., Mvere, D. A., Brinkmann, B., Naber, C., et al. (1999). Worldwide Ethnic Distribution Of The G Protein Beta3 Subunit 825 T Allele And Its Association With Obesity In Caucasian, Chinese, And Black African Individuals. *J. Am. Soc. Nephrol.* 10, 1921–1930.

Sohail, M., Maier, R. M., Ganna, A., Bloemendal, A., Martin, A. R., Turchin, M. C., et al. (2019). Polygenic Adaptation On Height Is Overestimated Due To Uncorrected Stratification In Genome-Wide Association Studies. *Elife* 8:39702. doi: 10.7554/Elife.39702

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., et al. (2010). Association Analyses Of 249,796 Individuals Reveal 18 New Loci Associated With Body Mass Index. *Nat. Genet.* 42, 937–948. doi: 10.1038/Ng.686

Spracklen, C. N., Horikoshi, M., Kim, Y. J., Lin, K., Bragg, F., Moon, S., et al. (2020). Identification of type 2 diabetes loci in 433,540 East Asian individuals. *Nature*. doi: 10.1038/s41586-020-2263-3

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., et al. (2007). Relative Impact Of Nucleotide And Copy Number Variation On Gene Expression Phenotypes. *Science* 315, 848–853. doi: 10.1126/Science.1136678

Stunkard, A. J. (1986). A Twin Study Of Human Obesity. *JAMA* 256:51. doi: 10.1001/Jama.1986.03380010055024

Stunkard, A. J., Sørensen, T. I., Hanis, C., Teasdale, T. W., Chakraborty, R., Schull, W. J., et al. (1986). An Adoption Study Of Human Obesity. *N. Engl. J. Med.* 314, 193–198. doi: 10.1056/Nejm198601233140401

Sun, C., Cao, M., Shi, J., Li, L., Miao, L., Hong, J., et al. (2013). Copy Number Variations Of Obesity Relevant Loci Associated With Body Mass Index In Young Chinese. *Gene* 516, 198–203. doi: 10.1016/J.Gene.2012.12.081

Takeuchi, F., Katsuya, T., Kimura, R., Nabika, T., Isomura, M., Ohkubo, T., et al. (2017). The Fine-Scale Genetic Structure And Evolution Of The Japanese Population. *PLoS One* 12:E0185487. doi: 10.1371/Journal.Pone.0185487

Tan, A., Sun, J., Xia, N., Qin, X., Hu, Y., Zhang, S., et al. (2012). A Genome-Wide Association And Gene-Environment Interaction Study For Serum Triglycerides

Levels In A Healthy Chinese Male Population. *Hum. Mol. Genet.* 21, 1658–1664. doi: 10.1093/Hmg/Ddr587

Tenesa, A., Campbell, H., Theodoratou, E., Dunlop, L., Cetnarskyj, R., Farrington, S. M., et al. (2009). Common Genetic Variants At The Mc4r Locus Are Associated With Obesity, But Not With Dietary Energy Intake Or Colorectal Cancer In The Scottish Population. *Int. J Obes* 33, 284–288. doi: 10.1038/Ijo.2008.257

Thorleifsson, G., Walters, G. B., Gudbjartsson, D. F., Steinthorsdottir, V., Sulem, P., Helgadottir, A., et al. (2009). Genome-Wide Association Yields New Sequence Variants At Seven Loci That Associate With Measures Of Obesity. *Nat. Genet.* 41, 18–24. doi: 10.1038/Ng.274

Tuzun, E., Sharp, A. J., Bailey, J. A., Kaul, R., Morrison, V. A., Pertz, L. M., et al. (2005). Fine-Scale Structural Variation Of The Human Genome. *Nat. Genet.* 37, 727–732. doi: 10.1038/Ng1562

Ullrich, S., Su, J., Ranta, F., Wittekindt, O. H., Ris, F., Rösler, M., et al. (2005). Effects Of I(Ks) Channel Inhibitors In Insulin-Secreting Ins-1 Cells. *Pflugers Arch.* 451, 428–436. doi: 10.1007/S00424-005-1479-2

Unoki, H., Takahashi, A., Kawaguchi, T., Hara, K., Horikoshi, M., Andersen, G., et al. (2008). Snps In Kcnq1 Are Associated With Susceptibility To Type 2 Diabetes In East Asian And European Populations. *Nat. Genet.* 40, 1098–1102. doi: 10.1038/Ng.208

Usher, C. L., Handsaker, R. E., Esko, T., Tuke, M. A., Weedon, M. N., Hastie, A. R., et al. (2015). Structural Forms Of The Human Amylase Locus And Their Relationships To Snps, Haplotypes And Obesity. *Nat. Genet.* 47, 921–925. doi: 10.1038/Ng.3340

Van Gaal, L. F., Mertens, I. L., and Block, C. E. De. (2006). Mechanisms Linking Obesity With Cardiovascular Disease. *Nature* 444, 875–880. doi: 10.1038/Nature05487

Van Meir, E. G., Polverini, P. J., Chazin, V. R., Su Huang, H. J., Tribolet, N. De, And Cavenee, W. K. (1994). Release Of An Inhibitor Of Angiogenesis Upon Induction Of Wild Type P53 Expression In Glioblastoma Cells. *Nat. Genet.* 8, 171–176. doi: 10.1038/Ng1094-171

Vasan, S. K., Fall, T., Neville, M. J., Antonisamy, B., Fall, C. H., Geethanjali, F. S., et al. (2012). Associations Of Variants In Fto And Near Mc4r With Obesity Traits In South Asian Indians. *Obesity* 20, 2268–2277. doi: 10.1038/Oby.2012.64

Villalobos-Comparán, M., Teresa Flores-Dorantes, M., Teresa Villarreal-Molina, M., Rodríguez-Cruz, M., García-Ulloa, A. C., Robles, L., et al. (2008). The Fto Gene Is Associated With Adulthood Obesity In The Mexican Population. *Obesity* 16, 2296–2301. doi: 10.1038/Oby.2008.367

Voight, B. F., Scott, L. J., Steinthorsdottir, V., Morris, A. P., Dina, C., Welch, R. P., et al. (2010). Twelve Type 2 Diabetes Susceptibility Loci Identified Through Large-Scale Association Analysis. *Nat. Genet.* 42, 579–589. doi: 10.1038/Ng.609

Wang, J., Mei, H., Chen, W., Jiang, Y., Sun, W., Li, F., et al. (2012). Study Of Eight GWAS-Identified Common Variants For Association With Obesity-Related Indices In Chinese Children At Puberty. *Int. J. Obes.* 36, 542–547. doi: 10.1038/Ijo.2011.218

Wang, T.-N., Huang, M.-C., Chang, W.-T., Ko, A. M.-S., Tsai, E.-M., Liu, C.-S., et al. (2006). G-2548a Polymorphism Of The Leptin Gene Is Correlated With Extreme Obesity In Taiwanese Aborigines. *Obesity* 14, 183–187. doi: 10.1038/Oby.2006.23

Wang, Y., and Beydoun, M. A. (2007). The Obesity Epidemic In The United States–Gender, Age, Socioeconomic, Racial/Ethnic, And Geographic Characteristics: a Systematic Review And Meta-Regression Analysis. *Epidemiol. Rev.* 29, 6–28. doi: 10.1093/Epirev/Mxm007

Wang, H., Zhang, F., Zeng, J., Wu, Y., Kemper, K. E., Xue, A., et al. (2019). Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. *Sci. Adv.* 5:eaaw3538. doi: 10.1126/sciadv.aaw3538

Watson, P., Townley, A. K., Koka, P., Palmer, K. J., And Stephens, D. J. (2006). Sec16 Defines Endoplasmic Reticulum Exit Sites And Is Required For Secretory Cargo Export In Mammalian Cells. *Traffic* 7, 1678–1687. doi: 10.1111/J.1600-0854.2006.00493.X

Wen, W., Cho, Y.-S., Zheng, W., Dorajoo, R., Kato, N., Qi, L., et al. (2012). Meta-Analysis Identifies Common Variants Associated With Body Mass Index In East Asians. *Nat. Genet.* 44, 307–311. doi: 10.1038/Ng.1087

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., Speliotes, E. K., Loos, R. J. F., Li, S., Lindgren, C. M., Heid, I. M., et al. (2009). Six New Loci Associated With Body Mass Index Highlight A Neuronal Influence On Body Weight Regulation. *Nat. Genet.* 41, 25–34. doi: 10.1038/Ng.287

Winkler, T. W., Justice, A. E., Graff, M., Barata, L., Feitosa, M. F., Chu, S., et al. (2015). The Influence Of Age And Sex On Genetic Associations With Adult Body Size And Shape: a Large-Scale Genome-Wide Interaction Study. *PLoS Genet.* 11:E1005378. doi: 10.1371/Journal.Pgen.1005378

Wood, A. R., Tyrrell, J., Beaumont, R., Jones, S. E., Tuke, M. A., Ruth, K. S., et al. (2016). Variants In The Fto And Cdkal1 Loci Have Recessive Effects On Risk Of Obesity And Type 2 Diabetes, Respectively. *Diabetologia* 59, 1214–1221. doi: 10.1007/S00125-016-3908-5

Worzfeld, T., and Offermanns, S. (2014). Semaphorins And Plexins As Therapeutic Targets. *Nat. Rev. Drug. Discov.* 13, 603–621. doi: 10.1038/Nrd4337

Yajnik, C. S., Janipalli, C. S., Bhaskar, S., Kulkarni, S. R., Freathy, R. M., Prakash, S., et al. (2009). Fto Gene Variants Are Strongly Associated With Type 2 Diabetes In South Asian Indians. *Diabetologia* 52, 247–252. doi: 10.1007/S00125-008-1186-6

Yang, T.-L., Guo, Y., Li, S. M., Li, S. K., Tian, Q., Liu, Y.-J., et al. (2013). Ethnic Differentiation Of Copy Number Variation On Chromosome 16p12.3 For Association With Obesity Phenotypes In European And Chinese Populations. *Int. J. Obes.* 37, 188–190. doi: 10.1038/Ijo.2012.31

Yasuda, K., Miyake, K., Horikawa, Y., Hara, K., Osawa, H., Furuta, H., et al. (2008). Variants In Kcnq1 Are Associated With Susceptibility To Type 2 Diabetes Mellitus. *Nat. Genet.* 40, 1092–1097. doi: 10.1038/Ng.207

Yengo, L., Sidorenko, J., Kemper, K. E., Zheng, Z., Wood, A. R., Weedon, M. N., et al. (2018). Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. *Hum. Mol. Genet.* 27, 3641–3649. doi: 10.1093/hmg/ddy271

Yong, R. Y. Y., Mustaffa, S. A. B., Wasan, P. S., Sheng, L., Marshall, C. R., Scherer, S. W., et al. (2016). Complex Copy Number Variation Of Amy1 Does Not Associate With Obesity In Two East Asian Cohorts. *Hum. Mutat.* 37, 669–678. doi: 10.1002/Humu.22996

Yoon, K.-H., Lee, J.-H., Kim, J.-W., Cho, J. H., Choi, Y.-H., Ko, S.-H., et al. (2006). Epidemic Obesity And Type 2 Diabetes In Asia. *Lancet* 368, 1681–1688. doi: 10.1016/S0140-6736(06)69703-1

Young, A. I., Nehzati, S. M., Lee, C., Benonisdottir, S., Cesarini, D., Benjamin, D. J., et al. (2020). *Mendelian Imputation Of Parental Genotypes For Genome-Wide Estimation Of Direct And Indirect Genetic Effects*. bioRxiv[Preprint]

Young, A. I., Wauthier, F., and Donnelly, P. (2016). Multiple Novel Gene-By-Environment Interactions Modify The Effect Of Fto Variants On Body Mass Index. *Nat. Commun.* 7:12724. doi: 10.1038/Ncomms12724

Yu, Y.-H., Liao, P.-R., Guo, C.-J., Chen, C.-H., Mochly-Rosen, D., And Chuang, L.-M. (2016). Pkc-Aldh2 Pathway Plays A Novel Role In Adipocyte Differentiation. *PLoS One* 11:E0161993. doi: 10.1371/Journal.Pone.0161993

Zaitlen, N., Paaniuc, B., Gur, T., Ziv, E., and Halperin, E. (2010). Leveraging Genetic Variability Across Populations For The Identification Of Causal Variants. *Am. J. Hum. Genet.* 86, 23–33. doi: 10.1016/J.Ajhg.2009.11.016

Zhang, D., Li, Z., Wang, H., Yang, M., Liang, L., Fu, J., et al. (2015). Interactions Between Obesity-Related Copy Number Variants And Dietary Behaviors In Childhood Obesity. *Nutrients* 7, 3054–3066. doi: 10.3390/Nu7043054

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: 25 March 2021
doi: 10.3389/fgene.2021.636542

Exploring a Region on Chromosome 8p23.1 Displaying Positive Selection Signals in Brazilian Admixed Populations: Additional Insights Into Predisposition to Obesity and Related Disorders

*Rodrigo Secolin[1], Marina C. Gonsales[1], Cristiane S. Rocha[1], Michel Naslavsky[2], Luiz De Marco[3], Maria A. C. Bicalho[4], Vinicius L. Vazquez[5], Mayana Zatz[2], Wilson A. Silva[6] and Iscia Lopes-Cendes[1]**

[1] Department of Medical Genetics and Genomic Medicine, Brazilian Institute of Neuroscience and Neurotechnology (BRAINN), University of Campinas – UNICAMP, Campinas, Brazil, [2] Departament of Genetics and Evolutive Biology, Human Genome and Stem Cell Research Center, Institute of Bioscience, University of São Paulo (USP), São Paulo, Brazil, [3] Department of Surgery, Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil, [4] Department of Clinical Medicine, Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil, [5] Molecular Oncology Research Center (CPOM) – Barretos Cancer Hospital, Barretos, Brazil, [6] Department of Genetics, Ribeirão Preto Medical School, University of São Paulo at Ribeirão Preto (USP), Ribeirão Preto, Brazil

OPEN ACCESS

Edited by:
Diego Ortega-Del Vecchyo, National Autonomous University of Mexico, Mexico

Reviewed by:
Fernando Villanea, University of Colorado Boulder, United States
Austin Reynolds, Baylor University, United States

***Correspondence:**
Iscia Lopes-Cendes
icendes@unicamp.br

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: 01 December 2020
Accepted: 04 March 2021
Published: 25 March 2021

Citation:
Secolin R, Gonsales MC, Rocha CS, Naslavsky M, De Marco L, Bicalho MAC, Vazquez VL, Zatz M, Silva WA and Lopes-Cendes I (2021) Exploring a Region on Chromosome 8p23.1 Displaying Positive Selection Signals in Brazilian Admixed Populations: Additional Insights Into Predisposition to Obesity and Related Disorders. Front. Genet. 12:636542. doi: 10.3389/fgene.2021.636542

We recently reported a deviation of local ancestry on the chromosome (ch) 8p23.1, which led to positive selection signals in a Brazilian population sample. The deviation suggested that the genetic variability of candidate genes located on ch 8p23.1 may have been evolutionarily advantageous in the early stages of the admixture process. In the present work, we aim to extend the previous work by studying additional Brazilian admixed individuals and examining DNA sequencing data from the ch 8p23.1 candidate region. Thus, we inferred the local ancestry of 125 exomes from individuals born in five towns within the Southeast region of Brazil (São Paulo, Campinas, Barretos, and Ribeirão Preto located in the state of São Paulo and Belo Horizonte, the capital of the state of Minas Gerais), and compared to data from two public Brazilian reference genomic databases, BIPMed and ABraOM, and with information from the 1000 Genomes Project phase 3 and gnomAD databases. Our results revealed that ancestry is similar among individuals born in the five Brazilian towns assessed; however, an increased proportion of sub-Saharan African ancestry was observed in individuals from Belo Horizonte. In addition, individuals from the five towns considered, as well as those from the ABRAOM dataset, had the same overrepresentation of Native-American ancestry on the ch 8p23.1 locus that was previously reported for the BIPMed reference sample. Sequencing analysis of ch 8p23.1 revealed the presence of 442 non-synonymous variants, including frameshift, inframe deletion, start loss, stop gain, stop loss, and splicing site variants, which occurred in 24 genes. Among these genes, 13 were associated with obesity, type II diabetes, lipid levels, and waist circumference (*PRAG1*, *MFHAS1*, *PPP1R3B*, *TNKS*, *MSRA*, *PRSS55*, *RP1L1*, *PINX1*,

MTMR9, *FAM167A*, *BLK*, *GATA4*, and *CTSB*). These results strengthen the hypothesis that a set of variants located on ch 8p23.1 that result from positive selection during early admixture events may influence obesity-related disease predisposition in admixed individuals of the Brazilian population. Furthermore, we present evidence that the exploration of local ancestry deviation in admixed individuals may provide information with the potential to be translated into health care improvement.

Keywords: population genomics, Latin American populations, complex diseases, risk stratification, genomic medicine, precision medicine

INTRODUCTION

Admixture between different continental populations generates mosaic chromosomes comprised of genomic segments with different ancestry, which is defined as local ancestry (Seldin et al., 2011). As a result, admixed populations may present marked differences in local ancestry patterns (Browning et al., 2016; Deng et al., 2016; Martin et al., 2017; Secolin et al., 2019). These differences may impact disease incidence and genetic risk prediction across populations (Myles et al., 2008; Moonesinghe et al., 2012; Martin et al., 2017). Thus, enhancing our knowledge of the effect of local ancestry is crucial for the development of adequate precision health programs in admixed populations (Aronson and Rehm, 2015; Hindorff et al., 2018).

The Brazilian population was formed via an admixture process comprised mostly of European, sub-Saharan African, and Native-American population ancestry. In terms of global ancestry inference, studies have shown a predominance of European ancestry, followed by sub-Saharan African and Native-American (Kehdy et al., 2015; Lima-Costa et al., 2015; Rodrigues de Moura et al., 2015). Furthermore, a recent study about local ancestry inferences reported that the Native-American component predominated on the chromosome (ch) 8p23.1 due to positive selection (Secolin et al., 2019) (3). Ch 8p23.1 has undergone inversion events stratified across continental populations (Salm et al., 2012), which may influence the recombination landscape (Alves et al., 2014).

Interestingly, the ch 8p23.1 region found to be under positive selection in the Brazilian population has been reported to contain genes previously associated with type 2 diabetes and overweight/obesity in admixed Americans (Dunn et al., 2006; Flores et al., 2016). Indeed, studies taking admixture into account have shown that type 2 diabetes, insulin secretion, body mass index, obesity, and adiposity are the main clinical phenotypes associated with metabolic disorders (Dunn et al., 2006; Hayes et al., 2013; Goetz et al., 2014; Flores et al., 2016; Mehta et al., 2017). Thus, we hypothesize that variants in candidate genes located on ch 8p23.1 could have provided an evolutionary advantage in a restrictive diet environment in the early stages of the Brazilian admixture. However, in the present high caloric diet environment, this genetic variability can result in an increased number of obesity-related traits in admixed Brazilian individuals.

Therefore, our objective was to expand our knowledge of the effects of admixture by describing the genetic variability of ch 8p23.1 from admixed Brazilian exomes compared with global populations. To achieve our aim, we first extended our study to additional admixed exomes from other southeastern Brazil towns. Second, we identified and analyzed sequencing variants identified in the candidate region of ch 8p23.1.

MATERIALS AND METHODS

Subjects

We evaluated 257 individuals from BIPMed (Rocha et al., 2020), 609 from ABraOM (Naslavsky et al., 2017), and 88 additional exomes from individuals born in the following towns within southeastern Brazil: Barretos (N = 30); Ribeirão Preto (N = 30), located in the state of São Paulo; and Belo Horizonte (N = 28), the capital of the state of Minas Gerais (**Supplementary Figure 1**). Among individuals included in BIPMed, the birthplace of 193 individuals were included; thus, we were able to extract 21 individuals born in São Paulo city and 37 from Campinas to increase the power of regional comparisons. No information regarding place of birth was obtained from the ABraOM dataset. Permission to use raw, anonymized data from BIPMed and ABraOM public databases and raw, anonymized data associated with the 88 exomes of individuals from Barretos, Ribeirão Preto, and Belo Horizonte was obtained. This study was approved by the University of Campinas's Research Ethics Committee (UNICAMP, Campinas, São Paulo, Brazil). All methods were performed following relevant guidelines and regulations.

Exome Processing

Exome data were stored in variant call format (VCF) files created using the GRCh37 assembly. We used PLINK 1.9 (Purcell et al., 2007) software to convert VCF to PLINK files, variant and individual filtering, and data merging (Anderson et al., 2010). First, we removed ambiguous variants (with G/C or A/T alleles) from VCFs associated with each town, BIPMed, and ABraOM. Next, we merged all Brazilian VCFs, maintaining only biallelic variants, autosomal variants, variants in Hardy–Weinberg equilibrium (Anderson et al., 2010), and removal of missing data (> 10%). These filters were used only to analyze population structure and local ancestry and were removed in the analysis to identify variants in the candidate region at ch 8p23.1.

We evaluated the heterozygosity rate of each individual to search for inbreeding (low heterozygosity rate) or sample contamination (high heterozygosity rate) (Anderson et al., 2010), and individuals with a heterozygosity rate higher or lower than three standard deviations (SDs) from the mean

were removed. We also removed individuals with genomic relatedness matrix estimations higher than 0.125, which is the expected genomic relatedness of third-degree relatives (Anderson et al., 2010). The genomic relatedness matrix estimation used a greedy algorithm implemented using the PLINK 1.9 software to maximize the sample size.

After genotype and individual filtering, a total of 893 exomes and 661,617 variants remained in the Brazilian datasets analyzed. We merged this dataset with the 1000 Genome project data phase 3 (1KGP) (1000 Genomes Project Consortium et al., 2015) and removed SNPs with a minor allele frequency (MAF) < 0.01. As a result, 225,997 variants for local ancestry inference were used. We also removed variants in linkage disequilibrium (LD) from the MAF-filtered dataset (parameters: window size = 50 SNPs; shift step = 5 SNPs; and r2 = 0.5) (Anderson et al., 2010), which left 127,172 SNPs for an investigation of population structure.

Population Structure

To evaluate whether our Brazilian sample (BRS) presents a geographical substructure based on birthplace, we performed the analysis of molecular variance (AMOVA) using the poppr.amova package in R software (Excoffier et al., 1992), which compares the genetic distance among birthplace/town groups based on a set of 10,000 random SNPs across the genome. In addition, we compared the BRS data classified by birthplace to the 1KGP dataset via principal component analysis (PCA) using PLINK v1.9 software to evaluate the presence of population-based outliers in the BRS dataset.

Local Ancestry Inference and Positive Selection Test

We phased SNPs without LD pruning using the SHAPEIT2 v2.r387 software with default parameters (O'Connell et al., 2014). After phasing, we converted the output data from SHAPEIT2 to input files required by RFMix v.1.5.4 software (Maples et al., 2013) using a pipeline previously reported[1] (Martin et al., 2017).

Previous studies showed that using Peruvian individuals from the 1KGP with a high degree of Native-American ancestry as a Native-American reference produced the same result as using Native-American indigenous individuals (Secolin et al., 2019). Therefore, we inferred the local ancestry of 23 Peruvian individuals who possessed a > 0.95 proportion of Native-American ancestry (NAT) (Secolin et al., 2019), 23 random Europeans (EUR), and 23 random sub-Saharan Africans from the 1KGP (AFR). The size sample of ancestry references was selected based on the 23 NAT to avoid biases due to unbalanced reference panel sizes of ancestry references, according to the RFMix v.1.5.4. Manual (Maples et al., 2013). We ran RFMix in PopPhased mode with a minimum window size of 0.2 cM, using one EM iteration and node size 5. The reference panel was maintained after the initial inference step, and forward-backward probabilities were saved. We analyzed the proportion of EUR, AFR, and NAT ancestry in the BRS dataset for each variant across the genome using in-house-developed R scripts (Secolin et al., 2019), and results were plotted using the man package in R software (Turner, 2014). In order to evaluate the presence of ch 8p32.1 inversions, we performed an inversion inference using the invClust package in R software (Cáceres and González, 2015), as performed in our previous work (Secolin et al., 2019). Since we have individuals that overlap the previous paper, we decided to compare the inversion inference between the SNP array data (Secolin et al., 2019) and the exome data from the same individuals to evaluate whether the inversion analysis generated a perfect match.

We tested our exome sample for positive selection by the same approach used previously (Patin et al., 2017). Briefly, this approach combines the results of five neutrality statistics (intrapopulation absolute integrated haplotype scores (| iHS|, | ΔiHH|) (Voight et al., 2006; Sabeti et al., 2007), interpopulation integrated haplotype score (| $\Delta iHH_{derived}$|) (Grossman et al., 2010), interpopulation extended haplotype homozygosity (XP-EHH) (Sabeti et al., 2007), and population branch statistics (PBS) (Yi et al., 2010) based on Hudson's Fst (Bhatia et al., 2013) in a single Fisher combined score (FCS) (Deschamps et al., 2016). The variants with values of FCS higher than 99% of the SNP FCS values across the genome (i.e., the 1% highest FCS values) were defined as outliers. Then, we split the genome into 100-variants blocks. Finally, we estimated the proportion of outliers within each block. If a block presents a proportion of outliers higher than the 99.5th percentile (the highest 0.5%), it was defined as a region under positive selection.

Analysis of Chromosome 8p23.1

We extracted the ch 8p23.1 region (8092025–11859740 bp) (Secolin et al., 2019) from the VCF file of each sample using vcftools (Danecek et al., 2011). Variant consequences from each gene region were annotated using the ANNOVAR software (version 2019Oct24) (Wang et al., 2010) with the following flags: -other info (to include our sample AF); -one transcript; -buildver hg19; -remove; -protocol refGene,gnomad211_exome,ALL.sites.2015_08,EUR.sites.2015_08,AFR.sites.2015_08,AMR.sites.2015_08,EAS.sites.2015_08,SAS.sites.2015_08,dbnsfp35a; -operation g,f,f,f,f,f,f,f,f; and -nastring.

We included the allele frequency (AF) information from African/African-American (AFR/AFA), Latino/admixed American (LAT/AMR), East Asian (EAS), non-Finish European (NFE), and South Asian (SAS) populations from the gnomAD exome dataset (Karczewski et al., 2020); sub-Saharan African (AFR), Europeans (EUR), admixed Americans (AMR), East Asians (EAS), and South Asians (SAS) from 1KGP, which are publicly available in ANNOVAR resource data. In addition, we annotated variants that were not identified via ANNOVAR using the Variant Effect Prediction (VEP) algorithm (McLaren et al., 2016) with the following parameters: a buffer_size 500; –canonical; –distance 5000; –regulatory; –species homo_sapiens; –symbol.

To predict the impact of non-synonymous variants identified on protein function, we analyzed the information provided by the use of 12 algorithms, which included PolyPhen2 (Adzhubei et al., 2013), Sort Intolerant from Tolerant (SIFT) (Sim et al., 2012), MutationTaster (Schwarz et al., 2010), PROVEAN (Choi et al., 2012), Combined Annotation Dependent Depletion (CADD)

[1]https://github.com/armartin/ancestry_pipeline

(Rentzsch et al., 2019), MutPred[2]. Functional Analysis through Hidden Markov Models (FATHMM) (Shihab et al., 2015), PhD-SNPg (Capriotti and Fariselli, 2017), Condel (González-Pérez and López-Bigas, 2011), PANTHER (Mi et al., 2013), Align Grantham Variation/Grantham Difference score (GVGD) (Tavtigian et al., 2006), and SNPs&GO (Calabrese et al., 2009). For Align-GVGD, variants graded higher than C35 were classified as deleterious. For MutPred2, variants with a score higher than 0.5 were considered pathogenic. For all other algorithms, we used default classifications.

Associated trait information for genes located on ch 8p23.1 was accessed from the NHGRI-EBI GWAS Catalog (Buniello et al., 2019) on October 30, 2020, and results were complemented by a search of the PubMed® database.

RESULTS

Population Structure

AMOVA results obtained using 10,000 random SNPs showed that 99.21% of observed variation occurred within groups and 0.79% occurred among groups (total φ-statistics = 0.0079; p = 0.0001), indicating the absence of a population substructure, which is consistent with the lack of clusters observed in the PCA of the BRS sample that was based on birthplace (**Figure 1**). In addition, the global PCA of the 1KGP dataset (**Supplementary Figure 2**) showed that our sample consisted of a mixture of European, sub-Saharan African, and Native-American/East Asian individuals, similar to other admixed American populations. However, the population was distributed mainly between Europeans and sub-Saharans rather than Native-Americans, consistent with previous studies (Ruiz-Linares et al., 2014; Kehdy et al., 2015; Secolin et al., 2019; Rocha et al., 2020).

[2]http://mutpred.mutdb.org

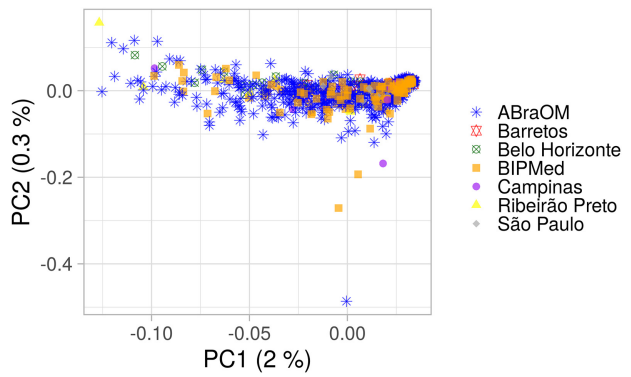

FIGURE 1 | PCA plot of the BRS sample. The *x*-axis and *y*-axis show the first and second principal components (PC1 and PC2) and their respective percentage variability. Each point represents one individual, and each shape and color indicate a birthplace or database. Only information from individuals with a known place of birth was included.

Local Ancestry Inference

The proportion of the BRS sample, which included individuals born in the different towns and two public datasets, had an average local ancestry proportion for its EUR component of 74.6% (*SD* = 1.4%). The proportion of the sample that comprised the AFR component was 16.0% (*SD* = 1.1%), and the NAT component was 9.4% (*SD* = 1.1%) (**Figure 2A** and **Supplementary Figure 3**). We observed differences in EUR and AFR ancestry proportions among towns, with the Belo Horizonte population containing the lowest EUR component (mean = 66.8%; *SD* = 6.0%), and the highest AFR component (mean = 26.9%; *SD* = 5.6%). São Paulo, in contrast, had the greatest proportion of EUR ancestry (mean = 87.8%; *SD* = 5.7%) and the lowest AFR proportion (mean = 6.2%; *SD* = 4.1%). The NAT component of the sample remained constant among individuals from the different towns and the two public Brazilian databases, ranging from a mean of 4.6% (*SD* = 2.8%) in Barretos to 8.2% (*SD* = 1.3%) in the BIPMed sample (**Figure 2A**). Our assessment revealed a decreased EUR component on ch 8p23.1, and an elevated NAT in individuals born in Campinas, São Paulo, Barretos, and Belo Horizonte, as well as in those included in the ABraOM dataset (**Figure 2B** and **Supplementary Figures 3–9**).

We inferred the inversion events on ch 8p23.1 from the exome data by the same approach used in our previous report (Secolin et al., 2019), with the invClust package in R (Cáceres and González, 2015). The results showed that 48.9% of the inferred inversions in the exome data matched the results previously obtained with the SNP-array dataset of the BIPMed sample used in our previous work (Secolin et al., 2019; **Supplementary Table 1**).

We tested for positive selection in the exome dataset using the Fisher combined scores (FCS). FCS, which includes PBS tests in the calculation, this is the same approach used in our previous work (Secolin et al., 2019). However, the results did not recapture the same positive selection signal on ch 8p23.1 previously observed (Secolin et al., 2019) (**Supplementary Figures 3–10**).

Analysis of Chromosome 8p23.1

We found 17,536 variants within ch 8p23.1. We focused on the following variants with the potential to impact gene function: 414 non-synonymous variants, ten frameshifts, eight inframe deletions, one start loss, five stop gains, one stop loss, and five splicing sites. The variants affected 24 genes and two open reading frames (**Supplementary Data Sheet 1**). Among these variants, 355 were also found in gnomAD and/or 1KGP databases, and 44 such variants were determined to be common with an alternative allele frequency (AAF) $>$ 0.01 in the BRS dataset but rare (AAF $<$ 0.01) in gnomAD and 1KGP (**Table 1**). Also, we identified nine common variants (AAF $>$ 0.01) among the 89 variants exclusive to the Brazilian population. The AF comparison of these 89 variants separated by Brazilian cities and datasets showed that the *RP1L1* gene in the ABraOM database contained the largest number of exclusive Brazilian variants (**Supplementary Figure 11**).

We observed that 374 of the 414 non-synonymous variants, in genes located at ch 8p23.1, were classified as deleterious via *in silico* prediction of at least one algorithm

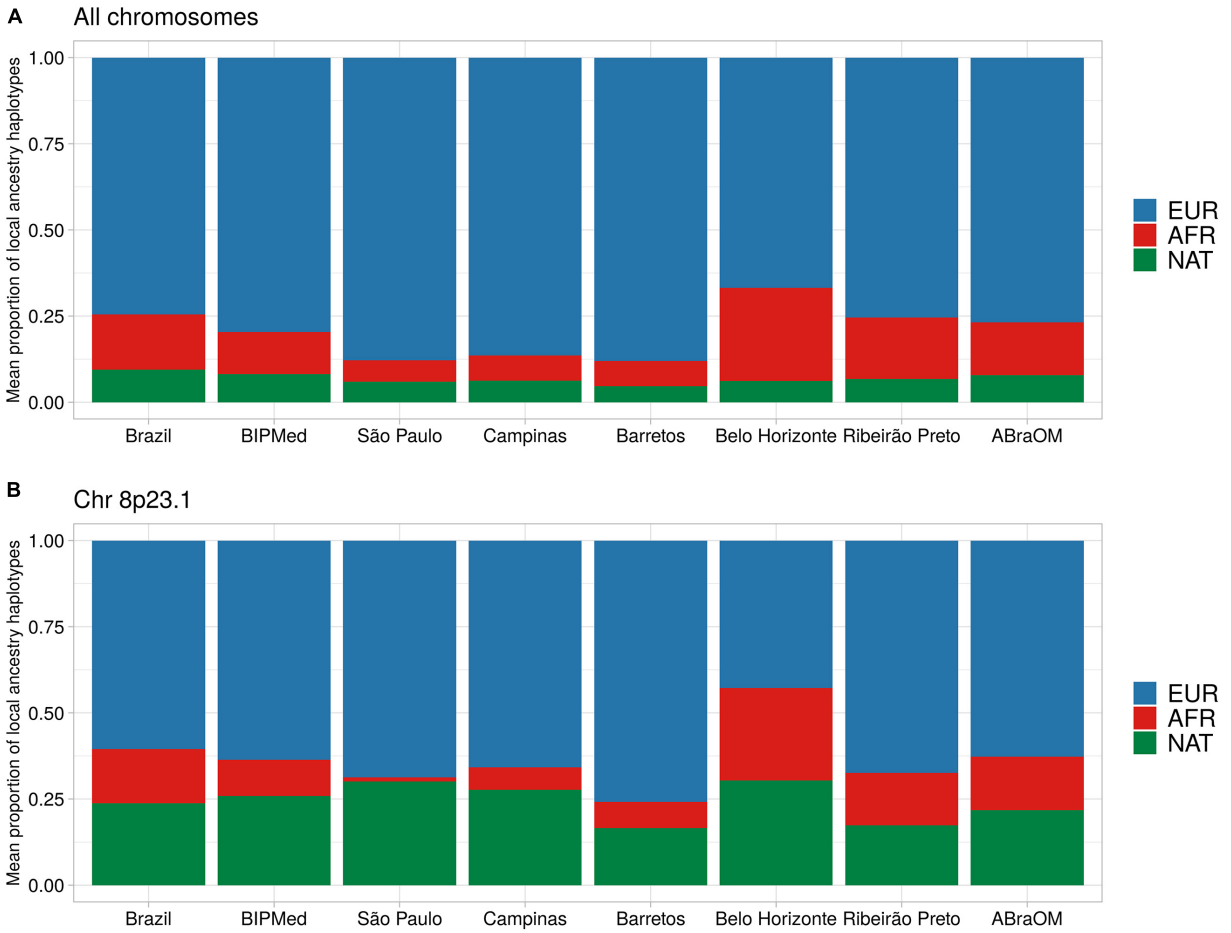

FIGURE 2 | Barplot of the mean proportion of local ancestry haplotypes. Each bar represents one Brazilian sample. **(A)** Mean local ancestry haplotypes across all exomes. **(B)** Mean local ancestry haplotypes on chr8p23.1. EUR, European ancestry component; AFR, African ancestry component; NAT, Native-American ancestry component.

(**Supplementary Data Sheet 2**), and 19 of these were predicted to be pathogenic with an 80% concordance among the different algorithms; these were present in five genes *PRSS55*, *RP1L1*, *SOX7*, *GATA4*, and *CTSB*. As shown in **Figure 3** and **Supplementary Table 2**, we found 167 non-synonymous variants predicted to be benign by at least one algorithm; these were present in 16 different genes. Also, there were 309 variants predicted to be deleterious and 50 variants predicted to be benign when considering less than 20% concordance among the algorithms.

Interestingly, 140 of the variants predicted to be benign by at least one algorithm (140/167) were found in 13 genes, which were previously associated with metabolic phenotypes such as type 1 diabetes mellitus (*T1DM*), type 2 diabetes mellitus (*T2DM*), obesity, insulin resistance, body mass index (BMI), body fat distribution, waist circumference, and diet measurement (*MFHAS1*, *ERI1*, *TNKS*, *PRSS55*, *RP1L1*, *PINX1*, *XKR6*, *FAM167A*, *BLK*, *GATA4*, and *CTSB* genes), **Table 2**. Also, 45 of these variants were located in the following eight genes with an AAF > 0.01, considering the BRSs and gnomAD/1KGP databases: *MFHAS1*, *ERI1*, *PRSS55*, *RP1L1*, *PINX1*, *FAM167A*, *GATA4*, and *CTSB* (**Supplementary Data Sheet 2**).

TABLE 1 | Distribution of genetic variants found in the candidate region of ch 8p23.1, classified according to allele frequencies (AF) observed in the different datasets studied.

AF distribution ($p = 2.2e^{-16}$)*	Common in the Brazilian sample	Rare in the Brazilin sample	Total
Common in gnomAD and/or 1KGP	69 (19.4%)	3 (0.9%)	72 (20.3%)
Rare in gnomAD and/or 1KGP	44 (12.4%)	239 (67.3%)	283 (79.7%)
Total	113 (31.8%)	242 (68.2%)	355 (100%)

**Calculated using Fisher's exact test. 1KGP, 1000 Genome Project phase 3.*

DISCUSSION

The sequencing analysis of ch 8p23.1 performed in the current study revealed the presence of 442 non-synonymous variants,

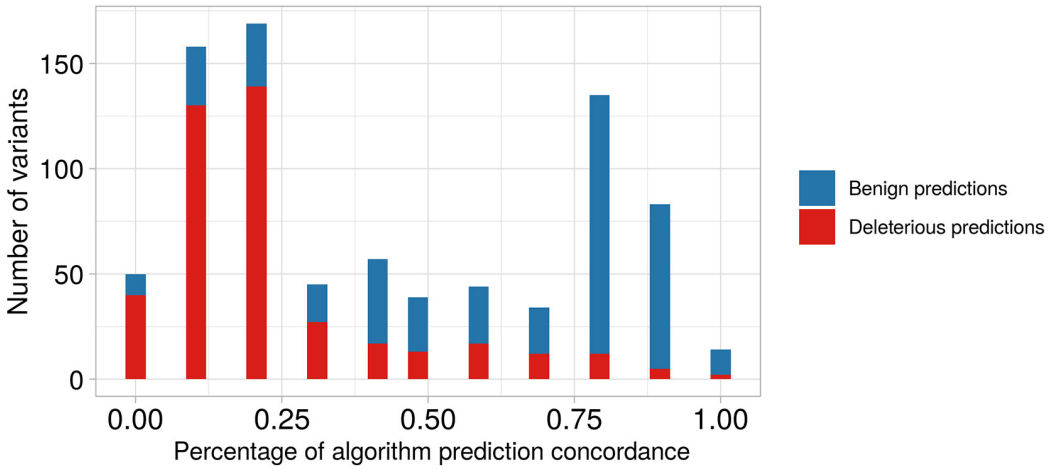

FIGURE 3 | Barplot of predictive algorithm concordance between benign versus deleterious variant predictions. On the *x*-axis, we show the percentage of concordance among the different algorithms. On the *y*-axis, we show the number of predicted variants. For example, the second bar represents the number of variants predicted with low concordance among different algorithms (∼12.5%), and we observe that the number predicted to be deleterious is higher than that predicted to be benign. In contrast, the tenth bar shows predictions with high concordance among algorithms (∼87.5%), and we observe that the number of predicted benign variants is higher than predicted deleterious variants.

including frameshift, inframe deletion, start loss, stop gain, stop loss, and splicing site variants, which occurred in 24 genes and two open reading frames. Among the genes, 13 were associated with obesity, type II diabetes, lipid levels, and waist circumference (*PRAG1*, *MFHAS1*, *PPP1R3B*, *TNKS*, *MSRA*, *PRSS55*, *RP1L1*, *PINX1*, *MTMR9*, *FAM167A*, *BLK*, *GATA4*, and *CTSB*).

The inversion event on ch 8p23.1 generated a large haplotype, which was able to be traced through continental populations globally (Salm et al., 2012). It presented us with an opportunity to investigate how admixture events and evolutionary processes have affected variability within non-inverted and inverted haplotypes in admixed populations. Previously, two independent studies reported local ancestry deviation on ch 8p23.1 in admixed American populations, likely due to inversion events (Guan, 2014; Secolin et al., 2019). Furthermore, our own work using SNP-array data demonstrated that the proportion of non-inverted haplotypes inherited from Native-Americans is higher than those inherited from Europeans in admixed Brazilian individuals (Secolin et al., 2019). Here, we replicated these findings using exome datasets in populations originating from an extended geographic region in the southeastern region of Brazil. Besides, since we evaluated individuals with unknown information regarding the presence of obesity-related disorders, our study is not biased toward a specific phenotype, and it is suitable to assess the genetic variability of the candidate region on ch 8p23.1.

We observed that the results from the inversion inference on ch 8p23.1 obtained in the present work, using the exome data, did not completely match that resulted from the analysis using the SNP-array dataset (Secolin et al., 2019). However, it is noteworthy that the invClust package was developed to be used with SNP-array data, and to our knowledge, there is no reference to its use with exome datasets. Thus, it is possible that inferences of chromosomal inversions using exome data may not be accurate with the invClust package. As pointed out in our previous work (Secolin et al., 2019), it is not likely that inversion bias would influence the high NAT proportions observed in the sample. However, we agree that this is a limitation of our current work. Further analysis, in which inverted and non-inverted genotypes are unequivocally identified, would help evaluate the distribution of the inflation in NAT ancestry in inverted and non-inverted genotypes.

There is evidence that the deviation towards Native-American ancestry on ch 8p23.1 could be due to positive selection events after the Brazilian admixture (Secolin et al., 2019). Indeed, previous studies suggested that Native-American ancestry was admixed early in the European colonization in Brazil (approximately 18 to 16 generations ago) and was followed by the posterior depletion of NAT (Kehdy et al., 2015). This early admixture could have catalyzed positive selection events among the first admixed Brazilian individuals. Although environmental causes that drove this positive selection remain unknown, studies had identified variants associated with type 2 diabetes mellitus, insulin secretion, body mass index, obesity, and adiposity, when admixture was considered (Dunn et al., 2006; Hayes et al., 2013; Goetz et al., 2014; Flores et al., 2016; Mehta et al., 2017). Therefore, the large number of genes located on ch 8p23.1 related to diet and metabolic traits suggest that positive selection may have occurred due to the restrictive diet environment and severe famine periods in early admixed Brazilian individuals (Davis, 2001).

In the present work, our results did not recapture the same positive selection signal detected previously (Secolin et al., 2019) (**Supplementary Figures 3–10**). However, since FCS has only been used with whole-genome sequencing (Deschamps et al., 2016; Patin et al., 2017) and SNP-array datasets

TABLE 2 | Genes associated with obesity-related traits that localize to the candidate region on ch 8p23.1 and are found to contain genetic variants in the Brazilian datasets analyzed in the present work.

Gene	Variant count		Associated trait	Population analyzed	References
	Benign	Deleterious			
CLDN23	3	–	–	–	–
MFHAS1	4	–	T2DM; cooked vegetable consumption; fish- and plant-related diet	European; African American; Hispanic; Asians; East Asian; South Asian;	Niarchou et al., 2020; Vujkovic et al., 2020
ERI1	1	–	Obesity; BMI; body fat distribution	European; Asian; Hispanic; Native-American; Oceanian	Pulit et al., 2019; Rask-Andersen et al., 2019; Schlauch et al., 2020
TNKS	2	–	T2DM; BMI; Early-onset extreme obesity	European; French and German groups	Scherag et al., 2010; Xue et al., 2018; Wang et al., 2019
PRSS55	9	3	Waist circumference	Hispanic obesity children	Comuzzie et al., 2012
RP1L1	107	2	Waist circumference	Waist circumference	Comuzzie et al., 2012
C8orf74	7	–	–	–	–
SOX7	–	2	–	–	–
PINX1	8	–	T2DM; Lipid levels	European, South Asian, East Asian, African	Willer et al., 2013; Xue et al., 2018; Wang et al., 2019
XKR6	1	–	T2DM; BMI; body fat distribution; raw vegetable consumption; processed meat consumption; fish- and plant-related diet	European	Mahajan et al., 2018; Pulit et al., 2019; Rask-Andersen et al., 2019; Niarchou et al., 2020
SLC35G5	13	–	–	–	–
FAM167A	4	–	T2DM	African American; Caribbean	Divers et al., 2017
BLK	1	–	T2DM	European	Borowiec et al., 2009
GATA4	1	2	T1DM; Neonatal and Childhood-Onset diabetes; fruit consumption, processed meat consumption	European	Sartori et al., 2014; Shaw-Smith et al., 2014; Niarchou et al., 2020
NEIL2	2	–	–	–	–
CTSB	2	1	Obesity; visceral obesity in T2DM; non-alcoholic fatty liver disease	Danes; Finnish; European	Peltola et al., 2006; Andreasen et al., 2009; Chalasani et al., 2010
DEFB136	2	-	-	-	-

T1DM, type 1 diabetes mellitus; T2DM, type 2 diabetes mellitus; BMI, body mass index.

(Secolin et al., 2019), we believe that the decrease in genetic variability present in exome data may render FCS less suitable for this type of analysis.

Increasing fat and glucose storage could increase body fat, glucose storage, and obesity-related diseases in individuals who eat a fat and glucose-rich diet today; findings consistent with previous association studies (Pulit et al., 2019; Rask-Andersen et al., 2019; Schlauch et al., 2020). Indeed, we identified 89 variants with the potential to impact gene function that were found exclusively in the admixed Brazilian sample. Unfortunately, we cannot define the correct phase for the allele variants and the ancestry block by the RFMix algorithm. However, we know that the 89 variants are in the region, presenting 60.69% of EUR ancestry proportion, followed by 15.47% of AFR and 23.83% of NAT ancestry proportions. The comparison of these proportions with the average EUR ancestry proportion among Brazilian genomes genomes (74.6%), AFR (16.0%), and NAT (9.4%) suggests that these variants present exclusively in the Brazilian samples could, most likely, be the main contributors to the signals of selection identified, and are possibly influencing obesity-related phenotypes. Therefore, we consider the region of ch 8p23.1 a hotspot for genetic variants that predispose individuals to obesity disorders. It may be useful, as a first strategy, to concentrate efforts on studying effects of non-synonymous variants identified within the 13 candidate genes of the region, *PRAG1*, *MFHAS1*, *PPP1R3B*, *TNKS*, *MSRA*, *PRSS55*, *RP1L1*, *PINX1*, *MTMR9*, *FAM167A*, *BLK*, *GATA4*, and *CTSB*. It may also be useful to expand genetic studies to include patients with obesity-related phenotypes and studying the expression levels of candidate genes in relevant tissue may also give additional clues regarding their roles in disease-related phenotypes.

In **Table 2**, we present the list of 17 genes that have been linked to diet patterns in large association studies and are located in the candidate region on ch 8p23.1. Seven of these large studies included Hispanic, Native-American, and Caribbean populations (Comuzzie et al., 2012; Divers et al., 2017; Pulit et al., 2019; Rask-Andersen et al., 2019; Niarchou et al., 2020; Schlauch et al., 2020; Vujkovic et al., 2020), and three contained association signals in the *ERI1* gene (Pulit et al., 2019; Rask-Andersen et al., 2019; Schlauch et al., 2020), which is located within the region and was determined to possess the greatest degree of positive selection in admixed Brazilians in our previous work (Secolin et al., 2019).

Finally, it is also important to study the non-synonymous variants identified in the candidate genes on ch 8p23 and predicted to be benign. These variants were identified in 11 of the candidate genes listed in **Table 2**. Currently, we cannot exclude the possibility that even though these variants are not predicted to

affect protein function individually, they may contribute to a polygenic phenotype.

Furthermore, when considering a polygenic phenotype, one aspect that we should take into account is the presence of epistatic interactions. Thus, we could argue that an increase in the frequency of the genes of NAT ancestry on ch 8p23.1 could be due to the breakup of negative epistatic interactions among genes on other regions from NAT genomes and the genes on ch 8p23.1, which are currently coupled with AFR and EUR ancestry tracts, and could lead to increased fitness. We count the number of AFR-NAT-AFR, EUR-NAT-EUR, and NAT-NAT-NAT haplotypes, including ch 8p23.1 and adjacent regions (approximately 3.7Mb upstream and downstream ch 8p23.1). However, we did not observe an overrepresentation of AFR-NAT-AFR (n = 32) or EUR-NAT-EUR (n = 98) ancestry haplotypes compared to NAT-NAT-NAT haplotypes (n = 118). Therefore, our results suggest that interactions among EUR or AFR ancestry genes in adjacent regions on ch 8p23.1 with the NAT ancestry core seems not to be enough to boost negative or positive selection in our sample. However, gene interactions can occur among genes on ch 8p23.1 and genes on other regions of the genome, and further studies should be performed to clarify this issue.

CONCLUSION

We successfully replicated previous results that identified local ancestry deviation on ch 8p23.1, which seems to have occurred in populations from the southeastern region of Brazil, including the states of São Paulo and Minas Gerais. Thus, the candidate region on ch 8p23.1 emerges as a hotspot for obesity-related genes in admixed Brazilians, which should be further explored. In particular, the information presented here could be used in the future to support risk stratification and implement personalized public health policies and preventive medical treatments.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories have been included in the **Supplementary Material**. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Comitê de Ética da Universidade Estadual de Campinas. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

RS created the study design, conceptualized the work, and performed data acquisition and analysis. MG performed the *in silico* prediction analysis. CR participated in BIPMed data acquisition. MN and MZ participated in the ABraOM data acquisition. LD and MB participated in the Belo Horizonte data acquisition and determined sample information. VV aided with the Barretos data acquisition and provided sample information. WS aided in the Ribeirão Preto data acquisition and provided sample information. IL-C conceptualized the work and served as the principal investigator. All authors reviewed and approved the final version of the manuscript.

FUNDING

This work was supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, grant number 2013/07559-3). RS was supported by FAPESP (grant number 2019/08526-8). IL-C was supported by CNPq (grant number 311923/2019-4).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.636542/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

Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* 76, 7.20.1–7.20.41. doi: 10.1002/0471142905.hg0720s76

Alves, J. M., Chikhi, L., Amorim, A., and Lopes, A. M. (2014). The 8p23 inversion polymorphism determines local recombination heterogeneity across human populations. *Genome Biol. Evol.* 6, 921–930. doi: 10.1093/gbe/evu064

Anderson, C., Pettersson, F., Clarke, G., Cardon, L., Morris, A., and Zondervan, K. (2010). Data quality control in genetic case-control association studies. *Nat. Protoc.* 5, 1564–1573. doi: 10.1038/nprot.2010.116

Andreasen, C. H., Mogensen, M. S., Borch-Johnsen, K., Sandbaek, A., Lauritzen, T., Almind, K., et al. (2009). Studies of CTNNBL1 and FDFT1 variants and measures of obesity: analyses of quantitative traits and case-control studies in 18,014 Danes. *BMC Med. Genet.* 10:17. doi: 10.1186/1471-2350-10-17

Aronson, S. J., and Rehm, H. L. (2015). Building the foundation for genomics in precision medicine. *Nature* 526, 336–342. doi: 10.1038/nature15816

Bhatia, G., Patterson, N., Sankararaman, S., and Price, A. L. (2013). Estimating and interpreting FST: the impact of rare variants. *Genome Res.* 23, 1514–1521. doi: 10.1101/gr.154831.113

Borowiec, M., Liew, C. W., Thompson, R., Boonyasrisawat, W., Hu, J., Mlynarski, W. M., et al. (2009). Mutations at the BLK locus linked to maturity onset diabetes of the young and beta-cell dysfunction. *Proc. Natl. Acad. Sci. U.S.A.* 106, 14460–14465. doi: 10.1073/pnas.0906474106

Browning, S. R., Grinde, K., Plantinga, A., Gogarten, S. M., Stilp, A. M., Kaplan, R. C., et al. (2016). Local ancestry inference in a large US-Based Hispanic/Latino study: hispanic community health study/study of latinos (HCHS/SOL). *G3 (Bethesda)* 6, 1525–1534. doi: 10.1534/g3.116.028779

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2019). The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47, D1005–D1012. doi: 10.1093/nar/gky1120

Cáceres, A., and González, J. R. (2015). Following the footprints of polymorphic inversions on SNP data: from detection to association tests. *Nucleic Acids Res.* 43:e53. doi: 10.1093/nar/gkv073

Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L., and Casadio, R. (2009). Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.* 30, 1237–1244. doi: 10.1002/humu.21047

Capriotti, E., and Fariselli, P. (2017). PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants. *Nucleic Acids Res.* 45, W247–W252. doi: 10.1093/nar/gkx369

Chalasani, N., Guo, X., Loomba, R., Goodarzi, M. O., Haritunians, T., Kwon, S., et al. (2010). Genome-wide association study identifies variants associated with histologic features of nonalcoholic Fatty liver disease. *Gastroenterology* 139:e1-e6. doi: 10.1053/j.gastro.2010.07.057

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., and Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS One* 7:e46688. doi: 10.1371/journal.pone.0046688

Comuzzie, A. G., Cole, S. A., Laston, S. L., Voruganti, V. S., Haack, K., Gibbs, R. A., et al. (2012). Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. *PLoS One* 7:e51954. doi: 10.1371/journal.pone.0051954

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Davis, M. (2001). *Late Victorian Holocausts: El Niño Famines and the Making of the Third World*. London: Verso.

Deng, L., Ruiz-linares, A., Xu, S., and Wang, S. (2016). Ancestry variation and footprints of natural selection along the genome in Latin American populations. *Sci. Rep.* 6, 1–7. doi: 10.1038/srep21766

Deschamps, M., Laval, G., Fagny, M., Itan, Y., Abel, L., Casanova, J. L., et al. (2016). Genomic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. *Am. J. Hum. Genet.* 98, 5–21. doi: 10.1016/j.ajhg.2015.11.014

Divers, J., Palmer, N. D., Langefeld, C. D., Brown, W. M., Lu, L., Hicks, P. J., et al. (2017). Genome-wide association study of coronary artery calcified atherosclerotic plaque in African Americans with type 2 diabetes. *BMC Genet.* 18:105. doi: 10.1186/s12863-017-0572-9

Dunn, J. S., Mlynarski, W. M., Pezzolesi, M. G., Borowiec, M., Powers, C., Krolewski, A. S., et al. (2006). Examination of PPP1R3B as a candidate gene for the type 2 diabetes and MODY loci on chromosome 8p23. *Ann. Hum. Genet.* 70, 587–593. doi: 10.1111/j.1469-1809.2005.00248.x

Excoffier, L., Smouse, P. E., and Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131, 479–491. doi: 10.5962/bhl.title.86657

Flores, Y. N., Velázquez-Cruz, R., Ramírez, P., Bañuelos, M., Zhang, Z. F., Yee, H. F., et al. (2016). Association between PNPLA3 (rs738409), LYPLAL1 (rs12137855), PPP1R3B (rs4240624), GCKR (rs780094), and elevated transaminase levels in overweight/obese Mexican adults. *Mol. Biol. Rep.* 43, 1359–1369. doi: 10.1007/s11033-016-4058-z

Goetz, L. H., Uribe-Bruce, L., Quarless, D., Libiger, O., and Schork, N. J. (2014). Admixture and clinical phenotypic variation. *Hum. Hered.* 77, 73–86. doi: 10.1159/000362233

González-Pérez, A., and López-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score. *Condel. Am. J. Hum. Genet.* 88, 440–449. doi: 10.1016/j.ajhg.2011.03.004

Grossman, S. R., Shylakhter, I., Karlsson, E. K., Byrne, E. H., Morales, S., Frieden, G., et al. (2010). A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* 327, 883–886. doi: 10.1126/science.1183863

Guan, Y. (2014). Detecting structure of haplotypes and local ancestry. *Genetics* 196, 625–642. doi: 10.1534/genetics.113.160697

Hayes, M. G., Urbanek, M., Hivert, M. F., Armstrong, L. L., Morrison, J., Guo, C., et al. (2013). Identification of HKDC1 and BACE2 as genes influencing glycemic traits during pregnancy through genome-wide association studies. *Diabetes* 62, 3282–3291. doi: 10.2337/db12-1692

Hindorff, L. A., Bonham, V. L., Brody, L. C., Ginoza, M. E. C., 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

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443. doi: 10.1038/s41586-020-2308-7

Kehdy, F. S. G., Gouveia, M. H., Machado, M., Magalhães, W. C. S., Horimoto, A. R., Horta, B. L., et al. (2015). Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. *Proc. Natl. Acad. Sci. U.S.A.* 112, 8696–8701. doi: 10.1073/pnas.1504447112

Lima-Costa, M. F., Rodrigues, L. C., Barreto, M. L., Gouveia, M., Horta, B. L., Mambrini, J., et al. (2015). Genomic ancestry and ethnoracial self-classification based on 5,871 community-dwelling Brazilians (The Epigen Initiative). *Sci. Rep.* 5:9812. doi: 10.1038/srep09812

Mahajan, A., Taliun, D., Thurner, M., Robertson, N. R., Torres, J. M., Rayner, N. W., et al. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* 50, 1505–1513. doi: 10.1038/s41588-018-0241-6

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

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

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., et al. (2016). The ensembl variant effect predictor. *Genome Biol.* 17:122. doi: 10.1186/s13059-016-0974-4

Mehta, M. B., Shewale, S. V., Sequeira, R. N., Millar, J. S., Hand, N. J., and Rader, D. J. (2017). Hepatic protein phosphatase 1 regulatory subunit 3B (Ppp1r3b) promotes hepatic glycogen synthesis and thereby regulates fasting energy homeostasis. *J. Biol. Chem.* 292, 10444–10454. doi: 10.1074/jbc.M116.766329

Mi, H., Muruganujan, A., and Thomas, P. D. (2013). PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* 41, D377–D386. doi: 10.1093/nar/gks1118

Moonesinghe, R., Ioannidis, J. P. A., Flanders, W. D., Yang, Q., Truman, B. I., and Khoury, M. J. (2012). Estimating the contribution of genetic variants to difference in incidence of disease between population groups. *Eur. J. Hum. Genet.* 20, 831–836. doi: 10.1038/ejhg.2012.15

Myles, S., Davison, D., Barrett, J., Stoneking, M., and Timpson, N. (2008). Worldwide population differentiation at disease-associated SNPs. *BMC Med. Genomics* 1:22. doi: 10.1186/1755-8794-1-22

Naslavsky, M. S., Yamamoto, G. L., de Almeida, T. F., Ezquina, S. A. M., Sunaga, D. Y., Pho, N., et al. (2017). Exomic variants of an elderly cohort of Brazilians in the ABraOM database. *Hum. Mutat.* 38, 751–763. doi: 10.1002/humu.23220

Niarchou, M., Byrne, E. M., Trzaskowski, M., Sidorenko, J., Kemper, K. E., McGrath, J. J., et al. (2020). Genome-wide association study of dietary intake in the UK biobank study and its associations with schizophrenia and other traits. *Transl. Psychiatry* 10:51. doi: 10.1038/s41398-020-0688-y

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., et al. (2014). A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* 10:e1004234. doi: 10.1371/journal.pgen.1004234

Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., et al. (2017). Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science* 356, 543–546. doi: 10.1126/science.aal1988

Peltola, P., Pihlajamäki, J., Koutnikova, H., Ruotsalainen, E., Salmenniemi, U., Vauhkonen, I., et al. (2006). Visceral obesity is associated with high levels of serum squalene. *Obesity (Silver Spring)* 14, 1155–1163. doi: 10.1038/oby.2006.132

Pulit, S. L., Stoneman, C., Morris, A. P., Wood, A. R., Glastonbury, C. A., Tyrrell, J., et al. (2019). Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum. Mol. Genet.* 28, 166–174. doi: 10.1093/hmg/ddy327

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., 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

Rask-Andersen, M., Karlsson, T., Ek, W. E., and Johansson, Å (2019). Genome-wide association study of body fat distribution identifies adiposity loci and sex-specific genetic effects. *Nat. Commun.* 10:339. doi: 10.1038/s41467-018-08000-4

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 47, D886–D894. doi: 10.1093/nar/gky1016

Rocha, C. S., Secolin, R., Rodrigues, M. R., Carvalho, B. S., and Lopes-Cendes, I. (2020). The Brazilian Initiative on Precision Medicine (BIPMed): fostering genomic data-sharing of underrepresented populations. *NPJ Genomic Med.* 5:42. doi: 10.1038/s41525-020-00149-6

Rodrigues de Moura, R., Coelho, A. V. C., de Queiroz Balbino, V., Crovella, S., and Brandão, L. A. C. (2015). Meta-analysis of Brazilian genetic admixture and comparison with other Latin America countries. *Am. J. Hum. Biol.* 27, 674–680. doi: 10.1002/ajhb.22714

Ruiz-Linares, A., Adhikari, K., Acuña-Alonzo, V., Quinto-Sanchez, M., Jaramillo, C., Arias, W., et al. (2014). Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS Genet.* 10:e1004572. doi: 10.1371/journal.pgen.1004572

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 913–918. doi: 10.1038/nature06250

Salm, M. P. A., Horswell, S. D., Hutchison, C. E., Speedy, H. E., Yang, X., Liang, L., et al. (2012). The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. *Genome Res.* 22, 1144–1153. doi: 10.1101/gr.126037.111

Sartori, D. J., Wilbur, C. J., Long, S. Y., Rankin, M. M., Li, C., Bradfield, J. P., et al. (2014). GATA factors promote ER integrity and β-cell survival and contribute to type 1 diabetes risk. *Mol. Endocrinol.* 28, 28–39. doi: 10.1210/me.2013-1265

Scherag, A., Dina, C., Hinney, A., Vatin, V., Scherag, S., Vogel, C. I., et al. (2010). Two new Loci for body-weight regulation identified in a joint analysis of genome-wide association studies for early-onset extreme obesity in French and german study groups. *PLoS Genet.* 6:e1000916. doi: 10.1371/journal.pgen.1000916

Schlauch, K. A., Read, R. W., Lombardi, V. C., Elhanan, G., Metcalf, W. J., Slonim, A. D., et al. (2020). A comprehensive genome-wide and phenome-wide examination of BMI and obesity in a Northern Nevadan Cohort. *G3* 10, 645–664. doi: 10.1534/g3.119.400910

Schwarz, J. M., Rödelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* 7, 575–576. doi: 10.1038/nmeth0810-575

Secolin, R., Mas-Sandoval, A., Arauna, L. R., Torres, F. R., de Araujo, T. K., Santos, M. L., et al. (2019). Distribution of local ancestry and evidence of adaptation in admixed populations. *Sci. Rep.* 9:13900. doi: 10.1038/s41598-019-50362-2

Seldin, M. F., Pasaniuc, B., and Price, A. L. (2011). New approaches to disease mapping in admixed populations. *Nat. Rev. Genet.* 12, 523–528. doi: 10.1038/nrg3002

Shaw-Smith, C., De Franco, E., Lango Allen, H., Batlle, M., Flanagan, S. E., Borowiec, M., et al. (2014). GATA4 mutations are a cause of neonatal and childhood-onset diabetes. *Diabetes* 63, 2888–2894. doi: 10.2337/db14-0061

Shihab, H. A., Rogers, M. F., Gough, J., Mort, M., Cooper, D. N., Day, I. N. M., et al. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 31, 1536–1543. doi: 10.1093/bioinformatics/btv009

Sim, N. L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* 40, W452–W457. doi: 10.1093/nar/gks539

Tavtigian, S. V., Deffenbaugh, A. M., Yin, L., Judkins, T., Scholl, T., Samollow, P. B., et al. (2006). Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J. Med. Genet.* 43, 295–305. doi: 10.1136/jmg.2005.033878

Turner, S. (2014). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *BioRxiv* [Preprint] doi: 10.1101/005165

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biol.* 4:e72. doi: 10.1371/journal.pbio.0040072

Vujkovic, M., Keaton, J. M., Lynch, J. A., Miller, D. R., Zhou, J., Tcheandjieu, C., et al. (2020). Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat. Genet.* 52, 680–691. doi: 10.1038/s41588-020-0637-y

Wang, H., Zhang, F., Zeng, J., Wu, Y., Kemper, K. E., Xue, A., et al. (2019). Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. *Sci. Adv.* 5:eaaw3538. doi: 10.1126/sciadv.aaw3538

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38:e164. doi: 10.1093/nar/gkq603

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

Xue, A., Wu, Y., Zhu, Z., Zhang, F., Kemper, K. E., Zheng, Z., et al. (2018). Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat. Commun.* 9:2941. doi: 10.1038/s41467-018-04951-w

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329, 75–78. doi: 10.1126/science.1190371

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: 08 July 2021
doi: 10.3389/fgene.2021.672304

Association Analysis of Candidate Variants in Admixed Brazilian Patients With Genetic Generalized Epilepsies

*Felipe S. Kaibara[1], Tânia K. de Araujo[1], Patricia A. O. R. A. Araujo[1], Marina K. M. Alvim[2,3], Clarissa L. Yasuda[2,3], Fernando Cendes[2,3], Iscia Lopes-Cendes[1] and Rodrigo Secolin[1]**

[1] Department of Translational Medicine, School of Medical Sciences, University of Campinas (UNICAMP), Campinas, Brazil, [2] Brazilian Institute of Neuroscience and Neurotechnology (BRAINN), Campinas, Brazil, [3] Department of Neurology, School of Medical Sciences, University of Campinas (UNICAMP), Campinas, Brazil

OPEN ACCESS

Edited by:
Diego Ortega-Del Vecchyo, National Autonomous University of Mexico, Mexico

Reviewed by:
Arslan A. Zaidi, University of Pennsylvania, United States
Minhui Chen, University of Chicago, United States

***Correspondence:**
Rodrigo Secolin
rsecolin@gmail.com

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: 25 February 2021
Accepted: 11 June 2021
Published: 08 July 2021

Citation:
Kaibara FS, de Araujo TK, Araujo PAORA, Alvim MKM, Yasuda CL, Cendes F, Lopes-Cendes I and Secolin R (2021) Association Analysis of Candidate Variants in Admixed Brazilian Patients With Genetic Generalized Epilepsies. Front. Genet. 12:672304. doi: 10.3389/fgene.2021.672304

Genetic generalized epilepsies (GGEs) include well-established epilepsy syndromes with generalized onset seizures: childhood absence epilepsy, juvenile myoclonic epilepsy (JME), juvenile absence epilepsy (JAE), myoclonic absence epilepsy, epilepsy with eyelid myoclonia (Jeavons syndrome), generalized tonic–clonic seizures, and generalized tonic–clonic seizures alone. Genome-wide association studies (GWASs) and exome sequencing have identified 48 single-nucleotide polymorphisms (SNPs) associated with GGE. However, these studies were mainly based on non-admixed, European, and Asian populations. Thus, it remains unclear whether these results apply to patients of other origins. This study aims to evaluate whether these previous results could be replicated in a cohort of admixed Brazilian patients with GGE. We obtained SNP-array data from 87 patients with GGE, compared with 340 controls from the BIPMed public dataset. We could directly access genotypes of 17 candidate SNPs, available in the SNP array, and the remaining 31 SNPs were imputed using the BEAGLE v5.1 software. We performed an association test by logistic regression analysis, including the first five principal components as covariates. Furthermore, to expand the analysis of the candidate regions, we also interrogated 14,047 SNPs that flank the candidate SNPs (1 Mb). The statistical power was evaluated in terms of odds ratio and minor allele frequency (MAF) by the genpwr package. Differences in SNP frequencies between Brazilian and Europeans, sub-Saharan African, and Native Americans were evaluated by a two-proportion Z-test. We identified nine flanking SNPs, located on eight candidate regions, which presented association signals that passed the Bonferroni correction (rs12726617; rs9428842; rs1915992; rs1464634; rs6459526; rs2510087; rs9551042; rs9888879; and rs8133217; p-values $<3.55e^{-06}$). In addition, the two-proportion Z-test indicates that the lack of association of the remaining candidate SNPs could be due to different genomic backgrounds observed in admixed Brazilians. This is the first time that candidate SNPs for GGE are analyzed in

an admixed Brazilian population, and we could successfully replicate the association signals in eight candidate regions. In addition, our results provide new insights on how we can account for population structure to improve risk stratification estimation in admixed individuals.

Keywords: neurology, genetic generalized epilepsies, population genomics, admixed population, association studies

INTRODUCTION

Genetic generalized epilepsies (GGEs) are a group of epilepsy syndromes in which the main feature is the recurrence of generalized onset seizures with no known or suspected etiology other than possible genetic predisposition (Berg et al., 2010; Scheffer et al., 2017). GGEs are among the most common types of epilepsy, with an estimated prevalence of 190 per 100,000 individuals (Aaberg et al., 2017). They include well-established syndromes: childhood absence epilepsy, juvenile myoclonic epilepsy (JME), juvenile absence epilepsy (JAE), myoclonic absence epilepsy (a rare form of GGE), epilepsy with eyelid myoclonia (Jeavons syndrome), generalized tonic–clonic seizures, and generalized tonic–clonic seizures alone (Berg et al., 2010). These different GGE syndromes share most genetic susceptibility factors, suggesting an important correlation among the clinical subtypes (International League Against Epilepsy Consortium on Complex Epilepsies (ILAE Consortium on Complex Epilepsies)., 2018). The diagnosis of GGE relies mainly on clinical information and electroencephalographic examination (Scheffer et al., 2017).

Previous genome-wide association studies (GWASs) and exome sequencing analyses have identified 48 single-nucleotide polymorphisms (SNPs) putatively associated with susceptibility to the GGEs (EPICURE Consortium et al., 2012a,b; International League Against Epilepsy Consortium on Complex Epilepsies (ILAE Consortium on Complex Epilepsies), 2014; International League Against Epilepsy Consortium on Complex Epilepsies (ILAE Consortium on Complex Epilepsies)., 2018; Zhang et al., 2014; Wang et al., 2019). These SNPs are located in or near several genes encoding ion channels and synaptic vesicles, making them plausible candidates for the susceptibility to epilepsy (International League Against Epilepsy Consortium on Complex Epilepsies (ILAE Consortium on Complex Epilepsies)., 2018). However, most of these studies evaluated non-admixed populations, including five studies based on Europeans, three based on Asian populations, mainly Chinese, and two based on African populations (EPICURE Consortium et al., 2012a,b; International League Against Epilepsy Consortium on Complex Epilepsies (ILAE Consortium on Complex Epilepsies), 2014; International League Against Epilepsy Consortium on Complex Epilepsies (ILAE Consortium on Complex Epilepsies)., 2018; Zhang et al., 2014; Wang et al., 2019). It is well known that admixed American populations are underrepresented in GWASs, decreasing the accuracy of replicating, predicting, and estimating polygenic risks for complex disorders in these populations (Martin et al., 2017, 2019).

Therefore, this work aims to investigate if a genetic association exists between previously reported candidate SNPs and GGEs in a cohort of admixed Brazilians. To accomplish this goal, we first investigated the population structure of Brazilian patients with GGE. Subsequently, we performed an association study using the 48 previously reported candidate SNPs and their flanking regions.

MATERIALS AND METHODS

Subjects

We evaluated a total of 87 patients with GGE who were followed up prospectively in the outpatient epilepsy clinic of the University of Campinas (UNICAMP) hospital. All patients had the diagnosis of GGE according to criteria established by the International League Against Epilepsy (ILAE) (Berg et al., 2010; Fisher et al., 2014). Patients were compared with a group of 340 individuals without any neurological disorder from the BIPMed database (Rocha et al., 2020). Both samples are predominantly from the Southeastern region in Brazil. Among the patients with GGE, we found 63 with JME, 10 with JAE, four generalized tonic–clonic seizures alone, two with Jeavons syndrome, one with myoclonic absence epilepsy, one with epilepsy with generalized tonic–clonic seizures, and six patients in whom a specific GGE syndrome could not be determined. All research participants signed an informed consent form previously approved by our Institutional Research Ethics Committee (IRB # 12112913.3.0000.5404).

Single-Nucleotide Polymorphism Quality Control and Population Structure Analysis

We extracted the genotypes for the 48 candidate SNPs (**Table 1**) from the SNP-array data generated by the Genome-Wide Human SNP Array 6.0 (Affymetrix Inc., Thermo Fisher Scientific, Waltham, MA, United States). These SNP-array data contain 905,171 available SNPs (GRCh37 build). To obtain an unbiased estimation of the population structure of our samples, we processed the SNP-array dataset of the 87 patients with GGE and the 340 BIPMed controls according to previous processing recommendations and pipelines (Anderson et al., 2010; Secolin et al., 2019). First, we removed ambiguous variants (with G/C or A/T alleles) from each dataset. Next, we merged the two datasets into one larger admixed Brazilian dataset ($N = 427$), maintaining only biallelic SNPs, autosomal SNPs, SNPs without Hardy–Weinberg disequilibrium (p-value <0.000001), and missing data $<10\%$. Then, we estimated the heterozygosity rate for each sample and removed individuals with heterozygosity rates higher or lower than three standard deviations from the mean to avoid individuals with high inbreeding (low heterozygosity rates) or sample contamination (high heterozygosity rates). We also removed pairs of individuals who presented a proportion

TABLE 1 | Descriptive statistics and logistic regression analysis of the candidate SNPs and the nine flanking SNPs that passed the most stringent Bonferroni correction (p-value = 3.55e^{-06}; in bold).

Chr	Position (BP)	dbSNP	A1/A2	Reference (PMID)	Reference effect sizes	MAF (A1)	HWE p-value	OR (95% CI) (A1)	Nominal p-value
1	10046460	rs12136213*	G/A	22242659	–	0.268	0.7882	0.76 (0.49–1.19)	0.2285
1	239970097	rs12059546	G/A	25271899; 22949513	1.42 (1.26–1.61)	0.236	0.0562	0.86 (0.56–1.32)	0.4876
1	**240605694**	**rs12726617**	**C/T**	**–**	**–**	**0.412**	**0.0015**	**2.44 (1.96–3.54)**	**2.62e^{-06}**
1	**240610720**	**rs9428842**	**A/G**	**–**	**–**	**0.052**	**0.6117**	**5.86 (2.81–12.2)**	**2.29e^{-06}**
2	23898317	rs4665630	C/T	22242659	–	0.176	0.4723	1.11 (0.68–1.82)	0.6693
2	57934055	rs13026414*	T/C	25271899; 22242659	1.51 (0.81–2.83)	0.349	1.0000	0.97 (0.64–1.46)	0.8705
2	57950346	rs4671319*	G/A	22242659	–	0.422	0.1598	0.70 (0.47–1.04)	0.0765
2	58042241	rs1402398*	G/A	22242659	–	0.35	0.0147	0.77 (0.52–1.14)	0.1906
2	58051769	rs12185644	C/A	22242659	–	0.333	0.1557	1.03 (0.70–1.53)	0.8749
2	58059803	rs2947349*	C/A	25087078	1.23 (1.16–1.31)	0.307	0.0064	0.83 (0.55–1.25)	0.3731
2	145359909	rs10496964	T/C	25271899; 22949513	0.68 (0.60–0.78)	0.138	0.8271	1.29 (0.77–2.17)	0.3388
2	145381225	rs13020210*	G/A	22242659	–	0.213	0.4299	1.15 (0.73–1.80)	0.5395
2	166943277	rs11890028*	G/T	25271899; 22949513; 22242659	0.85 (0.79–0.92)	0.257	0.3361	1.08 (0.69–1.69)	0.7323
2	167084615	rs13406236*	C/T	22242659	–	0.294	1.0000	0.96 (0.63–1.46)	0.8425
2	191583507	rs887696*	C/T	22242659	–	0.401	0.8268	1.28 (0.86–1.88)	0.2198
3	**61699969**	**rs1915992**	**A/G**	**–**	**–**	**0.221**	**0.6469**	**3.62 (2.27–5.76)**	**5.77e^{-08}**
3	61733962	rs624755	G/T	22242659	–	0.368	1.0000	0.81 (0.53–1.22)	0.3040
3	63075267	rs1374679	C/T	22242659	–	0.379	0.315	0.97 (0.65–1.46)	0.8945
3	66326302	rs782728*	A/G	22242659	–	0.468	1.0000	0.89 (0.61–1.30)	0.5499
3	**167113205**	**rs1464634**	**T/G**	**–**	**–**	**0.186**	**0.0026**	**4.03 (2.35–6.93)**	**4.48e^{-07}**
3	167861408	rs111577701*	T/C	25087078	1.16 (1.09–1.24)	0.074	0.4246	0.90 (0.43–1.89)	0.7834
4	31147874	rs1044352	T/G	25087078; 22242659	1.13 (1.12–1.23)	0.454	0.3963	1.17 (0.79–1.73)	0.4305
4	31151357	rs28498976*	A/G	22242659	–	0.446	0.2002	1.26 (0.84–1.88)	0.2629
4	46240287	rs535066*	G/T	22242659	–	0.413	1.0000	0.95 (0.64–1.41)	0.8021
4	46397617	rs11943905*	T/C	22242659	–	0.293	0.3092	1.19 (0.80–1.78)	0.3828
5	114221505	rs4596374*	C/T	22242659	–	0.475	0.0263	0.89 (0.62–1.29)	0.5498
5	114268470	rs55670112*	C/A	25087078	1.18 (1.10–1.26)	0.482	0.0458	1.19 (0.83–1.72)	0.3456
5	150840380	rs357608*	T/C	22242659	–	0.481	0.2452	0.97 (0.67–1.40)	0.8506
5	162867195	rs2069347	C/T	22242659	–	0.475	0.8331	1.18 (0.80–1.73)	0.3981
5	166893257	rs1025482*	C/T	22242659	–	0.488	0.4619	1.03 (0.71–1.49)	0.8813
5	166932520	rs1432881	T/C	22242659	–	0.407	0.2745	1.03 (0.69–1.54)	0.8899
5	167913510	rs244903*	G/A	22242659	–	0.364	0.4278	1.14 (0.76–1.71)	0.5419
6	16971575	rs68082256*	A/G	22242659	–	0.183	0.2934	0.76 (0.44–1.29)	0.3077
6	**17155461**	**rs6459526**	**T/C**	**–**	**–**	**0.222**	**0.0934**	**3.57 (2.15–5.92)**	**8.49e^{-07}**
6	128309768	rs13200150*	G/A	22242659	–	0.222	0.1711	0.92 (0.57–1.49)	0.7490
11	102595135	rs1939012*	A/G	25087078; 22242659	1.12 (1.07–1.17)	0.446	0.6695	0.60 (0.40–0.90)	0.0133
11	**102948592**	**rs2510087**	**A/G**	**–**	**–**	**0.217**	**0.0865**	**3.17 (1.96–5.14)**	**2.76e^{-06}**
13	23966145	rs1008812*	A/G	22242659	–	0.465	1.0000	0.86 (0.59–1.26)	0.4385
13	**24615989**	**rs9551042**	**A/G**	**–**	**–**	**0.243**	**0.1530**	**0.22 (0.11–0.41)**	**2.81e^{-06}**
13	91417190	rs1332470	C/T	22242659	–	0.307	0.0001	1.10 (0.76–1.58)	0.6244
16	30914626	rs1046276	T/C	22242659	–	0.331	1.27e^{-13}	–	-
16	**31310372**	**rs9888879**	**C/T**	**–**	**–**	**0.317**	**0.3935**	**3.65 (2.27–5.85)**	**7.91e^{-08}**
16	50045839	rs4638568	A/G	22242659	–	0.079	0.7125	1.64 (0.84–3.18)	0.1466
17	46027565	rs12951323*	A/C	22949513; 22242659	0.79 (0.72–0.86)	0.206	0.1075	0.67 (0.39–1.15)	0.1429
17	46045495	rs4794333*	A/G	22242659	–	0.356	0.0511	0.64 (0.42–0.97)	0.0361
17	46123004	rs72823592*	A/G	25271899; 22949513	0.77 (0.71–0.83)	0.181	0.7209	0.65 (0.37–1.13)	0.1262
18	48402338	rs2665558*	T/C	30719716; 22242659	–	0.451	0.6707	0.95 (0.64–1.40)	0.7980
18	48404784	rs2255610*	G/A	30719716; 22242659	–	0.474	0.7518	1.14 (0.77–1.67)	0.5199
18	48407326	rs608781*	C/T	30719716; 22242659	–	0.118	0.4472	1.72 (0.98–3.02)	0.0610
18	48414235	rs2850545*	A/C	30719716; 22242659	–	0.471	0.5976	1.12 (0.75–1.65)	0.5845
18	48456903	rs645088	T/C	30719716; 22242659	–	0.339	0.4812	0.83 (0.55–1.26)	0.3873
18	48458662	rs649224	A/G	30719716; 22242659	–	0.107	0.7821	1.50 (0.83–2.70)	0.1834

(Continued)

TABLE 1 | Continued

Chr	Position (BP)	dbSNP	A1/A2	Reference (PMID)	Reference effect sizes	MAF (A1)	HWE *p*-value	OR (95% CI) (A1)	Nominal *p*-value
18	48464204	rs654136	T/C	30719716; 22242659	–	0.488	0.5266	1.02 (0.69–1.51)	0.9271
19	53719250	rs9788	A/G	22242659	–	0.315	0.9032	1.57 (1.05–2.34)	0.0274
21	32183996	rs2833098*	G/A	22242659	–	0.369	0.9099	1.11 (0.75–1.64)	0.6152
21	**48063151**	**rs8133217**	**G/A**	**–**	**–**	**0.214**	**0.0422**	**0.15 (0.07–0.32)**	**4.94e^{-07}**
21	48077812	rs2839377	T/C	22242659	–	0.497	0.4605	1.07 (0.72–1.57)	0.7491

The positions are based on GRCh37. SNPs with an asterisk (*) were obtained by BEAGLE imputation.
BP, base pairs; PMID, PUBMED ID publications; MAF, minor allele frequency; HWE, Hardy–Weinberg equilibrium; OR, odds ratio; CI, confidence interval.

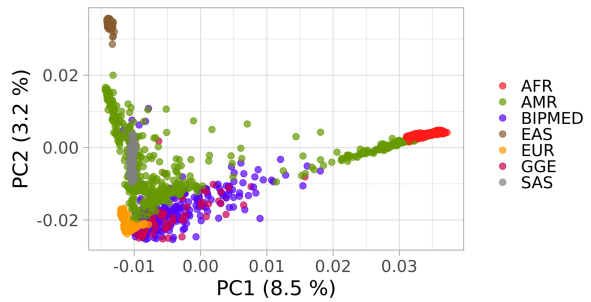

FIGURE 1 | Principal component analysis (PCA) plot of the Brazilian samples and 1000 Genomes Project (1KGP) dataset. The *x*-axis and *y*-axis show the first and second principal components (PC1 and PC2) and their respective percentage variability. Each point represents one individual and each color indicates patients with genetic generalized epilepsy (GGE), BIPMed controls, and the continental populations described in 1KGP. as follows: Sub-Saharan Africans (AFR); Europeans (EUR); admixed Americans (AMR); Southwestern Asians (SAS); Eastern Asians (EAS).

of identical-by-state (IBS) alleles >0.85, which could indicate duplicated samples, and individuals with genomic relatedness matrix estimations higher than 0.125, which is the expected genomic relatedness for third-degree relatives (Anderson et al., 2010). The merging process, genotyping, and sample filtering were performed using PLINK 1.9 software (Purcell et al., 2007).

Subsequently, we merged the filtered admixed Brazilian sample with the 1000 Genomes Project (1KGP) dataset (The 1000 Genomes Project Consortium et al., 2015), maintaining the SNPs present only in the admixed Brazilian sample. After merging, we removed SNPs with a minor allele frequency (MAF) < 0.01 and SNPs in linkage disequilibrium (LD), using the following parameters: window size = 50 SNPs, shift step = 5 SNPs, and r^2 = 0.5 (Anderson et al., 2010). We compared our dataset with the 1KGP data by principal component analysis (PCA) using PLINK v1.9 software (Purcell et al., 2007) to evaluate the presence of population-based outliers in the Brazilian samples.

To evaluate whether patients with GGE and BIPMed controls present population stratification, we performed the analysis of molecular variance (AMOVA) (Excoffier et al., 1992) using the poppr.amova R package and the RStudio interface, comparing the genetic distance among the two groups based on a set of 10,000 random SNPs across the genome. The AMOVA partitions the source of genetic variance (σ^2) into two components: within-groups and between-groups. The null hypothesis states that the samples were obtained from a global population, with variation due to random sampling in the construction of populations. Thus, we would expect a high heterogeneity within groups (σ^2 = 100%) and no heterogeneity between groups (σ^2 = 0%). On the other hand, under the alternative hypothesis, each group was obtained from different populations, and we would expect a low heterogeneity within groups (σ^2 < 100%) and high heterogeneity between groups (σ^2 > 0%) (Excoffier et al., 1992). Therefore, to evaluate the significance of σ^2 components, we generated a Monte Carlo null distribution of 10,000 variance components and tested against the observed variance components by the randtest function in the ade4 R package.

Single-Nucleotide Polymorphism Selection and Imputation

We observed that 31 SNPs were not found in the SNP-array dataset. Therefore, we performed an imputation of all 48 SNPs to obtain the missing SNPs and to evaluate the concordance between the imputed genotypes and the genotypes assessed by the SNP array. Since we analyzed a sample of admixed individuals, we elected to perform the imputation using two approaches. First, we phased and imputed the dataset using SHAPEIT2 v2.r387 (O'Connell et al., 2014) and BEAGLE v5.1 software (Browning et al., 2018) using the default software parameters for phasing and imputation. As a reference for the BEAGLE imputation, we used the 1KGP dataset (GRCh37/hg19 assembly) (The 1000 Genomes Project Consortium et al., 2015). To save on computation time, we imputed only the chromosomes in which the candidate SNPs are located (**Table 1**). We also evaluated whether the genotypes were successfully imputed by the correlation (in terms of r^2) of genotype dosage values between the imputed genotypes and true genotypes used as a reference from the 1KGP provided by the BEAGLE software. For the second imputation approach, we used the TOPMED Imputation Server (Das et al., 2016), with the TOPMed v.R2 on GRCh38 build (Kowalski et al., 2019). The TOPMED server imputation performed the liftover from GRCh37 to GRCh38 and the phasing using the EAGLE v.2.4 algorithm. Finally, imputation was performed by minimac4.

Candidate Single-Nucleotide Polymorphism Association Analysis

After genotype and individual filtering, 360 individuals remained (69 patients with GGE and 291 BIPMed controls), which were used in the association analysis. We estimated the statistical power of our sample by the genpwr package in R (Moore et al.,

TABLE 2 | AMOVA results.

Variance component	Variance σ^2	Percentage of variance	Total Φ-statistics	*p*-Value
Between samples	0.1884	0.39	0.00393	0.001
Within samples	47.7333	99.61		
Total	47.9217	100.00		

Variance component estimations are based on the genetic distance among patients with GGE and controls, including the Monte Carlo test p-value. Values were estimated based on 10,000 (10k) random autosome SNPs across the genome. AMOVA, analysis of molecular variance; SNPs, single-nucleotide polymorphisms; GGE, genetic generalized epilepsy.

2020), which analyzes the statistical power under the evaluation between true and test genetic models (Dominant, Additive, Recessive, 2df/unspecified model). In this case, we evaluate the statistical power using a vector of MAFs (from 0.05 to 0.45, by 0.05) and an odds ratios (from 1.5 to 2.0, by 0.1) since not all candidate SNPs presented OR estimations from previous studies. We also set the following parameters for genpwr: model = logistic; N = 360; case/control ratio = 69/291 = 0.237; and alpha = 0.05.

We evaluated candidate SNP association and OR estimation by logistic regression analysis using the PLINK v1.9 software (Purcell et al., 2007), including the first five PCs as covariates. We did not include age, age at seizure onset, and sex since these variables have not been correlated with the GGE phenotype (Berg et al., 2010; Scheffer et al., 2018).

It has been reported that SNPs found to be associated with the phenotype by GWAS in one population may be only nominally associated or non-associated in another population due to difference in LD across populations (Akiyama et al., 2019; Chen et al., 2020; Graff et al., 2021); however, it does not mean that an associated signal in the genomic region cannot be replicated. This is because the SNPs ascertained from GWAS are only tagging variants linked to causal ones. The lack of signals in the replication population could simply be caused by the broken linkage between tagging and causal variants. Therefore, to account for the difference in LD across populations and to investigate the transferability of previous GWAS signals, we used the SNP-array dataset, filtered for population structure and without LD pruning (652,883 SNPs), to interrogate the SNPs flanking the 1 Mb upstream and downstream the candidate SNPs by logistic regression. We assumed a *p*-value adjusted by the Bonferroni correction to avoid biased results due to the multiple comparisons. In this case, we used two thresholds: the first threshold took into account the 48 SNPs (*p*-value = 0.05/48 = 0.001), assuming one effective test per region, which is a reasonable assumption and may lead to more informative results. However, this threshold may not be stringent enough. Therefore, we also evaluate the results under a second threshold, considering all the 48 candidate SNPs and the additional flanking SNPs tested, and the results were plotted using the qqman package in R software (Turner, 2014).

Since previous studies of GGE were based on European populations and admixed Brazilians have a large proportion of European ancestry, we decided to evaluate whether the candidate SNP allele frequencies are similar between Brazilian and European populations. We extracted European allele frequencies from the gnomAD database (Karczewski et al., 2020) and performed a two-proportion Z-test using the prop.test function in R. Also, we included African populations from

FIGURE 2 | Statistical power estimation. The figure includes six panels **(A–F)** for the six OR values evaluated. Each panel shows the statistical power (*y*-axis) estimated by minor allele frequency (MAF) (*x*-axis) for the combination of the three true genetic models (Dominant, Additive, and Recessive) with the four test models (Dominant, Additive, Recessive, and 2df/unspecified genetic model). Each colored line represents the test model, and each point represents the MAF.

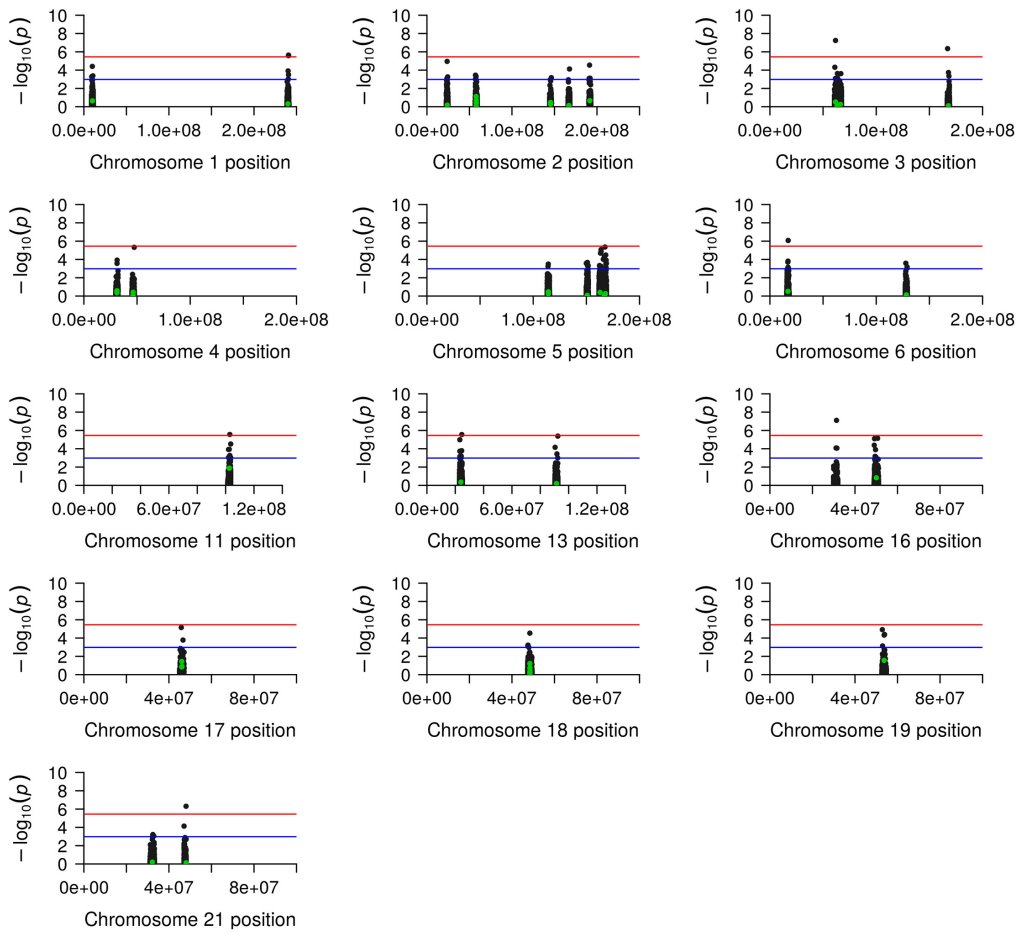

FIGURE 3 | Manhattan plot of genetic generalized epilepsy (GGE) candidate regions. The figure shows a plot for each candidate chromosome, including the chromosome position (*x*-axis) and the –log10(*p*-value) in the *y*-axis. The green and black points represent the candidate and the flanking single-nucleotide polymorphisms (SNPs), respectively. The blue line indicates the suggestive association signal based on the *p*-value adjusted by the Bonferroni correction under the 48 candidate SNPs. The red line indicates the association signal based on the *p*-value adjusted by Bonferroni under the 48 candidate SNPs plus the 14,047 flanking SNPs.

gnomAD in the analysis due to the sub-Saharan African ancestry component present in Brazilian populations. However, since gnomAD does not separate Native American populations in the database, we include the Latin population in the analysis as a proxy.

RESULTS

Population Structure Analysis

The principal components in the PCA plot indicate that both cases and controls clustered together and were spread between Europeans, sub-Saharan Africans, and other admixed American populations (**Figure 1**). The AMOVA results showed that 99.61% of the genetic variation was observed within groups (patients or controls), and only 0.39% of the genetic variation was observed between groups (**Table 2**). Because we have one hierarchical level of stratification (patients/controls), the poppr.amova package provided one total φ-statistics = 0.0031, with a *p*-value = 0.001 (**Table 2**), indicating evidence of population stratification between patients and controls and the necessity of population structure correction in further association tests.

Single-Nucleotide Polymorphism Selection and Imputation

According to the imputation results from the BEAGLE software (Browning et al., 2018), the correlation between the estimated allele dosage and the true allele dosage from the 1KGP is used as reference (in terms of r^2) and presented a minimum value of 95%. In addition, all 17 SNPs genotyped by the SNP array were correctly imputed by the BEAGLE software. However, we observed that the 17 SNPs genotyped by the SNP array presented only 45.7% of matching (on average) with genotypes imputed by the TOPMED server. Thus, we decided to perform further analysis using the imputed genotypes generated by the BEAGLE software.

Candidate Single-Nucleotide Polymorphism Association Analysis

As detailed in **Table 1**, one candidate SNP (rs1046276) was withdrawn from further association analysis due to the presence of the Hardy–Weinberg disequilibrium ($p < 0.000001$). According to the analysis performed using the genpwr package (Moore et al., 2020), we observed that the Additive model presented the highest power estimation. We did not observe 80% of statistical power for OR $\leq$ 1.6 ($\geq$ 0.62 for protection effect) (**Figures 2A,B**). However, we calculated that our study had 80% power to detect an increased risk in terms of OR $\geq$ 1.7 ($\leq$0.58 for protection effect) with MAF > 0.25 (**Figure 2C**), OR $\geq$ 1.8 ($\leq$0.55 for protection effect) with MAF > 0.2 (**Figure 2D**), and OR $\geq$ 1.9 ($\leq$0.52 for protection effect) with MAF > 0.15 (**Figures 2E,F**).

We identified suggestive evidence of a protective effect for the SNP rs1939012*A allele (MAF = 0.446; OR = 0.60; 95% CI = 0.40–0.90; nominal *p*-value = 0.0133) and rs4794333*A allele (MAF = 0.356; OR = 0.64; 95% CI = 0.42–0.97; nominal *p*-value = 0.0361) and an increased risk for rs9788*G (MAF = 0.315; OR = 1.57; 95% CI = 1.05–2.34, nominal *p*-value = 0.0274). However, these results did not pass the corrections for multiple comparisons by Bonferroni (**Table 1**). Interesting, we found 14,047 flanking SNPs, encompassing 29 candidate regions (**Supplementary Data**). As shown in **Figure 3**, under the *p*-value threshold = 0.001, we observed that the association signals in all candidate regions passed the Bonferroni correction. Adjusting the *p*-values by Bonferroni for 14,095 tests (*p*-value = $3.55e^{-06}$), we observed that nine flanking SNPs passed the multiple comparison adjustment: rs12726617 and rs9428842 on chromosome (chr.) 1q43; rs1915992 on chr. 3p14.2; rs1464634 on chr. 3q26.1; rs6459526 on chr. 6p22.3; rs2510087 on chr. 11q22.3; rs9551042 on chr. 13q12.12; rs9888879 on chr. 16p11.2; and rs8133217 on chr. 21q22.3 (**Figure 3** and **Table 1**).

Since most Brazilian ancestry is derived from European populations (Kehdy et al., 2015; Moura et al., 2015; Secolin et al., 2019), we could hypothesize that effect sizes in terms of OR would present a higher correlation with European effect sizes comparing with Chinese or European/African American samples from previous studies (EPICURE Consortium et al., 2012a,b; International League Against Epilepsy Consortium on Complex Epilepsies (ILAE Consortium on Complex Epilepsies), 2014; International League Against Epilepsy Consortium on Complex Epilepsies (ILAE Consortium on Complex Epilepsies)., 2018; Zhang et al., 2014; Wang et al., 2019). Thus, we show a comparison of the OR estimations of 11 SNPs, which were available from the previous studies, and the OR estimations in our admixed Brazilian samples (**Table 3**). Remarkably, Chinese and European/African American samples also presented similar OR estimations compared with admixed Brazilians. Two SNPs had different OR estimations for admixed Brazilians compared with European and Chinese samples (rs10496964 and rs11890028).

Furthermore, the two-proportion Z-test results showed that 25 candidate SNPs have allele frequencies that were different when comparing admixed Brazilian and the ancestral populations. Among them, 16 SNPs presented differences in allele frequencies comparing admixed Brazilian and European populations. All 25 SNPs presented different allele frequencies when comparing admixed Brazilian and African samples. Remarkably, we also found 15 candidate SNPs with different allele frequencies when comparing admixed Brazilians and the Latin American samples in the gnomAD database (**Table 4**).

TABLE 3 | Odds ratio comparison among studies.

SNP	Brazil	Population from previous studies		
		European (PMID: 22949513)	Chinese (PMID: 25271899)	European/African Americans (PMID: 25087078)
rs12059546	0.86 (0.56–1.32)	1.53 (1.32–1.79)	0.93 (0.57–1.53)	–
rs13026414	0.97 (0.64–1.46)	0.78 (0.71–0.86)	1.51 (0.81–2.83)	–
rs2947349	0.83 (0.55–1.25)	–	–	1.23 (1.16–1.31)
rs10496964	1.29 (0.77–2.17)	0.63 (0.52–0.76)	0.50 (0.18–1.40)	–
rs11890028	1.08 (0.69–1.69)	0.77 (0.70–0.85)	0.77 (0.26–2.24)	–
rs111577701	0.90 (0.43–1.89)	–	–	1.16 (1.09–1.24)
rs1044352	1.17 (0.79–1.73)	–	–	0·88 (0·82–0·93)
rs55670112	1.19 (0.83–1.72)	–	–	1.18 (1.10–1.26)
rs1939012	0.60 (0.40–0.90)	–	–	1.12 (1.07–1.17)
rs12951323	0.67 (0.39–1.15)	0.75 (0.66–0.84)	–	–
rs72823592	0.65 (0.37–1.13)	0.74 (0.66–0.83)	–	–

The table shows the OR estimation available in each study, including the 95% confidence interval in parentheses.
SNP, single-nucleotide polymorphism.

DISCUSSION

The Brazilian population was formed by an admixture of three main ancestry populations: Europeans, sub-Saharan Africans, and Native Americans (Kehdy et al., 2015; Moura et al., 2015; Secolin et al., 2019). In this scenario, it is important to explore whether candidate SNPs previously identified as associated with complex disorders in non-admixed populations also display association signals in the Brazilian admixed population. By doing so, one can better estimate the impact of population structure in estimating polygenic risks, avoiding misinterpretation of risk scores calculated in other populations.

Previous genetic association studies have identified 48 candidate SNPs associated with GGEs (EPICURE Consortium et al., 2012a,b; International League Against Epilepsy Consortium on Complex Epilepsies (ILAE Consortium on Complex Epilepsies), 2014; International League Against Epilepsy Consortium on Complex Epilepsies (ILAE Consortium on Complex Epilepsies)., 2018; Zhang et al., 2014; Wang et al., 2019). These studies were all performed in non-admixed populations, predominantly of European ancestry, raising the question of reproducibility of these results in other populations. Lack of transferability of GWAS results and polygenic risk scores

TABLE 4 | Two-proportion Z-test results comparing Brazilian samples with European, African, and Latin-American samples from gnomAD.

SNP ID	Allele	Brazilian samples	European vs. Brazilian samples		African vs. Brazilian samples		Latin American vs. Brazilian samples	
		Allele frequency	Allele frequency	*p*-Value	Allele frequency	*p*-Value	Allele frequency	*p*-Value
rs12136213	G	0.268	0.282	1	0.105	$3.66e^{-37}$	0.257	1
rs4665630	G	0.176	0.896	0	0.511	$6.45e^{-65}$	0.895	$6.43e^{-178}$
rs4671319	T	0.422	0.521	$6.37e^{-06}$	0.096	$2.79e^{-55}$	0.417	1
rs1402398	G	0.350	0.615	$4.35e^{-44}$	0.708	$3.69e^{-114}$	0.558	$7.64e^{-15}$
rs2947349	G	0.307	0.617	$5.11e^{-60}$	0.749	0	0.567	$8.40e^{-23}$
rs10496964	C	0.138	0.163	1	0.590	$3.37e^{-35}$	0.091	$1.24e^{-01}$
rs13020210	C	0.213	0.832	0	0.922	$1.36e^{-41}$	0.861	$1.31e^{-144}$
rs11890028	A	0.257	0.278	1	0.036	$3.32e^{-05}$	0.205	$4.62e^{-01}$
rs887696	G	0.401	0.660	$2.88e^{-44}$	0.479	$1.36e^{-83}$	0.481	$4.83e^{-02}$
rs111577701	G	0.074	0.131	$2.21e^{-04}$	0.183	$5.95e^{-17}$	0.057	1
rs28498976	C	0.446	0.379	$8.41e^{-03}$	0.689	0.00019	0.313	$2.12e^{-06}$
rs11943905	C	0.293	0.265	1	0.743	$6.70e^{-13}$	0.184	$1.25e^{-05}$
rs55670112	C	0.482	0.457	1	0.386	$6.63e^{-11}$	0.541	$5.62e^{-01}$
rs68082256	C	0.183	0.207	1	0.792	$9.55e^{-25}$	0.209	1
rs1939012	A	0.446	0.498	$1.82e^{-01}$	0.755	$1.01e^{-47}$	0.517	$1.60e^{-01}$
rs1008812	T	0.465	0.490	1	0.210	0.00884	0.579	$2.24e^{-04}$
rs4638568	T	0.079	0.057	$3.87e^{-01}$	0.553	$7.79e^{-20}$	0.031	$8.21e^{-04}$
rs12951323	A	0.206	0.788	$9.49e^{-280}$	0.533	$1.65e^{-88}$	0.882	$2.92e^{-158}$
rs4794333	G	0.356	0.390	1	0.704	0.00814	0.290	$1.62e^{-01}$
rs72823592	T	0.181	0.241	$6.02e^{-03}$	0.177	$6.01e^{-09}$	0.121	$2.91e^{-02}$
rs608781	C	0.118	0.928	0	0.618	$4.02e^{-194}$	0.900	$2.81e^{-208}$
rs645088	C	0.339	0.625	$5.61e^{-52}$	0.351	$1.73e^{-152}$	0.781	$4.82e^{-68}$
rs649224	T	0.107	0.074	$2.08e^{-02}$	0.296	$2.96e^{-28}$	0.064	$1.13e^{-04}$
rs9788	G	0.489	0.617	$4.18e^{-14}$	0.316	$6.18e^{-228}$	0.609	$9.30e^{-06}$
rs2839377	C	0.497	0.535	1	0.512	$1.21e^{-11}$	0.463	1

p-Values are adjusted by the Bonferroni correction. We show the results for the 25 SNPs presenting significant differences in allele frequencies among populations. SNP, single-nucleotide polymorphism.

obtained from Europeans and American admixed populations have previously been reported (Martin et al., 2017, 2019), making it important to investigate whether these SNPs are associated with GGEs in our admixed Brazilian sample.

An alternative explanation for the lack of reproducibility among populations relies on the observation that only tagging SNPs are ascertained in GWAS, and the lack of replication in different populations could be due to broken linkage between the tagging SNPs and the causal variants (Akiyama et al., 2019; Chen et al., 2020; Graff et al., 2021). Thus, we searched for SNPs flanking 1 Mb upstream and downstream of the candidate regions to investigate this issue. Indeed, we found 14,047 flanking SNPs, and nine of them presented statistically significant association signals after stringent corrections for multiple comparisons (p-value $< 3.55e^{-06}$). These nine SNPs encompass eight candidate regions (**Table 2** and **Figure 3**), which were previously found associated in European samples (1q43; 3p14.2; 3q26.1; 6p22.3; 11q22.3; 13q12.12; 16p.11.2; and 21q22.3), and two of them were also found associated in a mixed sample of European, African, and Asian populations (3q26.1 and 16p.11.2) (EPICURE Consortium et al., 2012a,b; International League Against Epilepsy Consortium on Complex Epilepsies (ILAE Consortium on Complex Epilepsies), 2014; International League Against Epilepsy Consortium on Complex Epilepsies (ILAE Consortium on Complex Epilepsies)., 2018; Wang et al., 2019). Therefore, we may suggest that polygenic risk scores calculated in European populations at these specific loci could indeed be transferable to admixed Brazilian individuals.

However, although all these 29 candidate regions passed the Bonferroni correction based on the 48 candidate SNPs (p-value = 0.001), we understand that this p-value threshold is not stringent. Thus, the lack of association signal cannot be discarded for the 20 remaining candidate regions. Thus, one may still speculate that the lack of reproducibility could be due to the absence of statistical power, population stratification, or the differences in the genomic structure of the admixed sample compared with the previously studied populations.

Although we have identified flanking SNPs in the neighborhood of the candidate regions, which presented 80% of statistical power to detect increased risk or protection allele effect, we acknowledge the limited statistical power provided by the cohort analyzed, with 87 patients with GGE and 340 controls.

Despite the observed high heterogeneity within groups (σ^2 = 99.61%) and low heterogeneity between patients and controls (σ^2 = 0.39%), the statistics based on AMOVA results

revealed evidence of population stratification between patients with GGE and the BIPMed controls. Thus, we corrected possible spurious association results by taking the first five principal components into account in the logistic regression model (Marchini et al., 2004; Price et al., 2010).

Indeed, the two-proportion Z-test showed that 16 SNPs presented different allele frequencies when comparing admixed Brazilian and European samples, further substantiating the hypothesis of lack of genetic association due to genetic differences when comparing the admixed Brazilians and Europeans.

It is important to note that 31 SNPs were not found in the SNP-array dataset, and we decided to impute them from all populations available in the 1KGP dataset (The 1000 Genomes Project Consortium et al., 2015). Previous studies have demonstrated that imputation accuracy for populations with a high proportion of European ancestry is higher than for populations with African or Native American ancestry (Martin et al., 2017). In addition, the EPIGEN-Brazil Initiative has also imputed admixed Brazilian samples from the 1KGP dataset with high confidence variants (Magalhães et al., 2018). However, the imputation by the TOPMED Consortium has demonstrated improved quality of variant imputation for admixed African and Hispanic/Latin populations compared with the 1KGP dataset (Kowalski et al., 2019). Thus, we also used this approach for comparison. We observed a perfect match between the SNPs genotyped in the SNP-array and their imputed correspondents for the BEAGLE imputation using the 1KGP as reference. By contrast, there was only 45.7% correspondence between the SNPs genotyped and the imputed SNPs using TOPMED. Thus, we can argue that Hispanic/Latin samples included in the TOPMED reference panel (Kowalski et al., 2019) may not represent the genomic structure of admixed Brazilians (Adhikari et al., 2016). This is an important finding and indicates that although allele frequencies of admixed Brazilian populations are different from other populations reported in public databases (Adhikari et al., 2016; Magalhães et al., 2018; Rocha et al., 2020), there is a remarkable accuracy in the SNP imputation for admixed Brazilian individuals based on populations from the 1KGP database, as demonstrated by our results and elsewhere (Magalhães et al., 2018).

In conclusion, we replicated association signals on eight candidate regions previously found in European populations, indicating the possibility of transferability of polygenic risk scores from European studies to admixed Brazilian populations in these specific candidate regions. In addition, we show evidence that differences in the genetic architecture of the population may hinder the replication of association results in admixed Brazilians for the remaining candidate regions, thus supporting the hypothesis of population differences influencing the association results in the present study. Also, we documented the effect of different methods/databases used for genotype imputation in admixed Brazilians. These results could be relevant to improving stratification risk estimation and future precision health applications in admixed Brazilian patients with GGEs and other complex disorders.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ebi.ac.uk/eva/, PRJEB39251; https://www.ebi.ac.uk/ena, PRJEB45235.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Comitê de Ética em Pesquisa da Universidade Estadual de Campinas. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

FK performed data processing, statistical analysis, and imputation. TA and PA performed data acquisition and SNP array genotyping. MA, CY, and FC performed the clinical analysis of GGE patients. FC and IL-C served as the principal investigators. RS conceptualized the work, created the study design, and served as a principal investigator. All authors reviewed and approved the final version of the manuscript.

FUNDING

This work was supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brazil, grant number 2013/07559-3; and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, grant number: 001), Brazil. FK was supported by a studentship from Conselho Nacional de Pesquisa (CNPq), Brazil. IL-C is supported by CNPq (grant number 311923/2019-4). RS was supported by FAPESP (grant number 2019/08526-8).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.672304/full#supplementary-material

REFERENCES

The 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

Aaberg, K. M., Surén, P., Søraas, C. L., Bakken, I. J., Lossius, M. I., Stoltenberg, C., et al. (2017). Seizures, syndromes, and etiologies in childhood epilepsy: the International League Against Epilepsy 1981, 1989, and 2017 classifications used in a population-based cohort. *Epilepsia* 58, 1880–1891. doi: 10.1111/epi.13913

Adhikari, K., Mendoza-Revilla, J., Chacón-Duque, J. C., Fuentes-Guajardo, M., Ruiz-Linares, A., Chacón-Duque, J. C., et al. (2016). Admixture in Latin America. *Curr. Opin. Genet. Dev.* 41, 106–114. doi: 10.1016/j.gde.2016.09.003

Akiyama, M., Ishigaki, K., Sakaue, S., Momozawa, Y., Horikoshi, M., Hirata, M., et al. (2019). Characterizing rare and low-frequency height-associated variants in the Japanese population. *Nat. Commun.* 10:4393. doi: 10.1038/s41467-019-12276-5

Anderson, C., Pettersson, F., Clarke, G., Cardon, L., Morris, A., and Zondervan, K. (2010). Data quality control in genetic case-control association studies. *Nat. Protoc.* 5, 1564–1573. doi: 10.1038/nprot.2010.116

Berg, A. T., Berkovic, S. F., Brodie, M. J., Buchhalter, J., Cross, J. H., Van Emde Boas, W., et al. (2010). Revised terminology and concepts for organization of seizures and epilepsies: report of the ILAE Commission on Classification and Terminology, 2005-2009. *Epilepsia* 51, 676–685. doi: 10.1111/j.1528-1167.2010.02522.x

Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* 103, 338–348. doi: 10.1016/j.ajhg.2018.07.015

Chen, M. H., Raffield, L. M., Mousas, A., Sakaue, S., Huffman, J. E., Moscati, A., et al. (2020). Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell* 182, 1198–1213.e14. doi: 10.1016/j.cell.2020.06.045

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

EPICURE Consortium, EMINet Consortium, Steffens, M., Leu, C., Ruppert, A. K., Zara, F., et al. (2012a). Genome-wide association analysis of genetic generalized epilepsies implicates susceptibility loci at 1q43, 2p16.1, 2q22.3 and 17q21.32. *Hum. Mol. Genet.* 21, 5359–5372. doi: 10.1093/hmg/dds373

EPICURE Consortium, Leu, C., de Kovel, C., Zara, F., Striano, S., Pezzella, M., et al. (2012b). Genome-wide linkage meta-analysis identifies susceptibility loci at 2q34 and 13q31.3 for genetic generalized epilepsies. *Epilepsia* 53, 308–318. doi: 10.1111/j.1528-1167.2011.03379.x

Excoffier, L., Smouse, P. E., and Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131, 479–491. doi: 10.5962/bhl.title.86657

Fisher, R. S., Acevedo, C., Arzimanoglou, A., Bogacz, A., Cross, J. H., Elger, C. E., et al. (2014). ILAE Official Report: a practical clinical definition of epilepsy. *Epilepsia* 55, 475–82. doi: 10.1111/epi.12550

Graff, M., Justice, A. E., Young, K. L., Marouli, E., Zhang, X., Fine, R. S., et al. (2021). Discovery and fine-mapping of height loci via high-density imputation of GWASs in individuals of African ancestry. *Am. J. Hum. Genet.* 108, 564–582. doi: 10.1016/j.ajhg.2021.02.011

International League Against Epilepsy Consortium on Complex Epilepsies (ILAE Consortium on Complex Epilepsies) (2014). Genetic determinants of common epilepsies: a meta-analysis of genome-wide association studies. *Lancet Neurol.* 13, 893–903. doi: 10.1016/S1474-4422(14)70171-1

International League Against Epilepsy Consortium on Complex Epilepsies (ILAE Consortium on Complex Epilepsies). (2018). Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies. *Nat. Commun.* 9:5269. doi: 10.1038/s41467-018-07524-z

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443. doi: 10.1038/s41586-020-2308-7

Kehdy, F. S. G., Gouveia, M. H., Machado, M., Magalhães, W. C. S., Horimoto, A. R., Horta, B. L., et al. (2015). Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. *Proc. Natl. Acad. Sci.U. S. A.* 112, 8696–8701. doi: 10.1073/pnas.1504447112

Kowalski, M. H., Qian, H., Hou, Z., Rosen, J. D., Tapia, A. L., Shan, Y., et al. (2019). Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet.* 15:e1008500. doi: 10.1371/journal.pgen.1008500

Magalhães, W. C. S., Araujo, N. M., Leal, T. P., Araujo, G. S., Viriato, P. J. S., Kehdy, F. S., et al. (2018). EPIGEN-Brazil Initiative resources: a Latin American imputation panel and the Scientific Workflow. *Genome Res.* 28, 1090–1095. doi: 10.1101/gr.225458.117

Marchini, J., Cardon, L. R., Phillips, M. S., and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nat. Genet.* 36, 512–517. doi: 10.1038/ng1337

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

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* 51, 584–591. doi: 10.1038/s41588-019-0379-x

Moore, C. M., Jacobson, S. A., and Fingerlin, T. E. (2020). Power and Sample Size Calculations for Genetic Association Studies in the Presence of Genetic Model Misspecification. *Hum. Hered.* 84, 1–16. doi: 10.1159/000508558

Moura, R. R., de, Coelho, A. V. C., de Queiroz Balbino, V., Crovella, S., Brandão, L. A. C., et al. (2015). Meta-analysis of Brazilian genetic admixture and comparison with other Latin America countries. *Am. J. Hum. Biol.* 27, 674–680. doi: 10.1002/ajhb.22714

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., et al. (2014). A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS Genet.* 10:e1004234. doi: 10.1371/journal.pgen.1004234

Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* 11, 459–463. doi: 10.1038/nrg2813

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., 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

Rocha, C. S., Secolin, R., Rodrigues, M. R., Carvalho, B. S., and Lopes-Cendes, I. (2020). The Brazilian Initiative on Precision Medicine (BIPMed): fostering genomic data-sharing of underrepresented populations. *NPJ Genom. Med.* 5:42. doi: 10.1038/s41525-020-00149-6

Scheffer, I. E., Berkovic, S., Capovilla, G., Connolly, M. B., French, J., Guilhoto, L., et al. (2017). ILAE classification of the epilepsies: position paper of the ILAE Commission for Classification and Terminology. *Epilepsia* 58, 512–521. doi: 10.1111/epi.13709

Scheffer, I. E., Berkovic, S., Capovilla, G., Connolly, M. B., French, J., Guilhoto, L., et al. (2018). ILAE classification of the epilepsies: position paper of the ILAE Commission for Classification and Terminology. *Z. Epileptol.* 31, 296–306. doi: 10.1007/s10309-018-0218-6

Secolin, R., Mas-Sandoval, A., Arauna, L. R., Torres, F. R., Araujo, T. K., Santos, M. L., et al. (2019). Distribution of local ancestry and evidence of adaptation in admixed populations. *Sci. Rep.* 9:13900. doi: 10.1038/s41598-019-50362-2

Turner, S. (2014). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *BioRxiv* [Preprint]. doi: 10.1101/005165

Wang, M., Greenberg, D. A., and Stewart, W. C. L. (2019). Replication, reanalysis, and gene expression: ME2 and genetic generalized epilepsy. *Epilepsia* 60, 539–546. doi: 10.1111/epi.14654

Zhang, Y., Qu, J., Mao, C. X., Wang, Z. B., Mao, X. Y., Zhou, B. T., et al. (2014). Novel Susceptibility Loci were Found in Chinese Genetic Generalized Epileptic Patients by Genome-wide Association Study. *CNS Neurosci. Ther.* 20, 1008–1010. doi: 10.1111/cns.12328

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

REVIEW
published: 04 August 2021
doi: 10.3389/fgene.2021.703541

Quantitative Human Paleogenetics: What can Ancient DNA Tell us About Complex Trait Evolution?

Evan K. Irving-Pease[1], Rasa Muktupavela[1], Michael Dannemann[2] and Fernando Racimo[1]**

[1] Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark, [2] Center for Genomics, Evolution and Medicine, Institute of Genomics, University of Tartu, Tartu, Estonia

Genetic association data from national biobanks and large-scale association studies have provided new prospects for understanding the genetic evolution of complex traits and diseases in humans. In turn, genomes from ancient human archaeological remains are now easier than ever to obtain, and provide a direct window into changes in frequencies of trait-associated alleles in the past. This has generated a new wave of studies aiming to analyse the genetic component of traits in historic and prehistoric times using ancient DNA, and to determine whether any such traits were subject to natural selection. In humans, however, issues about the portability and robustness of complex trait inference across different populations are particularly concerning when predictions are extended to individuals that died thousands of years ago, and for which little, if any, phenotypic validation is possible. In this review, we discuss the advantages of incorporating ancient genomes into studies of trait-associated variants, the need for models that can better accommodate ancient genomes into quantitative genetic frameworks, and the existing limits to inferences about complex trait evolution, particularly with respect to past populations.

Keywords: aDNA, paleogenetics, GWAS, polygenic adaptation, complex traits

OPEN ACCESS

Edited by:
Diego Ortega-Del Vecchyo, National Autonomous University of Mexico, Mexico

Reviewed by:
Iain Mathieson, University of Pennsylvania, United States
Gulsah Merve Kilinc, Hacettepe University, Turkey

****Correspondence:***
Evan K. Irving-Pease
evan.irvingpease@gmail.com
Fernando Racimo
fernandoracimo@gmail.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *30 April 2021*
Accepted: *08 July 2021*
Published: *04 August 2021*

Citation:
Irving-Pease EK, Muktupavela R, Dannemann M and Racimo F (2021) Quantitative Human Paleogenetics: What can Ancient DNA Tell us About Complex Trait Evolution? Front. Genet. 12:703541. doi: 10.3389/fgene.2021.703541

INTRODUCTION

The last decade has seen dramatic advances in our understanding of the genetic architecture of polygenic traits (Visscher et al., 2017). The advent of genome-wide association studies (GWAS), with large sample sizes and deep phenotyping of individuals, has led to the identification of thousands of loci associated with complex traits and diseases (MacArthur et al., 2017; Bycroft et al., 2018; Buniello et al., 2019). The resulting associations, and their inferred effect sizes, have enabled the development of so-called polygenic risk scores (PRS), which summarise either the additive genetic contribution of single nucleotide polymorphisms (SNPs) to a quantitative trait (e.g., height), or the increase in probability of a binary trait (e.g., major coronary heart disease) (Dudbridge, 2013). For some well-characterised medical traits, like cardiovascular disease, the predictive value of PRS has led to their adoption in clinical settings (Knowles and Ashley, 2018); however, the accuracy of PRS remains limited to populations closely related to the original GWAS cohort (Martin et al., 2019) and can vary within populations due to age, sex and socioeconomic status (Mostafavi et al., 2020). Ancient genomics has yielded considerable insights into natural selection on large-effect variants (Malaspinas, 2016; Dehasque et al., 2020), and an increasing number of studies are also now utilizing ancient genomes to learn about polygenic adaptation; the process by which natural selection acts on a trait with a large number of genetic loci, leading to changes in allele frequencies

at many sites across the genome. Among these studies, the most commonly inferred complex traits are pigmentation and standing height.

ANCIENT DNA AND COMPLEX TRAIT GENOMICS

Skin, hair and eye pigmentation are among the least polygenic complex traits; though more than a hundred pigmentation-associated loci have been found, their heritability is largely dominated by large-effect common SNPs (Sulem et al., 2007; Eiberg et al., 2008; Han et al., 2008; Sturm et al., 2008; Hider et al., 2013; Liu et al., 2015; O'Connor et al., 2019). Additionally, several of these variants have signatures of past selective sweeps detectable in present-day genomes (Lao et al., 2007; Sabeti et al., 2007; Pickrell et al., 2009; Rocha, 2020). Nevertheless, genomic analyses in previously understudied populations—like sub-Saharan African groups—suggest that perhaps hundreds of skin pigmentation alleles of small effect remain to be found (Martin et al., 2017b). Similarly, recent studies have shown that eye pigmentation is far more polygenic than previous thought (Simcoe et al., 2021). Recent quantitative and molecular genomic studies are painting an increasingly complex picture of the architecture of these traits, featuring more considerable roles for epistasis, pleiotropy and small-effect variants than were previously assumed (for an extensive review of skin pigmentation, see Quillen et al., 2019).

Recently, ancient DNA (aDNA) studies have attempted to reconstruct pigmentation phenotypes in ancient human populations, although the extent to which these predictions are accurate remains uncertain. These reconstructions have been mostly focused on ancient individuals from Western Eurasia, due to the relatively higher abundance of SNP-phenotype associations from European-centric studies, and the poor portability of gene-trait associations to more distantly related populations (Martin et al., 2017a, 2019). For example, Olalde et al. (2014) queried pigmentation-associated SNPs in genomes of Mesolithic hunter-gatherer remains from western and central Eurasia, and suggested that the lighter skin colour characteristic of Europeans today was not widely present in the continent before the Neolithic. González-Fortes et al. (2017) analysed Mesolithic and Eneolithic genomes from central Europe, and inferred dark hair, brown eyes and dark skin pigmentation for the Mesolithic individuals and dark hair, light eyes, and lighter skin pigmentation for an Eneolithic individual. Similarly, Brace et al. (2019) inferred pigmentation phenotypes for Mesolithic and Neolithic genomes from western Europe, and reported that the so-called "Cheddar Man," a Mesolithic individual found in England, had blue/green eyes and dark to black skin, in contrast to later Neolithic individuals with dark to intermediate skin pigmentation. Contrastingly, Günther et al. (2018) found elevated frequencies of light skin pigmentation alleles in individuals from the Scandinavian Mesolithic, suggestive of early environmental adaptation to life at higher latitudes. These reconstructions have also been carried out in individuals with no skeletal remains; for example, Jensen et al. (2019) used pigmentation-associated SNPs to infer the skin, hair and eye colour of a female individual whose DNA was preserved in a piece of birch tar "chewing gum."

Some aDNA studies have sought to systematically investigate how pigmentation-associated variants were introduced and evolved in the European continent. Wilde et al. (2014) was one of the first studies to provide aDNA-based evidence that skin, hair, and eye pigmentation-associated alleles have been under strong positive selection in Europe over the past 5,000 years. The first large-scale population genomic studies (Allentoft et al., 2015; Haak et al., 2015; Mathieson et al., 2015) showed that major effect alleles associated with light eye colour likely rose in frequency in Europe before alleles associated with light skin pigmentation. More recently, Ju and Mathieson (2021) argued that the increase in light skin pigmentation in Europeans was primarily driven by strong selection at a small proportion of pigmentation-associated loci with large effect sizes. When testing for polygenic adaptation using an aggregation of all known pigmentation-associated variants, they did not detect a statistically significant signature of selection.

The other trait that has shared comparable prominence with pigmentation in the aDNA literature is standing height. In contrast to pigmentation, the genetic architecture of height is highly polygenic (Yang et al., 2015; Bycroft et al., 2018; Yengo et al., 2018). The heritability of this trait is dominated by a large number of alleles with small effect sizes, and shows strong evidence for negative selection in present-day populations (O'Connor et al., 2019). Studies of the genetic component of height in ancient populations have shown that ancient West Eurasian populations were, on average, more highly differentiated for this trait than present-day West Eurasian populations, and more so than one would predict from genetic drift alone (Mathieson et al., 2015; Martiniano et al., 2017; Cox et al., 2019). Cox et al. (2019) compared predicted genetic changes in height in ancient populations to inferred height changes estimated via skeletal remains. They concluded that the changes in inferred standing height were partially predicted by genetics; with both measures remaining relatively constant between the Mesolithic and Neolithic, and increasing between the Neolithic and Bronze Age. A follow-up study by Cox et al. (2021) used polygenic scores for height to show that PRS predicts 6.8% of the observed variance in femur length in ancient skeletons, after controlling for other variables. This is approximately one quarter of the predictive accuracy of PRS in present-day populations; which the authors attribute to the low-coverage aDNA data used in their study. Contrastingly, Marciniak et al. (2021) used the discordance between PRS for height, calculated from aDNA, and height inferred from the corresponding skeletal remains, to argue that Neolithic individuals were shorter than expected due to either poorer nutrition or increased disease burden, relative to hunter-gatherer populations.

However, the inference of standing height from skeletal remains is not without its own problems. Both Cox et al. (2021) and Marciniak et al. (2021) used the method developed by Ruff et al. (2012) to estimate stature from skeletal remains. Nevertheless, their respective estimates of stature—based on femur length—varied between some of the individuals included in both studies. Where multiple skeletal elements were available for ancient individuals, Marciniak et al. (2021) also produced

separate stature estimates from femur, tibia, humerus and radius length, which varied substantially within some individuals; highlighting the uncertainty in estimates of stature from skeletal remains.

INFERRING COMPLEX TRAITS IN ARCHAIC HOMINIDS

The availability of genome sequences from archaic humans, like Neanderthals and Denisovans, has greatly expanded our understanding of their demographic history and interactions with modern humans (Meyer et al., 2012; Prüfer et al., 2014, 2017). However, little is known about complex traits in archaic humans, besides what can be inferred directly from their skeletal remains. In the case of Denisovans, such remains are presently limited to a few teeth, a mandible and other small bone fragments, making it difficult to make confident inferences of their biology (Meyer et al., 2012; Sawyer et al., 2015; Slon et al., 2017; Chen et al., 2019). However, past admixture events with archaic human groups have left a genetic legacy in present-day people, providing a possible inroad to study archaic human biology (Sankararaman et al., 2012). Today, around 2% of the genomes of non-African humans are known to be descended from Neandertals, and an additional ∼5% of the genomes of people in Oceania can be traced back to Denisovans (Sankararaman et al., 2014, 2016; Vernot and Akey, 2014; Vernot et al., 2016).

Knowledge about admixture between archaic and modern humans has led to a recent flurry of exploratory studies concerning the potential impact of archaic variants on complex traits in present-day populations. Various approaches have been used to identify introgressed archaic DNA putatively under positive selection in modern humans (Khrameeva et al., 2014; Sankararaman et al., 2014, 2016; Vernot and Akey, 2014; Perry et al., 2015; Gittelman et al., 2016; Vernot et al., 2016; Racimo et al., 2017b). Overall, these studies have shown that archaic DNA is linked to pathways related to metabolism, as well as skin and hair morphology. Via association studies, Neanderthal variants in specific loci have been shown to influence several disease and immune traits, as well as skin and hair colour, behavioural traits, skull shape, pain perception and reproduction (Sankararaman et al., 2014; Dannemann et al., 2016; Sams et al., 2016; Gunz et al., 2019; Skov et al., 2020; Zeberg and Pääbo, 2020, 2021; Zeberg et al., 2020a,b).

Additionally, comparisons between the combined phenotypic effects of Neandertal variants and frequency-matched non-archaic variants have revealed that Neanderthal DNA is over-proportionally associated with neurological and behavioural phenotypes, as well as viral immune responses and type 2 diabetes (Quach et al., 2016; Simonti et al., 2016; Dannemann and Kelso, 2017; Dannemann, 2021). These groups of phenotypes may be linked to environmental factors, such as ultraviolet light exposure, pathogen prevalence and climate, that substantially differed between Africa and Eurasia. It has been suggested that the over-proportional contribution of Neandertal DNA to immunity and behavioural traits in present-day humans might be a reflection of adaptive processes in Neandertals to these environmental differences. In comparison, much less is known about the impact of Denisovan DNA on complex traits, because limited phenotypic data are presently available from present-day populations. However, individual Denisovan-like haplotypes found in high frequencies in some human populations have been associated with high altitude adaptation and fat metabolism (Huerta-Sánchez et al., 2014; Racimo et al., 2017a).

One key limitation to these approaches is that only about 40–50% of the Neandertal genome can be recovered in present-day humans, and therefore discoverable in such analyses (Sankararaman et al., 2014; Vernot and Akey, 2014; Skov et al., 2020). Furthermore, the majority of tested cohorts used for such studies are of European ancestry, which limits analyses to archaic variants present in these populations. This is particularly notable since Neandertal phenotype associations in European and Asian populations have been shown to contain population-specific archaic variants (Dannemann, 2021). It has also been shown that negative selection, soon after admixture, has played an important role in removing some of the missing segments of archaic DNA (Harris and Nielsen, 2016; Juric et al., 2016; Petr et al., 2019). It is therefore possible that missing segments of archaic DNA had strong phenotypic effects. For archaic DNA that does persist in present-day populations, much of it is segregating at low allele frequencies, making it difficult to confidently link it to phenotypic effects.

Furthermore, it remains questionable how transferable any phenotypic associations are between modern and archaic humans, given the difficulties of transferring associations between present-day populations (Martin et al., 2017a; Duncan et al., 2019). All of the above studies have used gene-trait association information from analyses carried out in modern humans. It remains undetermined if the phenotypic effects of archaic DNA in present-day populations are a reliable proxy for phenotypic effects in archaic humans themselves.

Recent studies have also aimed to predict the phenotypic effects of archaic DNA without relying on introgression in present-day populations (see **Figure 1**). Colbran et al. (2019) used a machine learning algorithm, trained on genetic variation in present-day humans, to infer putative regulatory effects on variation present only in Neandertal genomes. Gokhman et al. (2020a,b) used aDNA damage patterns to infer a DNA methylation map of the Denisovan genome, and linked the inferred regulatory patterns to loss-of-function phenotypes, in order to predict their skeletal morphology and vocal and facial anatomy. It remains to be seen how successful these approaches are at predicting archaic human phenotypes. A possible inroad into validation could rest on functional assays for testing and evaluating the phenotypic impact of archaic DNA (Dannemann et al., 2020; Dannemann and Gallego Romero, 2021; Trujillo et al., 2021).

THE CHALLENGE OF DETECTING POLYGENIC ADAPTATION IN ANCIENT POPULATIONS

Perhaps the most fascinating question about the evolution of complex traits in humans is whether they were subject to natural selection. Current methods to detect polygenic

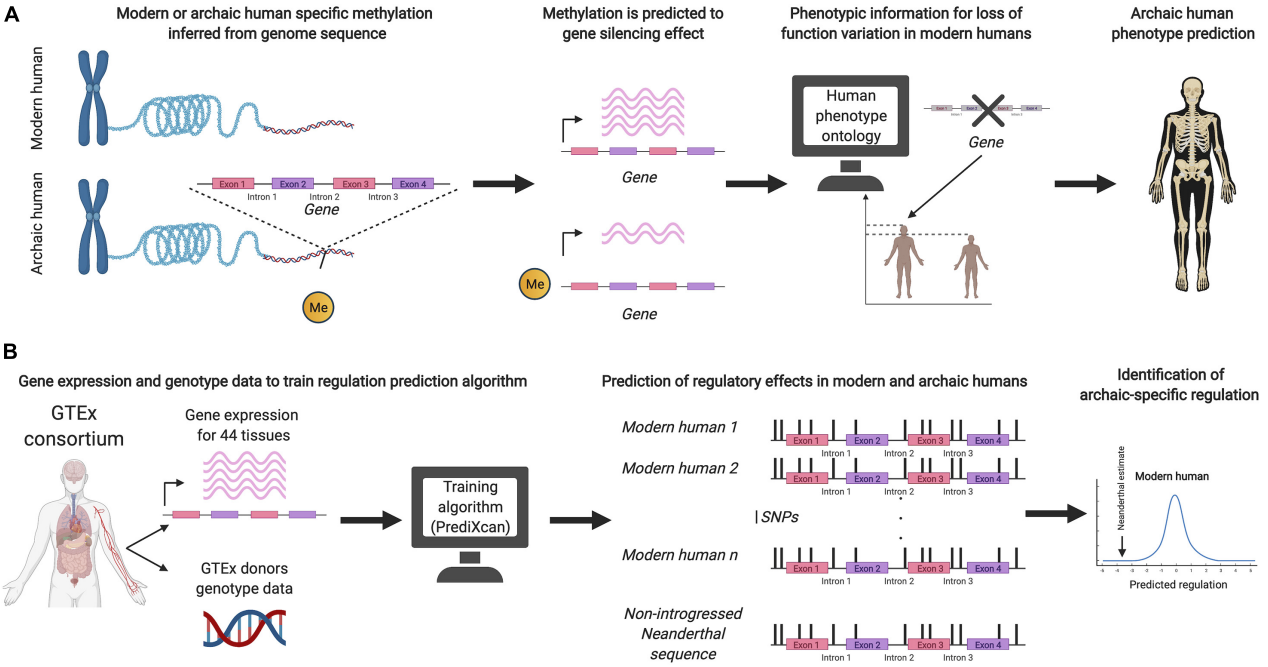

FIGURE 1 | (A) Schematic illustration of the prediction method used by Gokhman et al. (2020a) to infer archaic human phenotypes based on methylation maps. **(B)** Schematic illustration of the method by Colbran et al. (2019) to predict regulatory effects of non-introgressed archaic human DNA.

adaptation have mainly focused on present-day populations; using either differences between populations, or variation within them, to identify polygenic adaptation. For example, Berg and Coop (2014) developed a method that identifies over-dispersion of genetic values among populations, compared to a null distribution expected under a model of drift; which Racimo et al. (2018) extended to work with admixture graphs. Field et al. (2016) used the distribution of singletons around trait-associated SNPs, and Uricchio et al. (2019) used the joint distribution of variant effect sizes and derived allele frequencies (DAF). Whichever method is used, significant caveats must be addressed before attributing differences in such scores to polygenic adaptation (Novembre and Barton, 2018; Coop, 2019; Rosenberg et al., 2019). Most of these issues affect both present-day and ancient populations, but many are especially problematic when working with ancient genomes.

A prominently reported example of polygenic adaptation is that of selection for increasing height across a north-south gradient in Europe (Turchin et al., 2012; Berg and Coop, 2014; Robinson et al., 2015; Zoledziewska et al., 2015; Guo et al., 2018; Racimo et al., 2018; Berg et al., 2019b; Chen et al., 2020). Most studies which described this signal based their analyses on effect size estimates from the GIANT consortium, a GWAS meta-analysis encompassing 79 separate studies (Wood et al., 2014). Concerningly, follow-up work using the larger and more homogeneous UK Biobank cohort failed to replicate the signal of polygenic adaptation for height (Berg et al., 2019a; Sohail et al., 2019). A recent systematic comparison across a range of GWAS cohorts has further shown that the results of these tests are highly dependent on the ancestry composition of the cohort used to obtain the effect size estimates (Refoyo-Martínez et al., 2021). These analyses showed that residual stratification in GWAS meta- and mega-analyses can result in inflated effect size estimates that, in turn, can lead to spurious signals of selection. The effects of this residual stratification may be exacerbated for ancient populations with non-uniform relatedness to present-day GWAS cohorts (see **Figure 2**).

Residual stratification is a major concern for GWAS, even among a relatively homogeneous cohort like the UK Biobank. Zaidi and Mathieson (2020) used simulations to show that fine-scale recent demography can confound GWAS which has been corrected for stratification using common variants only. Failure to adequately correct for localised population structure can lead to spurious associations between a trait and low-frequency variants that happen to be common in areas of atypical environmental effect. This finding is problematic as most GWAS have been conducted on either SNP array data, or on genomes imputed from SNP array data (Visscher et al., 2017). For example, GWAS summary statistics from the UK Biobank are based on imputed genomes (Bycroft et al., 2018). A limitation of this approach is that the accuracy of imputed genotypes are inversely correlated with the minor allele frequencies (MAF) of variants in the reference panel. Additionally, rare variants that are not segregating in the reference panel cannot be imputed at all. As a result, imputed genomes are specifically depleted in the rare variants needed to adjust for stratification from recent demography.

For large sample sizes, low-frequency variants (MAF ≤ 0.05) make a significant contribution to the heritability of many complex traits (Mancuso et al., 2016; Hartman et al., 2019), but

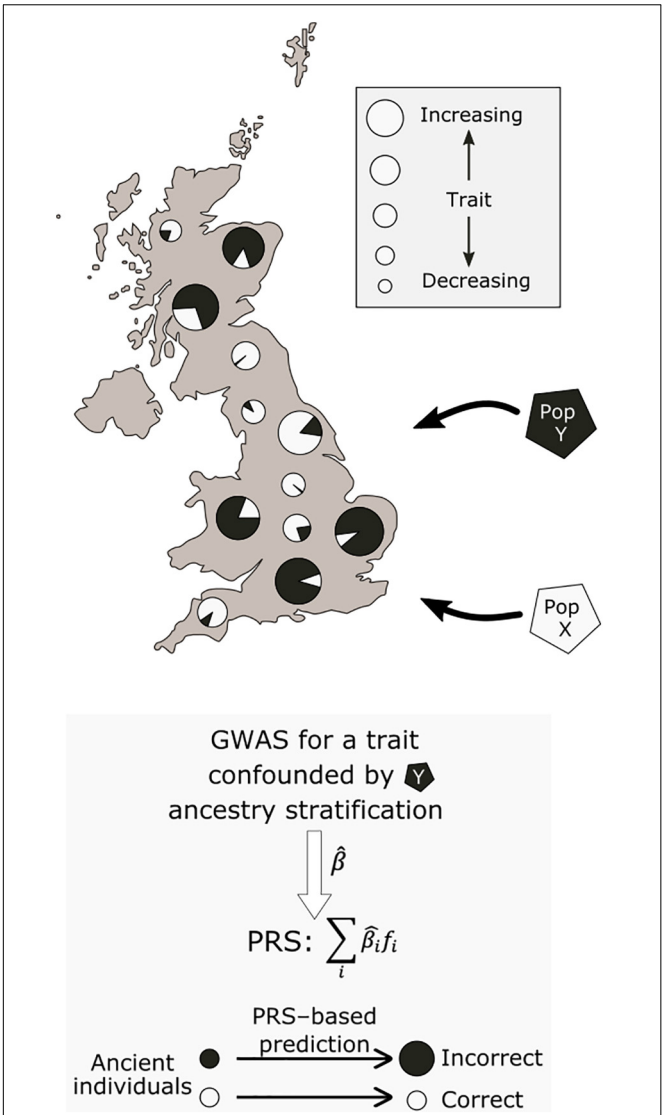

FIGURE 2 | Potential effect of population stratification in GWAS. Two ancestral populations, X and Y, have contributed differing ancestry proportions to present-day individuals. Due to non-genetic environmental effects, individuals with a larger proportion of population Y ancestry have higher values for a measured trait. This may lead to biased GWAS effect size estimates, which associate population Y ancestry with increasing values of the trait. When used to make inferences about the past, this would lead to systematically inflated polygenic scores for this trait in samples from population Y.

the role of rare variants is less well established. Both empirical and simulation studies have shown that for traits under either negative or stabilising selection, there is an inverse correlation between effect size and MAF (Simons et al., 2018; Schoech et al., 2019; Durvasula and Lohmueller, 2021). For the many traits thought to be under negative selection (O'Connor et al., 2019), large effect variants that are rare in present-day populations may have had higher allele frequencies in ancient populations due to selection. This makes polygenic scores for ancient individuals especially sensitive to bias from GWAS effect size estimates ascertained from common variants only. Conversely, where present-day rare variants with large-effect sizes are known, higher frequencies in ancient populations would result in more accurate PRS predictions, due to their larger contribution to the overall genetic variance.

A recent analysis indicated that a substantial component of the unidentified heritability for anthropometric traits like height and BMI lies within large effect rare variants, some with MAF as low as 0.01% (Wainschtein et al., 2019). However, using GWAS to recover variant associations for SNPs as rare as this would require hundreds of thousands of whole-genomes, substantially exceeding the largest whole-genome GWAS published to date (e.g., Taliun et al., 2021). The consequence of this missing heritability may be particularly acute for trait prediction in ancient samples, as large-effect rare variants which contributed to variability in the past may no longer be segregating in present-day populations. Indeed, simulations suggest that the genetic architecture of complex traits is highly specific to each population, and that negative selection enriches for private variants, which contribute to a substantial component of the heritability of each trait (Durvasula and Lohmueller, 2021). Empirical studies have also identified that functionally important regions, including conserved and regulatory regions, are enriched for population-specific effect sizes, and that this pattern may have been driven by directional selection (Shi et al., 2021).

In addition to these issues, the majority of SNP associations inferred from GWAS are likely not the causal alleles. Instead, GWAS predominantly identifies SNPs which are in high linkage disequilibrium (LD) with causal alleles. Most GWAS also assume a model in which all complex trait heritability is additive and well tagged by SNPs segregating in the cohort; although some GWAS do include non-additive models (e.g., Guindo-Martínez et al., 2021). Consequently, effect size estimates are contingent on the LD structure of the cohort in which they were ascertained. Due to recombination, this LD structure decays through time, and is reshaped by the population history in which selection processes are embedded.

Over the last decade, paleogenomic studies have repeatedly demonstrated that the evolutionary histories of human populations are characterized by recurrent episodes of divergence, expansion, migration and admixture (reviewed in Pickrell and Reich, 2014; Skoglund and Mathieson, 2018). For example, in West Eurasia, four major ancestry groups have contributed to the majority of present-day genetic variation (Jones et al., 2015). As such, the LD structure of present-day British individuals—which underpins effect size estimates from the UK Biobank—was substantially different prior to the Bronze Age, when the most recent of these major admixture episodes occurred (Allentoft et al., 2015; Haak et al., 2015). To improve ancestral trait prediction, new methods which explicitly model the haplotype structure of both ancient populations and present-day GWAS cohorts are needed.

In aggregate, these issues combine to substantially diminish the portability of polygenic scores between populations. Indeed, in present-day populations, the predictive accuracy of PRS degrades approximately linearly with increasing genetic distance from the cohort used to ascertain the GWAS (Scutari et al., 2016;

Martin et al., 2017a, 2019; Kim et al., 2018; Bitarello and Mathieson, 2020; Mostafavi et al., 2020; Majara et al., 2021). Even within a single ancestry group, the correlation between PRS calculated from different discovery GWAS shows considerable variance (Schultz et al., 2021). However, the extent to which the issue of PRS portability also affects ancient populations, which are either partially or directly ancestral to the GWAS cohort, are yet to be determined.

In cases where a robust signal of polygenic adaptation can be identified, care must still be taken when interpreting which trait was actually subject to directional selection. Due to the highly polygenic nature of most complex traits, there is a high rate of genetic correlation between phenotypes (Shi et al., 2017; Ning et al., 2020). This can occur when correlated traits share causal alleles (i.e., pleiotropy) or where casual alleles are in high LD with each other. Consequently, selection acting on one specific trait can generate a spurious signal of polygenic adaptation for multiple genetically correlated traits. Recently, Stern et al. (2021) developed a method for conditional testing of polygenic adaptation to address this problem. When considered in a joint test, previously identified signals of selection for educational attainment and hair colour in British individuals were significantly attenuated by the signal of selection for skin pigmentation (Stern et al., 2021). However, this approach can only untangle genetic correlations between traits which have been measured in GWAS cohorts, leaving open the possibility that selection is acting on an unobserved yet correlated trait. Indeed, many GWAS traits are either coarse proxy measures with substantial socio-economic confounding (e.g., educational attainment), or narrow physiological measurements (e.g., levels of potassium in urine); neither of which are likely to have been direct targets of polygenic adaptation. In practice, the truly adaptive phenotype is rarely directly observable, and all measured traits are genetically correlated proxies at various levels of abstraction.

LIMITATIONS AND CAVEATS SPECIFIC TO ANCIENT DNA

In addition to all of the general issues and caveats discussed above, working with ancient DNA also involves a range of issues that are particular to the degraded nature of the data; such as post-mortem damage, generally low average sequence coverage, short fragment lengths, reference bias, and microbial and human contamination (Gilbert et al., 2005; Dabney et al., 2013; Renaud et al., 2019; Peyrégne and Prüfer, 2020). All of these factors affect our ability to correctly infer ancient genotypes; and therefore, to construct accurate polygenic scores or infer polygenic adaptation.

A common strategy for dealing with the low endogenous fraction of aDNA libraries is to use in-solution hybridisation capture to retrieve specific loci, or a set of predetermined SNPs (Avila-Arcos et al., 2011; Cruz-Dávalos et al., 2017). This approach has substantial advantages in on-target efficiency, at the cost of ascertainment bias. For example, in the case of the popular "1240k" capture array, targeted SNPs were predominantly ascertained in present-day individuals (Fu et al., 2015; Haak et al., 2015). Consequently, an unknown fraction of the true ancestral variation is lost during capture. This is further exacerbated by the generally low coverage of most aDNA libraries; for which a common practice is to draw a read at random along each position in the genome, to infer "pseudo-haploid" genotypes. When used to compute polygenic scores for ancient populations, only a subset of GWAS variants can be used, which substantially reduces predictive accuracy. Cox et al. (2021) estimate that the combined effect of low-coverage and pseudo-haploid genotypes reduced their predictive accuracy by approximately 75%, when compared to present-day data.

An alternative approach is to perform low-coverage shotgun sequencing, followed by imputation, using a large reference panel (Ausmees et al., 2019; Hui et al., 2020). This has the dual advantages of reducing ascertainment bias and increasing the number of GWAS variants available to calculate polygenic scores. However, imputation itself introduces a new source of bias, particularly if the reference panel is not representative of the ancestries found in the low-coverage samples. Nevertheless, the level of imputation bias can be empirically estimated by downsampling high-coverage aDNA libraries and testing imputed genotypes against direct observations (e.g., Margaryan et al., 2020). Where a suitable reference panel exists, recently developed methods for imputation from low-coverage sequencing data (Davies et al., 2021; Rubinacci et al., 2021) show great promise for ancient DNA studies (e.g., Clemente et al., 2021).

Even under ideal conditions, in which exact polygenic scores for ancient populations are known *a priori*, interpreting differences in mean PRS between groups still requires careful consideration. For many polygenic traits, the variance between population means is lower than the variance within populations. As a result, differences in population level polygenic scores have limited predictive value for inferring the physiology or behaviour of individual people in the past. Genetics plays only a partial role in shaping phenotypic diversity, and differences in polygenic scores between individuals, or populations, does not automatically translate into differences in the expressed phenotype. Indeed, for some complex traits, an inverse correlation has been observed; in which polygenic scores have been steadily decreasing over recent decades, whilst the measured phenotype has been increasing [e.g., educational attainment (Kong et al., 2017; Abdellaoui et al., 2019)]. This highlights the substantial role of environmental variation in shaping phenotypic diversity. For ancient populations, we must also consider the wide variation in culture, diet, health, social organisation and climate which will have mediated any potential differences in population level polygenic scores. Furthermore, ancient populations are likely to have experienced a heterogeneous range of selective pressures. What we observe in present-day populations is not the result of a single directional process, but instead represents a mosaic of haplotypes which were shaped by different fitness landscapes, at varying levels of temporal depth.

Lastly, in most cases, we cannot directly observe phenotypes in the ancient individuals whose genomes have been studied. This greatly limits our ability to compare the genetically predicted value of a trait to its expressed phenotype, raising the

question: are predictions of most ancient phenotypes inherently unverifiable? For well-preserved traits, like standing height, there is considerable variability in estimates produced from different skeletal elements and between different studies (Cox et al., 2021; Marciniak et al., 2021). For traits that do not preserve well in the archaeological record, the prospects of validation are much poorer. These include not only soft tissue measurements (e.g., pigmentation or haemoglobin counts), but also personality and mental health traits that require an individual to be alive to be properly measured or diagnosed. Furthermore, some phenotypes are non-sensical outside of a modern context. Whilst it is possible to build a polygenic score for "time spent watching television" (UK Biobank code: 1070), it is not clear how to interpret any potential differences one might find between Mesolithic hunter-gatherers and Neolithic farmers. This problem extends more generally to all phenotypes which have strong gene–environment interactions, in which the expression of the trait may have been substantively different in the past due to diverse environmental conditions (e.g., the interplay between BMI and diet).

PROSPECTS FOR THE FUTURE

The growth in the number of ancient genomes currently shows little signs of slowing, nor does the increasing availability of gene-trait association data. Predictably, efforts to perform trait predictions in ancient individuals will also continue to grow. We believe that increased emphasis on limitations and caveats in the way we study and communicate these findings will enable a better understanding of what we can and cannot predict with existing models.

As a working assumption, polygenic scores from any single GWAS should be considered unreliable in an ancient trait reconstruction analysis. Researchers should only trust observed signals of trait evolution if those patterns hold across multiple independent GWAS (e.g., Chen et al., 2020), and preferably where each of these GWAS has been performed on a large cohort with homogeneous ancestry (Refoyo-Martínez et al., 2021).

We also need to better understand how well GWAS effect size estimates, ascertained in present-day populations, generalise to ancient populations that are only partially ancestral to the GWAS cohort. One approach to this would be to use simulations, under a plausible demographic scenario, to explore how the predictive accuracy of PRS degrades through time and across the boundaries of major ancestral migrations.

Traits that are preserved in the fossil record can provide a degree of partial benchmarking (Cox et al., 2019, 2021); however, the genetic components of variation are often only partially explained by polygenic scores, and environmental components almost always play large roles in expressed trait variation, often dwarfing the contribution of polygenic scores. Furthermore, only a few—largely osteological—traits are well preserved over time, so these comparisons will always be limited in scope.

That being said, there are several promising avenues of research that could serve to improve genetic trait prediction in ancient populations. An existing approach to improve the portability of PRS across ancestries is to prioritise variants with predicted functional roles (Amariuta et al., 2020; Weissbrod et al., 2020). This approach aims to improve PRS portability in present-day populations by reducing the fraction of spurious associations due to the cohort specific LD structure of the GWAS reference panel. Another promising approach is to jointly model PRS using GWAS summary statistics from multiple populations (Márquez-Luna et al., 2017; Ruan et al., 2021; Turley et al., 2021). By including information from genetically distant groups, these methods can account for the variance in effect sizes inferred between GWAS cohorts. This multi-ancestry approach holds particular promise for ancient populations, as it may help to identify variant associations which are segregating in only a subset of present-day populations, but which were more widespread in the past.

These studies also underscore the importance of studying the ancestral haplotype backgrounds on which beneficial, deleterious or neutral alleles spread. Recent studies have shown that tests of selection on individual loci can gain power by explicitly modelling patterns of ancestry across the genome (Pierron et al., 2018; Hamid et al., 2021). Strong selective signals might be masked by post-selection admixture processes, but might become evident once the ancestry of the selected haplotypes is explicitly modelled (Souilmi et al., 2020). This phenomenon is also likely to affect polygenic adaptation studies, particularly when the degree of correlation between genetic score differences and differences in ancestral haplotype backgrounds is expected to be high, for example, after admixture between populations that have been evolving in isolation for long periods of time.

A promising avenue of research is developing around new methods for approximately inferring ancestral recombination graphs (ARG) via the construction of tree sequences (Kelleher et al., 2019; Speidel et al., 2019), which have recently been extended to incorporate non-contemporaneous sampling (Speidel et al., 2021; Wohns et al., 2021). An ARG is a model which contains a detailed description of the genealogical relationships in a set of samples, including the full history of gene trees, ancestral haplotypes and recombination events which relate the samples to each other at every site in the genome (Griffiths and Marjoram, 1997). One potential advantage of an ARG is that it may be used to help mitigate issues with the portability of polygenic scores. By building an ARG composed of both ancient samples and the present-day cohorts used to ascertain the GWAS associations, one could potentially determine which haplotypes are shared between the GWAS cohort and the ancient populations; thereby reducing effect size bias in populations that are only partially ancestral to the GWAS cohort.

Another area in which ancient genomes offer unique potential is in detecting polygenic adaptation in response to environmental change. The time-series nature of ancient genomes provides the potential for the incorporation of paleoclimate reconstructions (e.g., Brown et al., 2018) into tests of polygenic adaptation, in a manner that is not possible with present-day data alone.

Ultimately, the ancient genomics community must come to terms with the limitations of genetic hindcasting. Ancient

genomes provide an unprecedented window into our past, but this window is often blurry and distorted. There is still a lot of information waiting to be obtained from ancient DNA, and some of the blurriness might ultimately come into focus as computational methods continue to improve. But we must also accept the fact that many aspects of past human biology—including physical characteristics and disease susceptibility—might be irrevocably lost to the tides of history. Ancient genome sequences are, after all, molecular fossils: imperfect and degraded records of lives that ceased to exist long ago.

AUTHOR CONTRIBUTIONS

EI-P and FR reviewed and edited the manuscript. All authors wrote the original draft of the manuscript and approved the submitted version.

FUNDING

EI-P was supported by the Lundbeck Foundation (grant R302-2018-2155) and the Novo Nordisk Foundation (grant NNF18SA0035006). FR and RM were supported by a Villum Fonden Young Investigator award to FR (project no. 00025300). Additionally, FR was supported by the COREX ERC Synergy grant (ID 951385). MD was supported by the European Union through the Horizon 2020 Research and Innovation Programme under grant no. 810645 and the European Regional Development Fund Project No. MOBEC008.

ACKNOWLEDGMENTS

We thank the members of the Racimo group for their helpful advice and discussions, and thank the reviewers and editor for their constructive feedback. **Figure 1** was created with Biorender.com.

REFERENCES

Abdellaoui, A., Hugh-Jones, D., Yengo, L., Kemper, K. E., Nivard, M. G., Veul, L., et al. (2019). Genetic correlates of social stratification in Great Britain. *Nat. Hum. Behav.* 3, 1332–1342. doi: 10.1038/s41562-019-0757-5

Allentoft, M. E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., et al. (2015). Population genomics of Bronze Age Eurasia. *Nature* 522, 167–172. doi: 10.1038/nature14507

Amariuta, T., Ishigaki, K., Sugishita, H., Ohta, T., Koido, M., Dey, K. K., et al. (2020). Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. *Nat. Genet.* 52, 1346–1354. doi: 10.1038/s41588-020-00740-8

Ausmees, K., Sanchez-Quinto, F., Jakobsson, M., and Nettelblad, C. (2019). *An empirical evaluation of genotype imputation of ancient DNA*. Available Online at: http://uu.diva-portal.org/smash/record.jsf?pid=diva2%3A1367434&dswid=5303 [Accessed January 16, 2021]

Avila-Arcos, M. C., Cappellini, E., Romero-Navarro, J. A., Wales, N., Moreno-Mayar, J. V., Rasmussen, M., et al. (2011). Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. *Sci. Rep.* 1:74. doi: 10.1038/srep00074

Berg, J. J., and Coop, G. (2014). A population genetic signal of polygenic adaptation. *PLoS Genet.* 10:e1004412. doi: 10.1371/journal.pgen.1004412

Berg, J. J., Harpak, A., Sinnott-Armstrong, N., Joergensen, A. M., Mostafavi, H., Field, Y., et al. (2019a). Reduced signal for polygenic adaptation of height in UK Biobank. *Elife* 8:47. doi: 10.7554/eLife.39725

Berg, J. J., Zhang, X., and Coop, G. (2019b). Polygenic adaptation has impacted multiple anthropometric traits. *bioRxiv* 2019:167551. doi: 10.1101/167551

Bitarello, B. D., and Mathieson, I. (2020). Polygenic scores for height in admixed populations. *G3* 10, 4027–4036. doi: 10.1534/g3.120.401658

Brace, S., Diekmann, Y., Booth, T. J., van Dorp, L., Faltyskova, Z., Rohland, N., et al. (2019). Ancient genomes indicate population replacement in Early Neolithic Britain. *Nat. Ecol. Evol.* 3, 765–771. doi: 10.1038/s41559-019-0871-9

Brown, J. L., Hill, D. J., Dolan, A. M., Carnaval, A. C., and Haywood, A. M. (2018). PaleoClim, high spatial resolution paleoclimate surfaces for global land areas. *Sci. Data* 5:180254. doi: 10.1038/sdata.2018.254

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47, D1005–D1012. doi: 10.1093/nar/gky1120

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203–209. doi: 10.1038/s41586-018-0579-z

Chen, F., Welker, F., Shen, C.-C., Bailey, S. E., Bergmann, I., Davis, S., et al. (2019). A late middle pleistocene denisovan mandible from the Tibetan Plateau. *Nature* 569, 409–412. doi: 10.1038/s41586-019-1139-x

Chen, M., Sidore, C., Akiyama, M., Ishigaki, K., Kamatani, Y., Schlessinger, D., et al. (2020). Evidence of polygenic adaptation in sardinia at height-associated loci ascertained from the Biobank Japan. *Am. J. Hum. Genet.* 107, 60–71. doi: 10.1016/j.ajhg.2020.05.014

Clemente, F., Unterländer, M., Dolgova, O., Amorim, C. E. G., Coroado-Santos, F., Neuenschwander, S., et al. (2021). The genomic history of the Aegean palatial civilizations. *Cell* 184, 2565–2586. doi: 10.1016/j.cell.2021.03.039

Colbran, L. L., Gamazon, E. R., Zhou, D., Evans, P., Cox, N. J., and Capra, J. A. (2019). Inferred divergent gene regulation in archaic hominins reveals potential phenotypic differences. *Nat. Ecol. Evol.* 3, 1598–1606. doi: 10.1038/s41559-019-0996-x

Coop, G. (2019). *Reading tea leaves? Polygenic scores and differences in traits among groups*. Available Online at: http://arxiv.org/abs/1909.00892

Cox, S. L., Moots, H., Stock, J. T., Shbat, A., Bitarello, B. D., Haak, W., et al. (2021). Predicting skeletal stature using ancient DNA. *bioRxiv* 2021:437877. doi: 10.1101/2021.03.31.437877

Cox, S. L., Ruff, C. B., Maier, R. M., and Mathieson, I. (2019). Genetic contributions to variation in human stature in prehistoric Europe. *Proc. Natl. Acad. Sci. U. S. A.* 116, 21484–21492. doi: 10.1073/pnas.1910606116

Cruz-Dávalos, D. I., Llamas, B., Gaunitz, C., Fages, A., Gamba, C., Soubrier, J., et al. (2017). Experimental conditions improving in-solution target enrichment for ancient DNA. *Mol. Ecol. Resour.* 17, 508–522. doi: 10.1111/1755-0998.12595

Dabney, J., Meyer, M., and Pääbo, S. (2013). Ancient DNA damage. *Cold Spring Harb. Perspect. Biol.* 5:a012567. doi: 10.1101/cshperspect.a012567

Dannemann, M. (2021). The population-specific impact of Neandertal introgression on human disease. *Genome Biol. Evol.* 13:evaa250. doi: 10.1093/gbe/evaa250

Dannemann, M., Andrés, A. M., and Kelso, J. (2016). Introgression of Neandertal- and Denisovan-like haplotypes contributes to adaptive variation in human Toll-like receptors. *Am. J. Hum. Genet.* 98, 22–33. doi: 10.1016/j.ajhg.2015.11.015

Dannemann, M., and Gallego Romero, I. (2021). Harnessing pluripotent stem cells as models to decipher human evolution. *FEBS J.* 2021:15885. doi: 10.1111/febs.15885

Dannemann, M., He, Z., Heide, C., Vernot, B., Sidow, L., Kanton, S., et al. (2020). Human stem cell resources are an inroad to neandertal DNA functions. *Stem Cell Rep.* 15, 214–225. doi: 10.1016/j.stemcr.2020.05.018

Dannemann, M., and Kelso, J. (2017). The contribution of neanderthals to phenotypic variation in modern humans. *Am. J. Hum. Genet.* 101, 578–589. doi: 10.1016/j.ajhg.2017.09.010

Davies, R. W., Kucka, M., Su, D., Shi, S., Flanagan, M., Cunniff, C. M., et al. (2021). Rapid genotype imputation from sequence with reference panels. *Nat. Genet.* 53, 1–8. doi: 10.1038/s41588-021-00877-0

Dehasque, M., Ávila-Arcos, M. C., Díez-Del-Molino, D., Fumagalli, M., Guschanski, K., Lorenzen, E. D., et al. (2020). Inference of natural selection from ancient DNA. *Evol. Lett.* 4, 94–108. doi: 10.1002/evl3.165

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* 9:e1003348. doi: 10.1371/journal.pgen.1003348

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., et al. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* 10, 1–9. doi: 10.1038/s41467-019-11112-0

Durvasula, A., and Lohmueller, K. E. (2021). Negative selection on complex traits limits phenotype prediction accuracy between populations. *Am. J. Hum. Genet.* 108, 620–631. doi: 10.1016/j.ajhg.2021.02.013

Eiberg, H., Troelsen, J., Nielsen, M., Mikkelsen, A., Mengel-From, J., Kjaer, K. W., et al. (2008). Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Hum. Genet.* 123, 177–187. doi: 10.1007/s00439-007-0460-x

Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., Golan, D., et al. (2016). Detection of human adaptation during the past 2000 years. *Science* 354, 760–764. doi: 10.1126/science.aag0776

Fu, Q., Hajdinjak, M., Moldovan, O. T., Constantin, S., Mallick, S., Skoglund, P., et al. (2015). An early modern human from Romania with a recent Neanderthal ancestor. *Nature* 524, 216–219. doi: 10.1038/nature14558

Gilbert, M. T. P., Bandelt, H.-J., Hofreiter, M., and Barnes, I. (2005). Assessing ancient DNA studies. *Trends Ecol. Evol.* 20, 541–544. doi: 10.1016/j.tree.2005.07.005

Gittelman, R. M., Schraiber, J. G., Vernot, B., Mikacenic, C., Wurfel, M. M., and Akey, J. M. (2016). Archaic hominin admixture facilitated adaptation to Out-of-Africa environments. *Curr. Biol.* 26, 3375–3382. doi: 10.1016/j.cub.2016.10.041

Gokhman, D., Mishol, N., de Manuel, M., de Juan, D., Shuqrun, J., Meshorer, E., et al. (2020a). Reconstructing denisovan anatomy using DNA methylation maps. *Cell* 180:601. doi: 10.1016/j.cell.2020.01.020

Gokhman, D., Nissim-Rafinia, M., Agranat-Tamir, L., Housman, G., García-Pérez, R., Lizano, E., et al. (2020b). Differential DNA methylation of vocal and facial anatomy genes in modern humans. *Nat. Commun.* 11:1189. doi: 10.1038/s41467-020-15020-6

González-Fortes, G., Jones, E. R., Lightfoot, E., Bonsall, C., Lazar, C., Grandal-d'Anglade, A., et al. (2017). Paleogenomic evidence for multi-generational mixing between neolithic farmers and mesolithic hunter-gatherers in the lower danube basin. *Curr. Biol.* 27, 1801–1810. doi: 10.1016/j.cub.2017.05.023

Griffiths, R. C., and Marjoram, P. (1997). An ancestral recombination graph. *Instit. Mathemat. Appl.* 87:257.

Guindo-Martínez, M., Amela, R., Bonàs-Guarch, S., Puiggròs, M., Salvoro, C., Miguel-Escalada, I., et al. (2021). The impact of non-additive genetic associations on age-related complex diseases. *Nat. Commun.* 12:2436. doi: 10.1038/s41467-021-21952-4

Günther, T., Malmström, H., Svensson, E. M., Omrak, A., Sánchez-Quinto, F., Kılınç, G. M., et al. (2018). Population genomics of mesolithic scandinavia: Investigating early postglacial migration routes and high-latitude adaptation. *PLoS Biol.* 16:e2003703. doi: 10.1371/journal.pbio.2003703

Gunz, P., Tilot, A. K., Wittfeld, K., Teumer, A., Shapland, C. Y., van Erp, T. G. M., et al. (2019). Neandertal introgression sheds light on modern human endocranial globularity. *Curr. Biol.* 29, 120–127. doi: 10.1016/j.cub.2018.10.065

Guo, J., Wu, Y., Zhu, Z., Zheng, Z., Trzaskowski, M., Zeng, J., et al. (2018). Global genetic differentiation of complex traits shaped by natural selection in humans. *Nat. Commun.* 9:1865. doi: 10.1038/s41467-018-04191-y

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522, 207–211. doi: 10.1038/nature14317

Hamid, I., Korunes, K. L., Beleza, S., and Goldberg, A. (2021). Rapid adaptation to malaria facilitated by admixture in the human population of Cabo Verde. *Elife* 10:e63177. doi: 10.7554/eLife.63177

Han, J., Kraft, P., Nan, H., Guo, Q., Chen, C., Qureshi, A., et al. (2008). A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. *PLoS Genet.* 4:e1000074. doi: 10.1371/journal.pgen.1000074

Harris, K., and Nielsen, R. (2016). The genetic cost of neanderthal introgression. *Genetics* 203, 881–891. doi: 10.1534/genetics.116.186890

Hartman, K. A., Rashkin, S. R., Witte, J. S., and Hernandez, R. D. (2019). Imputed genomic data reveals a moderate effect of low frequency variants to the heritability of complex human traits. *bioRxiv* 2019:879916. doi: 10.1101/2019.12.18.879916

Hider, J. L., Gittelman, R. M., Shah, T., Edwards, M., Rosenbloom, A., Akey, J. M., et al. (2013). Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian ancestry. *BMC Evol. Biol.* 13:150. doi: 10.1186/1471-2148-13-150

Huerta-Sánchez, E., Jin, X., Asan Bianba, Z., Peter, B. M., Vinckenbosch, N., et al. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* 512, 194–197. doi: 10.1038/nature13408

Hui, R., D'Atanasio, E., Cassidy, L. M., Scheib, C. L., and Kivisild, T. (2020). Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. *Sci. Rep.* 10:18542. doi: 10.1038/s41598-020-75387-w

Jensen, T. Z. T., Niemann, J., Iversen, K. H., Fotakis, A. K., Gopalakrishnan, S., Vågene, ÅJ., et al. (2019). A 5700 year-old human genome and oral microbiome from chewed birch pitch. *Nat. Commun.* 10:5520. doi: 10.1038/s41467-019-13549-9

Jones, E. R., Gonzalez-Fortes, G., Connell, S., Siska, V., Eriksson, A., Martiniano, R., et al. (2015). Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat. Commun.* 6:8912. doi: 10.1038/ncomms9912

Ju, D., and Mathieson, I. (2021). The evolution of skin pigmentation-associated variation in West Eurasia. *Proc. Natl. Acad. Sci. U. S. A.* 118:e2009227118. doi: 10.1073/pnas.2009227118

Juric, I., Aeschbacher, S., and Coop, G. (2016). The strength of selection against neanderthal introgression. *PLoS Genet.* 12:e1006340. doi: 10.1371/journal.pgen.1006340

Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., and McVean, G. (2019). Inferring whole-genome histories in large population datasets. *Nat. Genet.* 51, 1330–1338. doi: 10.1038/s41588-019-0483-y

Khrameeva, E. E., Bozek, K., He, L., Yan, Z., Jiang, X., Wei, Y., et al. (2014). Neanderthal ancestry drives evolution of lipid catabolism in contemporary Europeans. *Nat. Commun.* 5:3584. doi: 10.1038/ncomms4584

Kim, M. S., Patel, K. P., Teng, A. K., Berens, A. J., and Lachance, J. (2018). Genetic disease risks can be misestimated across global populations. *Genome Biol.* 19:179. doi: 10.1186/s13059-018-1561-7

Knowles, J. W., and Ashley, E. A. (2018). Cardiovascular disease: The rise of the genetic risk score. *PLoS Med.* 15:e1002546. doi: 10.1371/journal.pmed.1002546

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

Lao, O., de Gruijter, J. M., van Duijn, K., Navarro, A., and Kayser, M. (2007). Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. *Ann. Hum. Genet.* 71, 354–369. doi: 10.1111/j.1469-1809.2006.00341.x

Liu, F., Visser, M., Duffy, D. L., Hysi, P. G., Jacobs, L. C., Lao, O., et al. (2015). Genetics of skin color variation in Europeans: genome-wide association studies with functional follow-up. *Hum. Genet.* 134, 823–835. doi: 10.1007/s00439-015-1559-0

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

Majara, L., Kalungi, A., Koen, N., Zar, H., Stein, D. J., Kinyanda, E., et al. (2021). Low generalizability of polygenic scores in African populations due to genetic and environmental diversity. *bioRxiv* 2021:426453. doi: 10.1101/2021.01.12.426453

Malaspinas, A.-S. (2016). Methods to characterize selective sweeps using time serial samples: an ancient DNA perspective. *Mol. Ecol.* 25, 24–41. doi: 10.1111/mec.13492

Mancuso, N., Rohland, N., Rand, K. A., Tandon, A., Allen, A., Quinque, D., et al. (2016). The contribution of rare variation to prostate cancer heritability. *Nat. Genet.* 48, 30–35. doi: 10.1038/ng.3446

Marciniak, S., Bergey, C. M., Silva, A. M., Hałuszko, A., Furmanek, M., Veselka, B., et al. (2021). An integrative skeletal and paleogenomic analysis of prehistoric stature variation suggests relatively reduced health for early European farmers. *bioRxiv* 2021:437881. doi: 10.1101/2021.03.31.437881

Margaryan, A., Lawson, D. J., Sikora, M., Racimo, F., Rasmussen, S., Moltke, I., et al. (2020). Population genomics of the Viking world. *Nature* 585, 390–396. doi: 10.1038/s41586-020-2688-8

Márquez-Luna, C., Loh, P.-R., South Asian Type 2 Diabetes (SAT2D) Consortium, Sigma Type 2 Diabetes Consortium, and Price, A. L. (2017). Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.* 41, 811–823. doi: 10.1002/gepi.22083

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., et al. (2017a). 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

Martin, A. R., Lin, M., Granka, J. M., Myrick, J. W., Liu, X., Sockell, A., et al. (2017b). An unexpectedly complex architecture for skin pigmentation in Africans. *Cell* 171, 1340–1353. doi: 10.1016/j.cell.2017.11.015

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* 51, 584–591. doi: 10.1038/s41588-019-0379-x

Martiniano, R., Cassidy, L. M., Ó'Maoldúin, R., McLaughlin, R., Silva, N. M., Manco, L., et al. (2017). The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. *PLoS Genet.* 13:e1006852. doi: 10.1371/journal.pgen.1006852

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* 528, 499–503. doi: 10.1038/nature16152

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338, 222–226. doi: 10.1126/science.1224344

Mostafavi, H., Harpak, A., Agarwal, I., Conley, D., Pritchard, J. K., and Przeworski, M. (2020). Variable prediction accuracy of polygenic scores within an ancestry group. *Elife* 9:e48376. doi: 10.7554/eLife.48376

Ning, Z., Pawitan, Y., and Shen, X. (2020). High-definition likelihood inference of genetic correlations across human complex traits. *Nat. Genet.* 52, 859–864. doi: 10.1038/s41588-020-0653-y

Novembre, J., and Barton, N. H. (2018). Tread lightly interpreting polygenic tests of selection. *Genetics* 208, 1351–1355. doi: 10.1534/genetics.118.300786

O'Connor, L. J., Schoech, A. P., Hormozdiari, F., Gazal, S., Patterson, N., and Price, A. L. (2019). Extreme polygenicity of complex traits is explained by negative selection. *Am. J. Hum. Genet.* 105, 456–476. doi: 10.1016/j.ajhg.2019.07.003

Olalde, I., Allentoft, M. E., Sánchez-Quinto, F., Santpere, G., Chiang, C. W. K., DeGiorgio, M., et al. (2014). Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* 507, 225–228. doi: 10.1038/nature12960

Perry, G. H., Kistler, L., Kelaita, M. A., and Sams, A. J. (2015). Insights into hominin phenotypic and dietary evolution from ancient DNA sequence data. *J. Hum. Evol.* 79, 55–63. doi: 10.1016/j.jhevol.2014.10.018

Petr, M., Pääbo, S., Kelso, J., and Vernot, B. (2019). Limits of long-term selection against Neandertal introgression. *Proc. Natl. Acad. Sci. U. S. A.* 116, 1639–1644. doi: 10.1073/pnas.1814338116

Peyrégne, S., and Prüfer, K. (2020). Present-Day DNA contamination in ancient DNA datasets. *BioEssays* 42:2000081. doi: 10.1002/bies.202000081

Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., et al. (2009). Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* 19, 826–837. doi: 10.1101/gr.087577.108

Pickrell, J. K., and Reich, D. (2014). Toward a new history and geography of human genes informed by ancient DNA. *Trends Genet.* 30, 377–389. doi: 10.1016/j.tig.2014.07.007

Pierron, D., Heiske, M., Razafindrazaka, H., Pereda-Loth, V., Sanchez, J., Alva, O., et al. (2018). Strong selection during the last millennium for African ancestry in the admixed population of Madagascar. *Nat. Commun.* 9:932. doi: 10.1038/s41467-018-03342-5

Prüfer, K., de Filippo, C., Grote, S., Mafessoni, F., Korlević, P., Hajdinjak, M., et al. (2017). A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* 358, 655–658. doi: 10.1126/science.aao1887

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505, 43–49. doi: 10.1038/nature12886

Quach, H., Rotival, M., Pothlichet, J., Loh, Y.-H. E., Dannemann, M., Zidane, N., et al. (2016). Genetic adaptation and neandertal admixture shaped the immune system of human populations. *Cell* 167, 643–656. doi: 10.1016/j.cell.2016.09.024

Quillen, E. E., Norton, H. L., Parra, E. J., Lona-Durazo, F., Ang, K. C., Illiescu, F. M., et al. (2019). Shades of complexity: New perspectives on the evolution and genetic architecture of human skin. *Am. J. Phys. Anthropol.* 168, 4–26. doi: 10.1002/ajpa.23737

Racimo, F., Berg, J. J., and Pickrell, J. K. (2018). Detecting polygenic adaptation in admixture graphs. *Genetics* 208, 1565–1584. doi: 10.1534/genetics.117.300489

Racimo, F., Gokhman, D., Fumagalli, M., Ko, A., Hansen, T., Moltke, I., et al. (2017a). Archaic adaptive introgression in TBX15/WARS2. *Mol. Biol. Evol.* 34, 509–524. doi: 10.1093/molbev/msw283

Racimo, F., Marnetto, D., and Huerta-Sánchez, E. (2017b). Signatures of archaic adaptive introgression in present-day human populations. *Mol. Biol. Evol.* 34, 296–317. doi: 10.1093/molbev/msw216

Refoyo-Martínez, A., Liu, S., Jørgensen, A. M., Jin, X., Albrechtsen, A., Martin, A. R., et al. (2021). How robust are cross-population signatures of polygenic adaptation in humans? *bioRxiv* 2020:200030. doi: 10.1101/2020.07.13.200030

Renaud, G., Schubert, M., Sawyer, S., and Orlando, L. (2019). Authentication and assessment of contamination in ancient DNA. *Methods Mol. Biol.* 1963, 163–194. doi: 10.1007/978-1-4939-9176-1_17

Robinson, M. R., Hemani, G., Medina-Gomez, C., Mezzavilla, M., Esko, T., Shakhbazov, K., et al. (2015). Population genetic differentiation of height and body mass index across Europe. *Nat. Genet.* 47, 1357–1362. doi: 10.1038/ng.3401

Rocha, J. (2020). The evolutionary history of human skin pigmentation. *J. Mol. Evol.* 88, 77–87. doi: 10.1007/s00239-019-09902-7

Rosenberg, N. A., Edge, M. D., Pritchard, J. K., and Feldman, M. W. (2019). Interpreting polygenic scores, polygenic adaptation, and human phenotypic differences. *Evol. Med. Public Health* 2019, 26–34. doi: 10.1093/emph/eoy036

Ruan, Y., Anne Feng, Y.-C., Chen, C.-Y., Lam, M., Sawa, A., Martin, A. R., et al. (2021). Improving polygenic prediction in ancestrally diverse populations. *bioRxiv* 2020:20248738. doi: 10.1101/2020.12.27.20248738

Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J., and Delaneau, O. (2021). Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.* 53, 120–126. doi: 10.1038/s41588-020-00756-0

Ruff, C. B., Holt, B. M., Niskanen, M., Sladék, V., Berner, M., Garofalo, E., et al. (2012). Stature and body mass estimation from skeletal remains in the European Holocene. *Am. J. Phys. Anthropol.* 148, 601–617. doi: 10.1002/ajpa.22087

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 913–918. doi: 10.1038/nature06250

Sams, A. J., Dumaine, A., Nédélec, Y., Yotova, V., Alfieri, C., Tanner, J. E., et al. (2016). Adaptively introgressed Neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans. *Genome Biol.* 17:246. doi: 10.1186/s13059-016-1098-6

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., et al. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* 507, 354–357. doi: 10.1038/nature12961

Sankararaman, S., Mallick, S., Patterson, N., and Reich, D. (2016). The combined landscape of denisovan and neanderthal ancestry in present-day humans. *Curr. Biol.* 26, 1241–1247. doi: 10.1016/j.cub.2016.03.037

Sankararaman, S., Patterson, N., Li, H., Pääbo, S., and Reich, D. (2012). The date of interbreeding between neandertals and modern humans. *PLoS Genet.* 8:e1002947. doi: 10.1371/journal.pgen.1002947

Sawyer, S., Renaud, G., Viola, B., Hublin, J.-J., Gansauge, M.-T., Shunkov, M. V., et al. (2015). Nuclear and mitochondrial DNA sequences from two Denisovan individuals. *Proc. Natl. Acad. Sci. U. S. A.* 112, 15696–15700. doi: 10.1073/pnas.1519905112

Schoech, A. P., Jordan, D. M., Loh, P.-R., Gazal, S., O'Connor, L. J., Balick, D. J., et al. (2019). Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. *Nat. Commun.* 10:790. doi: 10.1038/s41467-019-08424-6

Schultz, L. M., Merikangas, A. K., Ruparel, K., Jacquemont, S., Glahn, D. C., Gur, R. E., et al. (2021). Stability of polygenic scores across discovery genome-wide association studies. *bioRxiv* 2021:449060. doi: 10.1101/2021.06.18.449060

Scutari, M., Mackay, I., and Balding, D. (2016). Using genetic distance to infer the accuracy of genomic prediction. *PLoS Genet.* 12:e1006288. doi: 10.1371/journal.pgen.1006288

Shi, H., Gazal, S., Kanai, M., Koch, E. M., Schoech, A. P., Siewert, K. M., et al. (2021). Population-specific causal disease effect sizes in functionally important regions

impacted by selection. *Nat. Commun.* 12:1098. doi: 10.1038/s41467-021-21286-1

Shi, H., Mancuso, N., Spendlove, S., and Pasaniuc, B. (2017). Local genetic correlation gives insights into the shared genetic architecture of complex traits. *Am. J. Hum. Genet.* 101, 737–751. doi: 10.1016/j.ajhg.2017.09.022

Simcoe, M., Valdes, A., Liu, F., Furlotte, N. A., Evans, D. M., Hemani, G., et al. (2021). Genome-wide association study in almost 195,000 individuals identifies 50 previously unidentified genetic loci for eye color. *Sci. Adv.* 7:eabd1239. doi: 10.1126/sciadv.abd1239

Simons, Y. B., Bullaughey, K., Hudson, R. R., and Sella, G. (2018). A population genetic interpretation of GWAS findings for human quantitative traits. *PLoS Biol.* 16:e2002985. doi: 10.1371/journal.pbio.2002985

Simonti, C. N., Vernot, B., Bastarache, L., Bottinger, E., Carrell, D. S., Chisholm, R. L., et al. (2016). The phenotypic legacy of admixture between modern humans and Neandertals. *Science* 351, 737–741. doi: 10.1126/science.aad2149

Skoglund, P., and Mathieson, I. (2018). Ancient genomics of modern humans: The first decade. *Annu. Rev. Genom. Hum. Genet.* 19, 381–404. doi: 10.1146/annurev-genom-083117-021749

Skov, L., Coll Macià, M., Sveinbjörnsson, G., Mafessoni, F., Lucotte, E. A., Einarsdóttir, M. S., et al. (2020). The nature of Neanderthal introgression revealed by 27,566 Icelandic genomes. *Nature* 582, 78–83. doi: 10.1038/s41586-020-2225-9

Slon, V., Viola, B., Renaud, G., Gansauge, M.-T., Benazzi, S., Sawyer, S., et al. (2017). A fourth denisovan individual. *Sci. Adv.* 3:e1700186. doi: 10.1126/sciadv.1700186

Sohail, M., Maier, R. M., Ganna, A., Bloemendal, A., Martin, A. R., Turchin, M. C., et al. (2019). Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *Elife* 8:e39702. doi: 10.7554/eLife.39702

Souilmi, Y., Tobler, R., Johar, A., Williams, M., Grey, S. T., Schmidt, J., et al. (2020). Ancient human genomes reveal a hidden history of strong selection in Eurasia. *bioRxiv* 2020:021006. doi: 10.1101/2020.04.01.021006

Speidel, L., Cassidy, L., Davies, R. W., Hellenthal, G., Skoglund, P., and Myers, S. R. (2021). Inferring population histories for ancient genomes using genome-wide genealogies. *bioRxiv* 2021:431573. doi: 10.1101/2021.02.17.431573

Speidel, L., Forest, M., Shi, S., and Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* 51, 1321–1329. doi: 10.1038/s41588-019-0484-x

Stern, A. J., Speidel, L., Zaitlen, N. A., and Nielsen, R. (2021). Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies. *Am. J. Hum. Genet.* 108, 219–239. doi: 10.1016/j.ajhg.2020.12.005

Sturm, R. A., Duffy, D. L., Zhao, Z. Z., Leite, F. P. N., Stark, M. S., Hayward, N. K., et al. (2008). A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. *Am. J. Hum. Genet.* 82, 424–431. doi: 10.1016/j.ajhg.2007.11.005

Sulem, P., Gudbjartsson, D. F., Stacey, S. N., Helgason, A., Rafnar, T., Magnusson, K. P., et al. (2007). Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat. Genet.* 39, 1443–1452. doi: 10.1038/ng.2007.13

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 590, 290–299. doi: 10.1038/s41586-021-03205-y

Trujillo, C. A., Rice, E. S., Schaefer, N. K., Chaim, I. A., Wheeler, E. C., Madrigal, A. A., et al. (2021). Reintroduction of the archaic variant of NOVA1 in cortical organoids alters neurodevelopment. *Science* 371:eaax2537. doi: 10.1126/science.aax2537

Turchin, M. C., Genetic Investigation of ANthropometric Traits (Giant) Consortium, Chiang, C. W. K., Palmer, C. D., Sankararaman, S., Reich, D., et al. (2012). Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat. Genet.* 44, 1015–1019. doi: 10.1038/ng.2368

Turley, P., Martin, A. R., Goldman, G., Li, H., Kanai, M., Walters, R. K., et al. (2021). Multi-Ancestry Meta-Analysis yields novel genetic discoveries and ancestry-specific associations. *bioRxiv* 2021:441003. doi: 10.1101/2021.04.23.441003

Uricchio, L. H., Kitano, H. C., Gusev, A., and Zaitlen, N. A. (2019). An evolutionary compass for detecting signals of polygenic selection and mutational bias. *Evol. Lett.* 3, 69–79. doi: 10.1002/evl3.97

Vernot, B., and Akey, J. M. (2014). Resurrecting surviving Neandertal lineages from modern human genomes. *Science* 343, 1017–1021. doi: 10.1126/science.1245938

Vernot, B., Tucci, S., Kelso, J., Schraiber, J. G., Wolf, A. B., Gittelman, R. M., et al. (2016). Excavating neandertal and denisovan DNA from the genomes of Melanesian individuals. *Science* 352, 235–239. doi: 10.1126/science.aad9416

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

Wainschtein, P., Jain, D. P., Yengo, L., Zheng, Z., TOPMed Anthropometry Working Group, Trans-Omics for Precision Medicine Consortium, et al. (2019). Recovery of trait heritability from whole genome sequence data. *bioRxiv* 2019:588020. doi: 10.1101/588020

Weissbrod, O., Hormozdiari, F., Benner, C., Cui, R., Ulirsch, J., Gazal, S., et al. (2020). Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* 52, 1355–1363. doi: 10.1038/s41588-020-00735-5

Wilde, S., Timpson, A., Kirsanow, K., Kaiser, E., Kayser, M., Unterländer, M., et al. (2014). Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proc. Natl. Acad. Sci. U. S. A.* 111, 4832–4837. doi: 10.1073/pnas.1316513111

Wohns, A. W., Wong, Y., Jeffery, B., Akbari, A., Mallick, S., Pinhasi, R., et al. (2021). A unified genealogy of modern and ancient genomes. *bioRxiv* 2021:431497. doi: 10.1101/2021.02.16.431497

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

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A. E., Lee, S. H., et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* 47, 1114–1120. doi: 10.1038/ng.3390

Yengo, L., Sidorenko, J., Kemper, K. E., Zheng, Z., Wood, A. R., Weedon, M. N., et al. (2018). Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. *Hum. Mol. Genet.* 27, 3641–3649. doi: 10.1093/hmg/ddy271

Zaidi, A. A., and Mathieson, I. (2020). Demographic history mediates the effect of stratification on polygenic scores. *Elife* 9:e61548. doi: 10.7554/eLife.61548

Zeberg, H., Dannemann, M., Sahlholm, K., Tsuo, K., Maricic, T., Wiebe, V., et al. (2020a). A Neanderthal sodium channel increases pain sensitivity in present-day humans. *Curr. Biol.* 30, 3465–3469. doi: 10.1016/j.cub.2020.06.045

Zeberg, H., Kelso, J., and Pääbo, S. (2020b). The neandertal progesterone receptor. *Mol. Biol. Evol.* 37, 2655–2660. doi: 10.1093/molbev/msaa119

Zeberg, H., and Pääbo, S. (2020). The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature* 587, 610–612. doi: 10.1038/s41586-020-2818-3

Zeberg, H., and Pääbo, S. (2021). A genomic region associated with protection against severe COVID-19 is inherited from Neandertals. *Proc. Natl. Acad. Sci. U. S. A.* 2021:118. doi: 10.1073/pnas.2026309118

Zoledziewska, M., UK10K Consortium, Sidore, C., Chiang, C. W. K., Sanna, S., Mulas, A., et al. (2015). Height-reducing variants and selection for short stature in Sardinia. *Nat. Genet.* 47, 1352–1356. doi: 10.1038/ng.3403

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

MINI REVIEW
published: 10 September 2021
doi: 10.3389/fgene.2021.722602

Current Developments in Detection of Identity-by-Descent Methods and Applications

*Evan L. Sticca[1], Gillian M. Belbin[2] and Christopher R. Gignoux[1]**

[1] Human Medical Genetics and Genomics Program and Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, United States, [2] Institute for Genomic Health, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, United States

Identity-by-descent (IBD), the detection of shared segments inherited from a common ancestor, is a fundamental concept in genomics with broad applications in the characterization and analysis of genomes. While historically the concept of IBD was extensively utilized through linkage analyses and in studies of founder populations, applications of IBD-based methods subsided during the genome-wide association study era. This was primarily due to the computational expense of IBD detection, which becomes increasingly relevant as the field moves toward the analysis of biobank-scale datasets that encompass individuals from highly diverse backgrounds. To address these computational barriers, the past several years have seen new methodological advances enabling IBD detection for datasets in the hundreds of thousands to millions of individuals, enabling novel analyses at an unprecedented scale. Here, we describe the latest innovations in IBD detection and describe opportunities for the application of IBD-based methods across a broad range of questions in the field of genomics.

Keywords: genetics, pedigree, relatedness inference, biobank, identity-by-descent

OPEN ACCESS

Edited by:
Diego Ortega-Del Vecchyo, National Autonomous University of Mexico, Mexico

Reviewed by:
Jazlyn Mooney, Stanford University, United States
Enrique Ambrocio-Ortiz, Instituto Nacional de Enfermedades Respiratorias (INER), Mexico

****Correspondence:***
Christopher R. Gignoux
chris.gignoux@cuanschutz.edu

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *09 June 2021*
Accepted: *24 August 2021*
Published: *10 September 2021*

Citation:
Sticca EL, Belbin GM and Gignoux CR (2021) Current Developments in Detection of Identity-by-Descent Methods and Applications. Front. Genet. 12:722602. doi: 10.3389/fgene.2021.722602

INTRODUCTION

The rapid growth and increasing availability of biobank-scale datasets has led to their increased utilization in human genetics studies, however, the demographic and evolutionary forces that underly genomic patterns within these data are often overlooked. Biases in sample recruitment has led to underrepresentation of non-European ancestry participants, limiting the scope and broad applicability of medical genomics and precision medicine. Additionally, standard genetic analytical frameworks often overlook the fine-scale population structure relevant to the segregation of rare variants, despite their role in common, complex diseases becoming increasingly apparent (Hernandez et al., 2019; Taliun et al., 2021). For these reasons, there is an increasing need for novel methods that can account for demographic substructure driving patterns of variation across the site frequency spectrum in large, diverse cohorts (Gravel et al., 2011). The principle of identity-by-descent (IBD) offers a framework through which we can interpret and leverage the demographic histories of large-scale human genomic data, and improve statistical power to detect causal variants.

Identity-by-descent is the shared inheritance of an identical portion of the genome between two individuals (Browning, 2008; Gusev et al., 2008; Browning and Browning, 2010, Browning and Browning, 2012; Henn et al., 2012; Thompson, 2013). This is distinct from identity-by-state (IBS), in which a portion of two individual's genomes may appear identical, but not necessarily

due to recent shared co-inheritance. Leveraging properties of IBD allows researchers to infer a vast amount of information about a population's demographic history (Carmi et al., 2013; Palamara and Pe'er, 2013; Nait Saada et al., 2020), allowing for evolutionary and pedigree-derived insights that can aid in the interpretation of genetic variation. Further, identifying these shared segments from a recent common ancestor can enrich for shared patterns of rare variation, due to the relationship between allele age and frequency (Slatkin and Rannala, 2000). In essence, inference of IBD sharing at the population level can allow for the same genetic frameworks behind pedigree studies and linkage analyses to be applied to large population-level genotyped or sequenced data sets. In this review, we explore the population genomic principles governing patterns of IBD sharing, past and recent methods for detecting IBD in population scale data, and downstream applications in contemporary human genomics.

GOVERNING EVOLUTIONARY POPULATION GENETICS PRINCIPLES

Methods of IBD detection, or the identification of haplotypes likely to arise from a recent common ancestor are well established in theory but are rarely applied to modern, biobank-scale datasets. These modern algorithms have been shown to have high accuracy and quick computational run times (Ramstetter et al., 2017). The underlying principle is that long haplotypes shared between individuals are statistically more likely to arise from relatedness due to deep, shared population history as opposed to random recombination or mutation (Browning, 2008; Browning and Browning, 2015). The more closely related individuals are, the higher the percentage of their genome will be shared IBD, since they share a common ancestor more recently in their genealogical history than two randomly sampled individuals. As populations both diverge and intermix over time, lengths of IBD segments will degrade due to recombination (Carmi et al., 2013; Palamara and Pe'er, 2013), therefore longer haplotypic segments tend to represent more recent relatedness due to there being a lower probability of recombination inducing a decay in their length over shorter spans of genealogical time (Henn et al., 2012). For a given set of observed genetic data and associated recombination rate estimates, the unknown population history can be modeled by the population genetics principle of the coalescent. This results in an abundance of information that can be inferred from the properties of the shared IBD segments. The length of a shared IBD segment serves as a proxy for age of the most recent common ancestor at that genomic region, i.e., a longer IBD segment reflects a more recent common ancestor. Therefore, by using IBD to measure local relatedness between individuals along the genome, it is possible to infer aspects of a population's demographic history. For instance, factors such as the effective population size over antecedent generations, bottlenecks and subsequent founder effects may be estimated given the distribution of observed IBD in a contemporary population (Browning and Browning, 2015). This has implications at the population level, as represented by patterns of IBD-sharing genome-wide, but can also be informative at specific loci along the genome, and can provide demographic and historical context to loci associated with complex traits. IBD can account for demography of a population for a given risk allele, that is, a variant arising through mutation or recombination, spreading and surviving in a population due to demographic events and genetic drift, has information that is encoded in the spanning inherited segment that is informative of evolutionary and complex disease processes (Nelson et al., 2018; Tian et al., 2019). With the concept of IBD explained, we will now offer some of the applications in contemporary human genomics.

A crucial goal in population genetics is the estimation of the mutation rate across the genome. IBD-based methods can augment mutation rate estimation approaches by leveraging IBD segments to condition on recent ancestry as part of the estimation process. Prior techniques involved using trios of parents and offspring to estimate mutation rate. However, this approach is difficult to implement due to the logistical challenges of recruiting trios, and is sensitive to genotyping errors or somatic mutations being incorrectly classified as *de novo* mutations (Shah et al., 2018; Tian et al., 2019). In identifying IBD segments, researchers can quantify the *de novo* mutation rate on each segment related to the degree of kinship between the samples to reduce the false positive rate, particularly when compared to small pedigree-based studies. Furthermore, IBD methods allow for the expansion beyond pedigree studies to large-scale population-based datasets by leveraging the inherent background IBD present in human populations, with recent investigations further narrowing the confidence in our estimation of mutation rates to between 1.02×10^{-8} and 1.56×10^{-8} (Campbell et al., 2012; Palamara et al., 2015). Other studies have shown that inferring short IBD segments into longer IBD segments can help to adjust estimations of the *de novo* mutation rate (Chiang et al., 2016). By leveraging IBD, the fundamental question of what mutation rates are across the genome can be more confidently assessed by creating more complete models of mutation, recombination and kinship.

Alongside interrogating the mutation rate of the genome, there has been significant interest in determining the variation in the recombination landscape among global human populations. In addition to having different population level prevalences, the same complex disease loci may exhibit local differences in linkage disequilibrium that directly impact fine-mapping and other common genetic analyses (Wojcik et al., 2019). This means that population-specific recombination maps will be important for fine-mapping both common and rare variants in complex diseases in diverse populations. One recent study showed that building a recombination map from IBD segments yields better estimation of recombinational endpoints and time-to-most-recent-common-ancestor when compared to LD- or admixture-based approaches (Zhou et al., 2020a). Here, IBD methods, particularly those that can work accurately and at scale, can help to create population specific recombination maps that will in turn allow for more accurate simulations of each specific population's demographic history, leading to other downstream applications such as improved imputation.

Identity-by-descent detection also plays into the recent advances in population structure estimation, particularly at fine

scale. Inherent to the idea of a population is the idea of shared ancestry and with this shared ancestry comes a higher probability of relatedness, and a larger portion of the genome shared IBD between any sampled individuals within the same population, when compared to two individuals sampled from between populations. We consider, as an example, the question of improving admixture inference accuracy. By identifying IBD segments among individuals in a population, admixture measurements can be considered with higher accuracy than just comparing genotypes, which may be additionally influenced by errors or somatic mutations. In addition, as studies grow larger, the search space for identifying shared cryptic ancestry as captured by IBD tends to scale quadratically (i.e., with the total pairs of individuals). Thus, a high degree of cryptic relatedness can be present in large-scale genetics studies when a prior, smaller study in the same population may have shown little to no cryptic relatedness. To account for this component of population structure, IBD methods allow researchers to reduce confounding in their study design and better reflect the populations' allele frequencies by matching cases and controls on the basis of genetic ancestry (Palin et al., 2011; Nelson et al., 2018; Sohail et al., 2019).

Concurrent with GWAS, mapping of genetic variants to IBD segments and/or clusters is an alternative method that can help to detect significant associations with a trait of interest. This is similar to how the technique of linkage mapping narrows the genetic signal to a linkage peak (Gusev et al., 2011; Browning and Thompson, 2012). Rare, causal variants preserved in the population while being affected by population demography, drift, selection and substructure have been shown to fall within segments of the genome that are IBD between pairs of individuals in study populations. Analysis of founder populations offer examples of how rare variants can be identified using IBD methods: one example showed how broadly rare European variants contribute disproportionately to disease risk in Quebec (Nelson et al., 2018). Similarly, the elevated IBD patterns present in island populations have empowered novel discoveries, such as the link between height-associated loci and a collagen disorder found in Puerto Ricans (Belbin et al., 2017). With increasing recognition of the role of rare variants in complex disease, and the highly structured manner in which they segregate, methods that leverage IBD for rare variant detection have the potential to be increasingly useful for rare variant discovery.

Finally, imputation can be dramatically improved when leveraging the population specific information inherent to IBD. With growing reference panels from global populations, imputation is resulting in more accurate haplotype matching (Kowalski et al., 2019). IBD can further improve this by noting how to match sample haplotypes to appropriate ancestral references for imputation in a concept called a Study-Specific Reference Panel (SSRP; Gusev et al., 2012; Uricchio et al., 2012; Abney and ElSherbiny, 2019). In practice, modern imputation methods hosted in current servers attempt to approximate this process, but do not recapitulate the augmentation of standard reference panels with appropriate SSRPs (Das et al., 2016). Even without a well annotated pedigree, modern IBD techniques show that imputation quality can be drastically improved when leveraging SSRPs above typical LD based imputation methods (Abney and ElSherbiny, 2019). Not only is IBD useful alone, but it also augments more standard imputation methods by improving imputation probabilities at difficult-to-impute SNPs. By creating custom SSRPs, recruitment efforts to improve representation of understudied populations in human genetics (Bustamante et al., 2011; Popejoy and Fullerton, 2016) can be efficiently leveraged for imputing rare variants, particularly those with greater population-specificity (Gravel et al., 2011).

With the utility of IBD detection outlined, we will next describe the theoretical, statistical and computational means through which IBD detection algorithms are implemented.

OVERVIEW OF METHODS

Both novel computational paradigms and improvements in computational architecture have led to scalable and accurate methods for IBD detection (**Table 1**). Originally, whether through strict string pattern matching or fuzzier matching, methods were not equipped to deal with the inherent quadratic scaling of IBD, limiting the size of initial investigations. The era of high-throughput IBD detection began with GERMLINE (Gusev et al., 2008) to detect variation in IBD patterns efficiently and explore how they are influenced by population processes. GERMLINE creates a hash table between short, exact matches of haplotypes and extending into longer, fuzzy (i.e., allowing for small SNP mismatches or genotype errors) IBD segments. This "seed and extend" paradigm, leveraging the inherent efficiency of short hashing functions for speedup beyond standard pairwise comparisons has been adopted by subsequent detection algorithms (Shemirani et al., 2019; Nait Saada et al., 2020), and improved efficiency over hidden Markov model (HMM)-based algorithms or simpler string matching approaches. The computational efficiency garnered by GERMLINE allows computational time to scale approximately linearly with the number of samples and genotyped variants. While GERMLINE demonstrated accuracy and efficiency in identifying known IBD from simulated datasets and early GWAS studies, it does not easily scale to sample sizes in the hundreds of thousands of individuals, as seen in many contemporary genetic cohorts [although it can provide meaningful insights into biobank-scale data with extensive parallelization (Sapin and Keller, 2021)]. Thus, the primary value in detailing GERMLINE is to describe how it influenced the current IBD calling algorithms outlined below. While GERMLINE works in both diploid and haploid modes, much recent work has been focused on recent haploid methods given the ubiquity of phasing in modern genomic analyses, although we discuss recent efforts in diploid IBD detection as well.

One of recent innovations in the rapid detection of IBD segments is the ILASH algorithm (Shemirani et al., 2019). ILASH works on the principle of locality sensitive hashing (Leskovec et al., 2020) to efficiently search the genome. It begins with a similar "seed and extend" hash table of two individuals in a data set via small stretches of DNA and extending data if the two stretches meet criteria matching IBD similarity. The locality sensitive hashing implemented in

TABLE 1 | Overview of IBD detection tools.

Tool Name	Underlying algorithm	Diploid/ Haploid	Citation	OS compatibility	Link
GERMLINE	Hash and extension	Diploid/Haploid	Gusev et al., 2008	UNIX, compile with make	http://gusevlab.org/projects/germline
RaPID	PBWT on phased haplotypes	Haploid	Naseri et al., 2019	UNIX, pre-compiled	https://github.com/ZhiGroup/RaPID
ILASH	Locality sensitive hashing, extension	Haploid	Shemirani et al., 2019	UNIX, CMAKE v3.5 or higher required	https://github.com/roohy/iLASH
Phaseibd	Templated positional Burrows–Wheeler transform (TPBWT)	Diploid/Haploid	Freyman et al., 2021	UNIX, requires the Python packages: Cython, numpy, and pandas	https://github.com/23andMe/phasedibd
FastSMC	Hash/Extend plus HMM for validation	Haploid	Nait Saada et al., 2020	Ubuntu, macOS, compile with cmake, some python dependencies	https://github.com/PalamaraLab/FastSMC
IBIS	Sliding Window Overlap on unphased genotypes	Diploid	Seidman et al., 2020	UNIX, compile with make	https://github.com/williamslab/ibis
Hap-IBD	Error adjusted PBWT	Haploid	Zhou et al., 2020b	UNIX, runs with java -jar	https://github.com/browning-lab/hap-ibd

PBWT, Positional Burroughs-Wheeler Transformation; HMM, Hidden Markov Model.

ILASH is scalable to IBD detection in tens to hundreds of thousands of individuals, such as in the PAGE Study and UK BioBank. Furthermore, it utilizes multiple parallelized computing across multiple stages of the algorithm to ensure optimization. While ILASH is optimized for the biobank era of genetics and proves easy to use in standard analysis pipelines, there are other algorithms with alternative mathematical and computational approaches.

Another solution to efficient IBD detection is RaPID (Naseri et al., 2019). Instead of locality sensitive hashing, RaPID works through random projections of the low-resolution genetic data and applying the Positional Burroughs-Wheeler Transformation (PBWT; Durbin, 2014)between phased individual haplotypes until a perfect match is obtained. These matches are also stored in a hash table and extended with further matches as previously detailed, combining those results into an IBD segment. While PBWT is an efficient data transformation for genetic data, a key additional step in RaPID incorporates the approximate matching needed to be added to tolerate small mismatches, while only adding trivially to the computational time. Furthermore, the accuracy of results can be improved by subsequent iterations of PWBT, albeit at the cost of longer analysis time. Developers also benchmarked RaPID on simulated and UK BioBank data, showing performance and accuracy results similar to those of ILASH.

Another method that has been developed on top of existing theory is hap-IBD (Zhou et al., 2020b). Building on extensive previous work in IBD estimation through the Beagle software program, researchers have made significant advances in haploid IBD speed. In their most recent efforts, they developed hap-IBD as an algorithm for implementing PBWT similar to RaPID. It differs from RaPID in that it controls for false positives of genotype error or mutation by allowing for small gaps of non-IBS between IBD segments. This allows the algorithm to account for gene conversion, a common phenomenon that can disrupt otherwise IBD segments. In addition, hap-IBD may run the PBWT in parallel, thus showing the best performance among algorithms benchmarked in UK BioBank data. Similarly, investigators at 23andMe leveraged the same PBWT to develop their new Templated PBWT framework (Freyman et al., 2021) with similar properties and efficient, scalable runtime. TPBWT is notable for attempting to identify and correct phase switch errors, thereby improving IBD tract length estimation and long-range phasing.

Another novel algorithmic extension that builds on IBD detection and that shows high performance in accuracy as well as speed is FastSMC (Nait Saada et al., 2020). FastSMC builds upon the hash table GERMLINE method as a first identification step by also including a validation step that uses a approximate coalescent HMM (Palamara et al., 2018). This second step distinguishes between segments of IBS and IBD by estimating the probability a shared IBS segment is due to recent common ancestry, thus allowing for IBD calls within shorter windows. This coalescence probability is reported as an IBD quality score, providing a further layer of information in addition to the IBD haplotypes themselves. By implementing this validation step, FastSMC shows higher accuracy in IBD identification at limited additional computational performance when compared to other algorithms. FastSMC is just one of many IBD identification tools that extend upon the frameworks originated in GERMLINE to improve performance and accuracy, and because of its two-step design, it could easily be adapted to utilize one of the newer IBD detection methods to further improve efficiency of the initial step.

While many IBD detection methods rely upon accurate phasing of alleles, one approach, IBIS, does not have this caveat. IBIS works through long range allelic sharing, detecting shared homozygous alleles between individuals and uses Boolean logic operators to determine IBD from a given rule set (Seidman et al., 2020). The main benefit of IBIS compared to other methods is the time and computational resources saved from not having to pre-phase the genetic data before IBD detection. The major

caveat behind this is that without phase information providing haplotype resolution, excess homozygosity within putative IBD segments can increase the false positive rate, and the shortest segments detectable in diploid IBD are larger than in haploid methods. However, this limitation on segment length (say ∼7 cM for diploid, versus 2–3 cM for haploid) can be acceptable for certain analyses. As previously stated, more recently related individuals share longer IBD segments which may empower risk allele identification or where measuring the length of long IBD segments is of particular importance. Researchers may be especially interested in IBIS as an intermediate analysis strategy, balancing accuracy and speed, for preliminary exploration of a dataset, or for applications that do not require phasing.

A final value to IBD is that in association studies looking for rare, causal variants in complex disease with large biobank sample sized data sets, IBD offers improved statistical power over traditional GWAS methods. This is because, rare variants are much more likely to be found within an IBD cluster (Nait Saada et al., 2020). Coalescence simulation-based work has shown the concordance between IBD and rare exomic variants (Nait Saada et al., 2020). Similarly, in the UK BioBank, researchers found significant associations to blood related traits otherwise not detected in exome-based tests by using IBD methods to predict sharing of ultra-rare, causal variants (MAF < 0.0001; Nait Saada et al., 2020). By identifying regions of IBD where rare, causal variants are likely to occur, the threshold for significance can be appropriately lowered, analogous to how a linkage peak narrows the search for a genetic signal. As a result of looking for associations between IBD segments and complex disease status, we propose the coining of the term "IBDWAS" to make the value of IBD-driven insights more pronounced.

CONCLUSION

To summarize, IBD has significant but often-overlooked meaning in human genetics studies in the context of biobank scale data. All genetic variants affecting traits are influenced by the combination of the evolutionary forces of selection and genetic drift. While in the past inferring the demographic history of a study's population was difficult, the field of genomics has reached datasets so large that ignoring underlying population history can lead to inappropriate conclusions in disease associations and pathogenicity adjudication. As biobank-scale datasets continue to grow, IBD-based analyses offer a paradigm to address unanswered questions within the field of genomics, and with recent advances in IBD-detection methods there are new opportunities to study these patterns of relatedness at scale. It is therefore relevant to incorporate methods of IBD detection into genetic studies to gain insights into the demographic history of variants of interest, to improve statistical power in detecting rare, causal variants, and to improve the accuracy of imputation, among other relevant analyses.

AUTHOR CONTRIBUTIONS

ES initially drafted the manuscript with edits and contributions from GB and CG. All authors contributed to the article and approved the submitted version.

FUNDING

This work was partially funded by the National Institutes of Health under R01HG011345 and U01HG009080.

REFERENCES

Abney, M., and ElSherbiny, A. (2019). Kinpute: using identity by descent to improve genotype imputation. *Bioinformatics* 35, 4321–4326. doi: 10.1093/bioinformatics/btz221

Belbin, G. M., Odgis, J., Sorokin, E. P., Yee, M. C., Kohli, S., Glicksberg, B. S., et al. (2017). Genetic identification of a common collagen disease in puerto ricans via identity-by-descent mapping in a health system. *Elife* 6:e25060. doi: 10.7554/eLife.25060.033

Browning, S. (2008). Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. *Genetics* 178, 2123–2132. doi: 10.1534/genetics.107.084624

Browning, S. R., and Browning, B. L. (2010). High-resolution detection of identity by descent in unrelated individuals. *Am. J. Hum. Genet.* 86, 526–539. doi: 10.1016/j.ajhg.2010.02.021

Browning, S. R., and Browning, B. L. (2012). Identity by descent between distant relatives: detection and applications. *Annu. Rev. Genet.* 46, 617–633. doi: 10.1146/annurev-genet-110711-155534

Browning, S. R., and Browning, B. L. (2015). Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *Am. J. Hum. Genet.* 97, 404–418. doi: 10.1016/j.ajhg.2015.07.012

Browning, S. R., and Thompson, E. A. (2012). Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics* 190, 1521–1531. doi: 10.1534/genetics.111.136937

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

Campbell, C. D., Chong, J. X., Malig, M., Ko, A., Dumont, B. L., Han, L., et al. (2012). Estimating the human mutation rate using autozygosity in a founder population. *Nat. Genet.* 44, 1277–1281. doi: 10.1038/ng.2418

Carmi, S., Palamara, P. F., Vacic, V., Lencz, T., Darvasi, A., and Pe'er, I. (2013). The variance of identity-by-descent sharing in the Wright-Fisher model. *Genetics* 193, 911–928. doi: 10.1534/genetics.112.147215

Chiang, C. W. K., Ralph, P., and Novembre, J. (2016). Conflation of short identity-by-descent segments bias their inferred length distribution. *G3* 6, 1287–1296. doi: 10.1534/g3.116.027581

Das, S., Forer, L., Schonherr, 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

Durbin, R. (2014). Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics* 30, 1266–1272. doi: 10.1093/bioinformatics/btu014

Freyman, W. A., McManus, K. F., Shringarpure, S. S., Jewett, E. M., Bryc, K., Me Research, T., et al. (2021). Fast and robust identity-by-descent inference with the templated positional burrows-wheeler transform. *Mol. Biol. Evol.* 38, 2131–2151. doi: 10.1093/molbev/msaa328

Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., et al. (2011). Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U.S.A.* 108, 11983–11988. doi: 10.1073/pnas.1019276108

Gusev, A., Kenny, E. E., Lowe, J. K., Salit, J., Saxena, R., Kathiresan, S., et al. (2011). DASH: a method for identical-by-descent haplotype mapping uncovers

association with recent variation. *Am. J. Hum. Genet.* 88, 706–717. doi: 10.1016/j.ajhg.2011.04.023

Gusev, A., Lowe, J., Stoffel, M., Daly, M., and Altshuler, D. (2008). Whole population, genomewide mapping of hidden relatedness. *Genome Res.* 19, 318–326. doi: 10.1101/gr.081398.108

Gusev, A., Shah, M. J., Kenny, E. E., Ramachandran, A., Lowe, J. K., Salit, J., et al. (2012). Low-pass genome-wide sequencing and variant inference using identity-by-descent in an isolated human population. *Genetics* 190, 679–689. doi: 10.1534/genetics.111.134874

Henn, B., Hon, L., Macpherson, J., and Eriksson, N. (2012). Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PLoS One* 7:e34267. doi: 10.1371/journal.pone.0034267

Hernandez, R. D., Uricchio, L. H., Hartman, K., Ye, C., Dahl, A., and Zaitlen, N. (2019). Ultrarare variants drive substantial cis heritability of human gene expression. *Nat. Genet.* 51, 1349–1355. doi: 10.1038/s41588-019-0487-7

Kowalski, M. H., Qian, H., Hou, Z., Rosen, J. D., Tapia, A. L., Shan, Y., et al. (2019). Use of >100,000 NHLBI trans-omics for precision medicine (TOPMed) consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet.* 15:e1008500. doi: 10.1371/journal.pgen.1008500

Leskovec, J., Rajaraman, A., and Ullman, J. D. (2020). *Mining of Massive Datasets*. New York, NY: Cambridge University Press. doi: 10.1017/9781108684163

Nait Saada, J., Kalantzis, G., Shyr, D., Cooper, F., Robinson, M., Gusev, A., et al. (2020). Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. *Nat. Commun.* 11:6130. doi: 10.1038/s41467-020-19588-x

Naseri, A., Liu, X., Tang, K., Zhang, S., and Zhi, D. (2019). RaPID: ultra-fast, powerful, and accurate detection of segments identical by descent (IBD) in biobank-scale cohorts. *Genome Biol.* 20:143. doi: 10.1186/s13059-019-1754-8

Nelson, D., Moreau, C., de Vriendt, M., Zeng, Y., Preuss, C., Vezina, H., et al. (2018). Inferring transmission histories of rare alleles in population-scale genealogies. *Am. J. Hum. Genet.* 103, 893–906. doi: 10.1016/j.ajhg.2018.10.017

Palamara, P. F., Francioli, L. C., Wilton, P. R., Genovese, G., Gusev, A., Finucane, H. K., et al. (2015). Leveraging distant relatedness to quantify human mutation and gene-conversion rates. *Am. J. Hum. Genet.* 97, 775–789. doi: 10.1016/j.ajhg.2015.10.006

Palamara, P. F., and Pe'er, I. (2013). Inference of historical migration rates via haplotype sharing. *Bioinformatics* 29, i180–i188. doi: 10.1093/bioinformatics/btt239

Palamara, P. F., Terhorst, J., Song, Y. S., and Price, A. L. (2018). High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nat. Genet.* 50, 1311–1317. doi: 10.1038/s41588-018-0177-x

Palin, K., Campbell, H., Wright, A. F., Wilson, J. F., and Durbin, R. (2011). Identity-by-descent-based phasing and imputation in founder populations using graphical models. *Genet. Epidemiol.* 35, 853–860. doi: 10.1002/gepi.20635

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

Ramstetter, M. D., Dyer, T. D., Lehman, D. M., Curran, J. E., Duggirala, R., Blangero, J., et al. (2017). Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. *Genetics* 207, 75–82. doi: 10.1534/genetics.117.1122

Sapin, E., and Keller, M. C. (2021). Novel approach for parallelizing pairwise comparison problems as applied to detecting segments identical by decent in whole-genome data. *Bioinformatics* 37, 2121–2125. doi: 10.1093/bioinformatics/btab084

Seidman, D. N., Shenoy, S. A., Kim, M., Babu, R., Woods, I. G., Dyer, T. D., et al. (2020). Rapid, phase-free detection of long identity-by-descent segments enables effective relationship classification. *Am. J. Hum. Genet.* 106, 453–466. doi: 10.1016/j.ajhg.2020.02.012

Shah, N., Hou, Y. C., Yu, H. C., Sainger, R., Caskey, C. T., Venter, J. C., et al. (2018). Identification of misclassified clinvar variants via disease population prevalence. *Am. J. Hum. Genet.* 102, 609–619. doi: 10.1016/j.ajhg.2018.02.019

Shemirani, R., Belbin, G. M., Avery, C. L., Kenny, E. E., Gignoux, C. R., and Ambite, J. L. (2019). Rapid detection of identity-by-descent tracts for mega-scale datasets. *bioRxiv* [Preprint]. doi: 10.1101/749507

Slatkin, M., and Rannala, B. (2000). Estimating allele age. *Annu. Rev. Genomics Hum. Genet.* 1, 225–249. doi: 10.1146/annurev.genom.1.1.225

Sohail, M., Maier, R. M., Ganna, A., Bloemendal, A., Martin, A. R., Turchin, M. C., et al. (2019). Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *Elife* 8:e39702. doi: 10.7554/eLife.39702

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 590, 290–299. doi: 10.1038/s41586-021-03205-y

Thompson, E. A. (2013). Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics* 194, 301–326. doi: 10.1534/genetics.112.148825

Tian, X., Browning, B. L., and Browning, S. R. (2019). Estimating the genome-wide mutation rate with three-way identity by descent. *Am. J. Hum. Genet.* 105, 883–893. doi: 10.1016/j.ajhg.2019.09.012

Uricchio, L. H., Chong, J. X., Ross, K. D., Ober, C., and Nicolae, D. L. (2012). Accurate imputation of rare and common variants in a founder population from a small number of sequenced individuals. *Genet. Epidemiol.* 36, 312–319. doi: 10.1002/gepi.21623

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature* 570, 514–518. doi: 10.1038/s41586-019-1310-4

Zhou, Y., Browning, B. L., and Browning, S. R. (2020a). Population-specific recombination maps from segments of identity by descent. *Am. J. Hum. Genet.* 107, 137–148. doi: 10.1016/j.ajhg.2020.05.016

Zhou, Y., Browning, S. R., and Browning, B. L. (2020b). A fast and simple method for detecting identity-by-descent segments in large-scale data. *Am. J. Hum. Genet.* 106, 426–437. doi: 10.1016/j.ajhg.2020.02.010

Author Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

PERSPECTIVE
published: 27 September 2021
doi: 10.3389/fgene.2021.643883

The Opportunities and Challenges of Integrating Population Histories Into Genetic Studies for Diverse Populations: A Motivating Example From Native Hawaiians

Charleston W. K. Chiang[1,2]*

[1] Department of Population and Public Health Sciences, Center for Genetic Epidemiology, Keck School of Medicine, University of Southern California, Los Angeles, CA, United States, [2] Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, United States

There is a well-recognized need to include diverse populations in genetic studies, but several obstacles continue to be prohibitive, including (but are not limited to) the difficulty of recruiting individuals from diverse populations in large numbers and the lack of representation in available genomic references. These obstacles notwithstanding, studying multiple diverse populations would provide informative, population-specific insights. Using Native Hawaiians as an example of an understudied population with a unique evolutionary history, I will argue that by developing key genomic resources and integrating evolutionary thinking into genetic epidemiology, we will have the opportunity to efficiently advance our knowledge of the genetic risk factors, ameliorate health disparity, and improve healthcare in this underserved population.

Keywords: population genetics, human genetics, genome-wide association studies, natural selection, Native Hawaiians, demographic history

OPEN ACCESS

Edited by:
Jeremy Berg,
University of Chicago, United States

Reviewed by:
Tony Merriman,
University of Otago, New Zealand
Levon Yepiskoposyan,
Armenian National Academy of Sciences, Armenia

***Correspondence:**
Charleston W. K. Chiang
charleston.chiang@med.usc.edu

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *19 December 2020*
Accepted: *19 August 2021*
Published: *27 September 2021*

Citation:
Chiang CWK (2021) The Opportunities and Challenges of Integrating Population Histories Into Genetic Studies for Diverse Populations: A Motivating Example From Native Hawaiians. Front. Genet. 12:643883. doi: 10.3389/fgene.2021.643883

INTRODUCTION

Genome-wide association studies (GWASs) have revealed the polygenic nature of human complex traits and diseases (Hirschhorn and Daly, 2005; McCarthy et al., 2008; Visscher et al., 2017), but these successes are heavily biased toward European-ancestry populations (Need and Goldstein, 2009; Popejoy and Fullerton, 2016; Spratt et al., 2016). To truly personalize medicine for everyone, we need to better understand both environmental/lifestyle risk factors and the genetic etiology of complex diseases, particularly in geographically diverse, often underserved, populations. It remains a challenge to attain sample sizes from diverse populations comparable to existing European-ancestry cohorts (>1 million individuals). Even when genetic data from understudied populations are included, they often comprise a small contributing part of a larger consortium, thereby masking any population-specific effects. There is thus a need to broadly include diverse populations in genomic studies through focused efforts. Whereas consortium-scale sample sizes are required to detect individual variants with ever-decreasing effect sizes associated with a complex trait, the genetic contributions to phenotypic differences among populations result from the distinct population history and unique interactions with the environment of the past or the present, which can be learned from moderately sized studies. For understudied populations, the focus is therefore

both to transfer knowledge gained from large-scale Euro-centric studies and to supplement our understanding with insights specific to the population at hand.

Genetic and phenotypic differences between populations can arise through two broad categories of evolutionary mechanisms: demographic events and natural selection. An example of demographic events is a population bottleneck. In a bottlenecked population, alleles with functional, deleterious, consequences can, by chance, overcome the impact of negative selection (Ohta, 1973) to reach higher frequencies and, in turn, explain a greater proportion of the heritability of a complex trait compared to alleles in a non-bottlenecked population (Lim et al., 2014; Lohmueller, 2014; Locke et al., 2019). An example of natural selection is local adaptation to selective pressures such as climate, diet, UV exposures, or pathogens (Fan et al., 2016; Mathieson, 2020; Rees et al., 2020). Alleles underlying adaptive traits will increase in frequency in the local population. But as the environment changed in modern societies, these adaptations could manifest as diseases and contribute to differences in genetic risk between populations (Greaves, 2007; Stearns et al., 2010; Fay, 2013). Leveraging these evolutionary events in practice has already identified population-enriched alleles disproportionately contributing to human complex traits in multiple populations around the globe (Zhernakova et al., 2010; Moltke et al., 2014; Sidore et al., 2015; Zoledziewska et al., 2015; Minster et al., 2016; Steri et al., 2017; Grarup et al., 2018a,b; Locke et al., 2019; Asgari et al., 2020; Lin et al., 2020). These discovered alleles are oftentimes rare and difficult to map in large continental populations, but were found using only a moderately sized (by GWAS standards) cohort. Therefore, a better understanding of our evolutionary past will enable better designs and interpretations of genetic epidemiology studies, provide an opportunity to better understand the biology of human traits and diseases, help explain the disparity in risks among populations today, and allow the incorporation of evolutionary insights into our clinical practice (Stearns et al., 2010). However, these questions have not been systematically investigated in geographically diverse populations around the globe.

As an illustrative and motivating example, I will describe the challenges and benefits to combine evolutionary insights and genetic studies with the Native Hawaiian population. Though they are one of the smallest ethnic minorities in the United States, consisting of 1.2 million individuals and 0.4% of the United States census in 2010, Native Hawaiians and other Pacific Islanders (alone or in combination with other races) showed the second fastest rate of growth at 40% between 2000 and 2010. Compared to European- or Asian-Americans, Native Hawaiians display alarming rates of obesity, diabetes, cardiovascular diseases, cancers, and other related chronic health conditions (Grandinetti et al., 2002; Pike et al., 2002; Maskarinec et al., 2009; Mau et al., 2009; Madan et al., 2012; Singh and Lin, 2013; Tung and Barnes, 2014; Braden and Nigg, 2016). Environmental and/or social factors undoubtedly play an important role for these disparity, but in some cases, the risks for diseases are elevated even after adjusting for BMI and other socioeconomic and lifestyle factors (Pike et al., 2002; Maskarinec et al., 2009; Madan et al., 2012; Singh and Lin, 2013). This suggests that systematic differences in the number, frequencies, or effects of genetic risk alleles could partly explain the differences in risk among populations. The history of Native Hawaiians exemplifies all major evolutionary mechanisms influencing the pattern of variations in humans – population size changes, adaptation, and recent admixture. I will describe the opportunities to leverage extensively characterized genetic history for understanding the Hawaiian-specific disease architecture, current challenges that inhibit large-scale and systematic genetic studies, and important considerations of partnering with Native Hawaiians to perform genetic research. While I focus on leveraging evolutionary insight to improve the design and interpretation of genomic studies in understudied populations, there are important ethical considerations of studies with indigenous communities. I describe briefly my own experience and approach, and note that a large body of literature exists (e.g., Claw et al., 2018; Merriman and Wilcox, 2018; Garrison et al., 2019; Fox, 2020; Hudson et al., 2020, among others) that could not be covered in detail here. Finally, the opportunities and challenges described here are not limited to Native Hawaiians and are generally applicable to other understudied populations around the globe.

DEMOGRAPHIC AND ADMIXTURE HISTORY OF NATIVE HAWAIIANS

There is no detailed characterization of the demographic history of Native Hawaiians using genetic data, though there are suggested models for Eastern Polynesians based on archeological findings, ancient and modern DNA studies, and oral history. Because of the shared genetic ancestry with aboriginal people in Island Southeast Asia, it has been hypothesized that Austronesian-speaking people from locations such as Taiwan or the Philippines migrated to the remote reaches of Oceania and Western Polynesia about 2,000–3,000 years ago (Bellwood, 2011; Skoglund et al., 2016; Gosling and Matisoo-Smith, 2018; Hudjashov et al., 2018; Lipson et al., 2018; Posth et al., 2018). These Austronesians settled in islands like Vanuatu, Tonga, and Samoa for nearly 1,000–2,000 years (Nordyke, 1989; Gosling et al., 2015), where they coinhabited with the Papuan-speaking natives of Northern Melanesia. Today, Polynesian populations [including the Native Hawaiians (Kim et al., 2012)] have varying levels of an ancestry found predominantly in present-day Papuans (Skoglund et al., 2016; Lipson et al., 2018; Posth et al., 2018). The ancient Polynesians began long-range seafaring to the vast stretches of the Pacific around 200 B.C. to 700 A.D., arriving at Hawai'i between 900 A.D. and 1300 A.D. (Kirch, 1985; Bellwood, 1987; Nordyke, 1989). Inter-island interactions were initially frequent but ceased by the 1400s perhaps due to the development of more complex sociopolitical structures. Native Hawaiians then became relatively isolated until the European settlers arrived (Nordyke, 1989; Gosling et al., 2015). Records of Native Hawaiian population sizes pre-European contact are unreliable, but the effective population sizes (N_e) for Native Hawaiians are likely small throughout history since a genetically estimated N_e as recent as 1,000 years ago was reported to be ∼1,000 for Melanesians and Samoans (Bergström et al., 2020;

Harris et al., 2020). Thus, the demographic history of the Native Hawaiians is likely characterized by multiple founding events and persistent small sizes, which would permit rare alleles to drift to higher frequencies and contribute uniquely to the genetic architecture. Like previous examples from Sardinia, Peru, and Samoa (Sidore et al., 2015; Zoledziewska et al., 2015; Minster et al., 2016; Asgari et al., 2020), a moderate-sized cohort of Native Hawaiians and other Polynesians could provide power to detect these population-specific associations.

Native Hawaiians are also recently admixed. The largest wave of migrants occurred following Captain James Cook's arrival in Hawai'i in 1778. Immigrants and missionaries from Europe and Americas as well as laborers from China and East Asia arrived throughout the 19th and 20th centuries. African-ancestry individuals began arriving on the island in the 20th century, mostly as part of the military force (Nordyke, 1989). Today, Native Hawaiians are the group most likely to report having two or more components of ancestry in the United States census (Humes et al., 2011), deriving major continental ancestry from the Polynesians, Europeans, and East Asians (Sun et al., 2021). Variations of these continental ancestries would also partly explain risks of diseases in Native Hawaiians. For example, an individual's proportion of Polynesian ancestry is associated with the risk of obesity, while both Polynesian and East Asian ancestries contribute to the risk of type 2 diabetes (T2D) (Sun et al., 2021; **Figure 1**). Note that Polynesian ancestry here is better considered as the component that spread across Polynesia from the initial settlements in remote Oceania. This component itself may be a mixture of the ancient Austronesians that showed close affinity to the East Asian ancestry, as well as the component ancestry native to Melanesia and found predominantly in Papuans today (Gosling et al., 2015; Skoglund et al., 2016). Moreover, while the associations of disease risks with Polynesian ancestry suggest the presence of Polynesian-specific genetic risk factors, the associations are also likely to reflect any cultural or environmental non-genetic factors correlated with Polynesian ancestry (e.g., diet). Nevertheless, past admixture events suggest that approaches such as admixture mapping (Winkler et al., 2010; Shriner, 2017) could identify regions of the genome disproportionately impacting the health of Native Hawaiians.

POTENTIAL ROLE OF ADAPTATION IN SHAPING THE GENETIC ARCHITECTURE

Adaptive events likely shaped the genetic architecture of complex traits in Native Hawaiians. The successful settlement of previously uninhabited Hawaiian archipelago likely involved adopting new subsistence strategies and overcoming famines, nutritional deficiencies, and higher tropical load of infections (Gosling et al., 2015). The encounter in the 18th century with Europeans and their pathogens deeply impacted the Native Hawaiians: historians have suggested that pathogens such as syphilis, gonorrhea, measles, whooping cough, mumps, cholera, or smallpox, among others, contributed to up to an 80% decrease in census size in Hawai'i between 1780 and 1850 (Nordyke, 1989). Diets and pathogens are well-known evolutionary forces that shaped the human genome and contributed to phenotypic differences between populations today (Fan et al., 2016; Mathieson, 2020; Rees et al., 2020). As such, adaptation, whether due to forces of nature or actions of the people, could also leave a lasting imprint on the health of Native Hawaiians. However, this hypothesis has not been systematically tested in Native Hawaiians or any Polynesian populations.

Native Hawaiians, and Polynesian populations at large, are more susceptible to metabolic diseases such as obesity and type 2 diabetes (Maskarinec et al., 2009, 2016; Madan et al., 2012; Gosling et al., 2015; Minster et al., 2016; Sun et al., 2021). One contested explanation for this elevated susceptibility is the "Thrifty Gene Hypothesis," which stipulates that efficient energy storage during times of famine in the past provided an evolutionary advantage that is no longer consistent with the present-day diets. This hypothesis could explain the higher burden of metabolic diseases observed in Polynesian populations today, but there are questions of whether the diversity of environments and genetic ancestries across the Pacific populations would all converge on the same manifestation of risk for metabolic syndromes (Gosling et al., 2015). Genetic support for the Thrifty Gene Hypothesis in other populations has been inconclusive (Ayub et al., 2014; Koh et al., 2014). Results from recent genomic data from Polynesian populations have also been inconsistent, though generally based on single or a few loci (Cadzow et al., 2016; Minster et al., 2016; Lin et al., 2020). Therefore, it is difficult to ascribe the hypothesized selective pressure to the genetic evidence of adaptation. Ultimately, the Thrifty Gene Hypothesis is just one possible reason for adaptation. The focus is not testing the Thrifty Gene Hypothesis, *per se*, but to understand the link between past adaptation and present-day health. Given the advancement in population genetic methods to detect selection across different time scales (Field et al., 2016; Palamara et al., 2018; Edge and Coop, 2019; Speidel et al., 2019), and the emerging genomic data from large epidemiological cohorts from Polynesian populations (Minster et al., 2016; Sun et al., 2021), there is now an opportunity to systematically survey the genome for signature of adaptation and assess their modern-day health consequences.

CHALLENGES IN GENOMIC STUDIES WITH NATIVE HAWAIIANS

One deterrent to including Native Hawaiians in genomic studies is the underdevelopment of genomic resources. For other continental populations, these resources have been abundant and publicly available, enabling large-scale collaborations and investigations. Development of these resources in Native Hawaiians or other Polynesian populations will similarly accelerate genetic research in these populations.

One sorely needed resource is a catalog of genetic variation, akin to gnomAD, which contains variation discovered from sequencing data of up to ∼141,000 individuals (Karczewski et al., 2020). This catalog has substantially improved clinicians' ability to interpret clinical sequencing data of severe and rare genetic

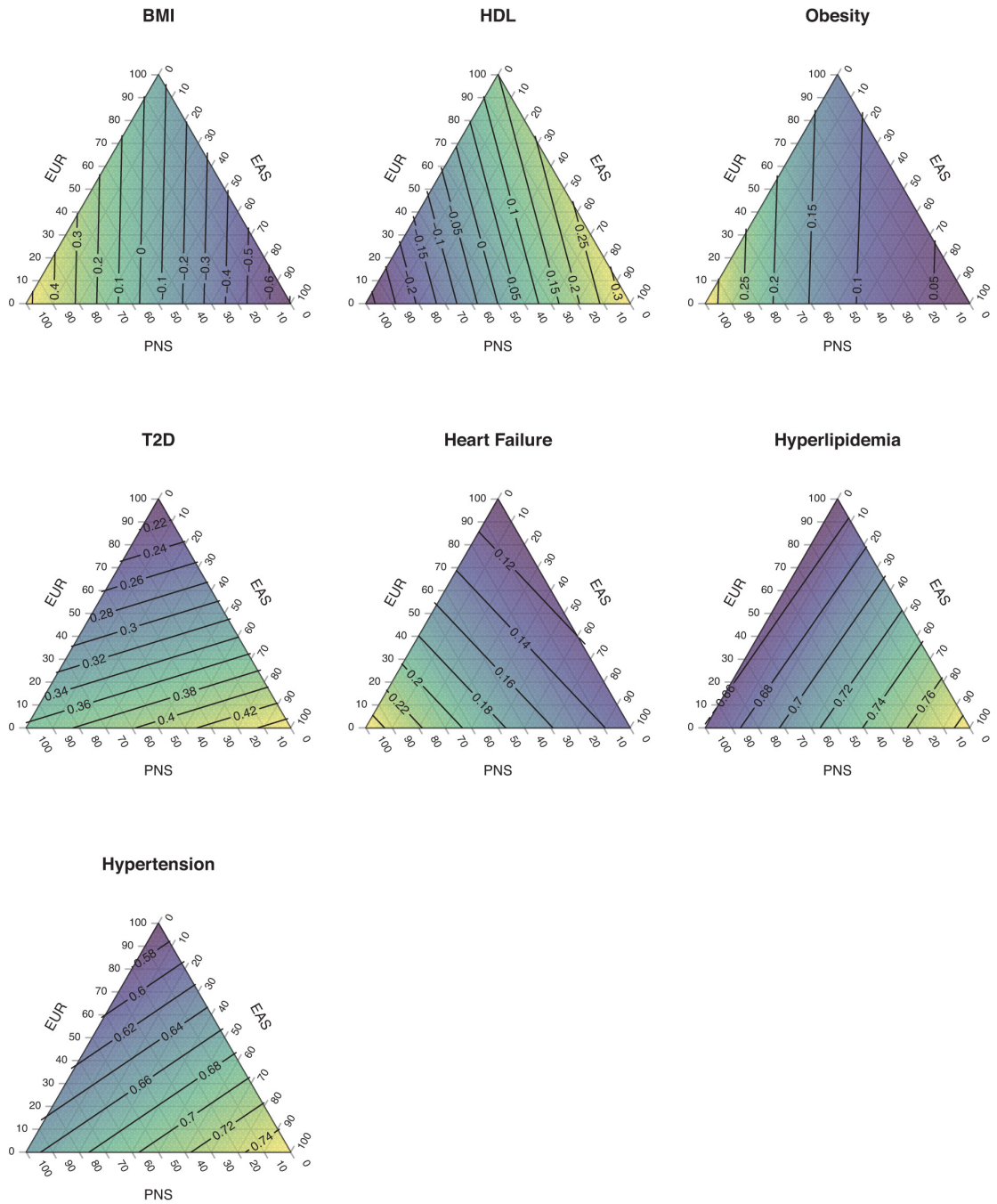

FIGURE 1 | Impact of ancestry components on complex traits and disease risks in Native Hawaiians. The distribution of estimated disease risk are shown as a function of a three-component ancestry model. The linear models used were described in Sun et al. (2021), where for each trait examined as the dependent variable, the effect sizes of the relevant independent variables (e.g., age, BMI, and estimated genetic ancestry as scalar variables, or education level as the categorical variable) were estimated from a Native Hawaiian cohort. Quantitative (BMI and HDL) traits were modeled using linear regression, which predicts the estimated trait value in units of standard deviations given the genetic ancestries. Binary [obesity, type 2 diabetes (T2D), heart failure, hyperlipidemia, and hypertension] traits were modeled using logistic regression, which predicts the probability of disease given genetic ancestries and other covariates. An adult male with age = 50 years, BMI = 30 units (excluded from the obesity model), and education level = college graduate was assumed for calculating probability of disease or estimated trait value. For simplicity, a three-component ancestry model with contributions only from European (EUR), East Asian (EAS), and Polynesian (PNS) ancestors was assumed for Native Hawaiians. The predicted values were interpolated across all possible combinations of ancestries and shown with contour lines. For example, a hypothetical individual with 80% PNS ancestry, 10% EAS, and 10% EUR ancestry aged 50 years, with BMI 30 and college degree, is predicted to have 35–36% chance of being affected with T2D. Similarly, someone with 10% PNS ancestry, 80% EAS, and 10% EUR ancestry of the same age, BMI, and education level is predicted to have ∼42% chance of being affected with T2D. Risk for T2D in Native Hawaiians increases with both PNS and EAS components of ancestry. Note that genetic ancestry captures both genetic and correlated environmental/cultural effects.

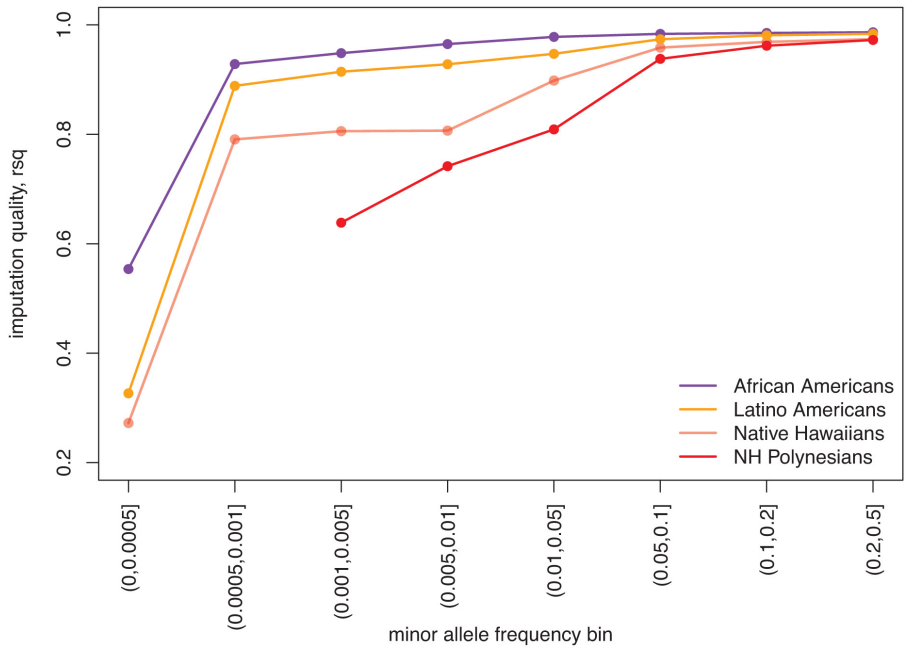

FIGURE 2 | Relatively poor imputation quality for Native Hawaiians due to underrepresentation in imputation reference panels. We imputed 5,325 African Americans, 2,838 Latino Americans, and 3,940 Native Hawaiians from the Multiethnic Cohort (Kolonel et al., 2000) using freeze 8 of the TOPMED imputation server (Taliun et al., 2021) (imputed in July 2020). Each population was genotyped on the MEGA array and subjected to the same QC filters. As measured by the mean imputation quality, R^2 (rsq), Native Hawaiian individuals are imputed more poorly than other United States ethnic minority populations, particularly for variants with minor allele frequency <5%. The disparity is even stronger when focusing on only the 178 Native Hawaiians with estimated Polynesian ancestry >90% (NH Polynesians) (Lin et al., 2020).

diseases and to reach a genetic diagnosis. Though still dominated by genomic data from European individuals, gnomAD does include data from ∼20,000 individuals of African ancestry, and similar catalogs are emerging from Asians as well (Chiang et al., 2018; Liu et al., 2018; GenomeAsia100K Consortium, 2019)[1]. However, Native Hawaiians, or Polynesians in general, are not yet represented in these catalogs. The publicly available sequencing data of Native Hawaiians are limited to data from a single individual in the Simons Genome Diversity Project (Mallick et al., 2016). [There are also ∼28 individuals across Oceania in the Human Genome Diversity Panel (Bergström et al., 2020).] Going forward, the sample size need not be large – even a few hundred individuals will allow one to detect nearly all common variations (with frequency >1%) in the population. Since many of these variants will be Polynesian-specific and have not been observed elsewhere in the world, such a catalog will further improve physicians' ability to interpret variants of unknown significance in the clinical setting to directly benefit the Polynesian community (Easteal et al., 2020).

To accelerate the discovery of genetic associations to diseases, we also need to improve Native Hawaiian representation in imputation reference panels. Genome-wide genotyping followed by imputation of the unobserved genetic variation is one of the most efficient approaches to conduct genetic association studies. Publicly available imputation reference panels are constantly growing in size, allowing investigators to query rarer variations that are usually absent on genotyping arrays. Because of the lack of representation in imputation reference panels, the quality of imputation in Native Hawaiians lags significantly behind that of other ethnic minorities (**Figure 2**). In a proof-of-principle study, it was shown that rs373863828 in *CREBRF* is associated with a large effect on BMI and T2D in Native Hawaiians, but could not be imputed or discovered using publicly available imputation resources at the time, despite the study having sufficient statistical power to do so (Lin et al., 2020). The lack of representation has thus contributed to the disparity in bringing genomic medicine to Native Hawaiians compared to other ethnic minorities in the United States.

Ultimately, larger cohorts will boost statistical power and undoubtedly enhance the genomic insights we can garner, but large recruitments in indigenous communities such as the Native Hawaiians have been challenging. The population sizes of any indigenous population are already small, and past mistakes by researchers, such as the Havasupai diabetes study that misused genetic information from the indigenous community in unconsented studies (Garrison et al., 2019), have also caused community mistrust in scientists. In a recent assessment of Pacific Islanders, over 65% of participants shared some reservation or reluctance about providing biospecimens for research, citing concerns due to spirituality, lack of knowledge of research, or invasion of privacy, among others (Kwan et al., 2015). With increasing awareness of these past mistakes, genome scientists should open dialog with the community early and

[1]Genome Medical alliance Japan Project. A Comprehensive Japanese Genetic Variation Database. Available Online at: https://togovar.biosciencedbc.jp/

often, respect both community and individual consent, and *partner with indigenous communities* rather than just enrolling them as participants (Claw et al., 2018; Garrison et al., 2019; Hudson et al., 2020).

DISCUSSION

Population genetic theories predict the existence of unique genetic variants segregating in the Native Hawaiian population that disproportionately impact their health. Identifying these variants could significantly improve healthcare practices and directly benefit this community. Though several challenges currently exist, the outlook for genetic research in Native Hawaiians and other diverse populations in general can be promising while requiring only a moderate level of funding commitments. Whole genome sequencing of only 150–200 Native Hawaiian individuals would already allow better imputation of Native Hawaiian individuals in a genetic study and accelerate the discovery of population-specific alleles of large effects (Jewett et al., 2012; Lin et al., 2020). The generation and aggregation of WGS data from multiple Polynesian populations will also provide the catalog of genetic variation currently lacking in Polynesian populations, make an immediate impact in the clinical care of Polynesian populations, and accelerate future large-scale genomic research in these populations. Deploying low-coverage sequencing as an alternative first step could also efficiently identify population-specific alleles (Sidore et al., 2015; Chiang et al., 2018; Martin et al., 2021). Importantly, this roadmap is cost-efficient, achievable by pooling resources from a handful of research labs. These are realistic outlooks over the next 5 years.

However, it is important to develop the partnership of the indigenous community in order for the research to proceed. Past exploitation of indigenous populations (Claw et al., 2018; Garrison et al., 2019; Hudson et al., 2020) and the lack of benefits sharing from lucrative pharmaceutical enterprises (Fox, 2020) have brooded mistrust between underprivileged communities and scientists. Research with the indigenous community must also have the community benefits in mind. Note that as health disparity between populations is also driven by non-genetic or social factors, the health benefits derived directly from genomic studies, if any, will likely be slow and not immediately apparent. Nevertheless, it is still important for genomic research to be inclusive if we want to achieve equity and representation; in fact, exclusion of a group of people from research may contribute to inequity in itself. In this context, it is often beneficial for research to be led by scientists of the indigenous community as they are more knowledgeable of the local cultural practices. Alas, there is a dearth of indigenous researchers in the specific research domain described here (see Popejoy and Fullerton, 2016; Merriman and Wilcox, 2018). Whereas pharmaceutical or biotech companies are positioned to directly benefit indigenous communities through proceeds distributions or profit sharing, individual researchers, including non-indigenous ones, are positioned to tailor their engagement to the unique circumstances of each community. By leveraging their long-term individualized interactions, individual researchers will be able to engage in outreach and develop improved and informed consent process, act in stewardship of indigenous data, and help build research capacity through training of the indigenous scientists.

Working within the framework of the Multiethnic Cohort (Kolonel et al., 2000) study, every one of my research projects with – and generally all research proposals utilizing biospecimen data from – the Native Hawaiian population is reviewed by the Native Hawaiian Community Advisory Board (NHCAB) composed of scholars and advocates from the community. A recent study from my group investigating the impact of genetic ancestry on risk of disease in Native Hawaiians (Sun et al., 2021) exemplifies how dialog with community representatives provided the appropriate cultural context. In this study, we observed that the Polynesian component of genetic ancestry (sometimes also with the East Asian component) is associated with risk to certain cardiometabolic diseases (Sun et al., 2021). Through constructive comments from the NHCAB on the early drafts of the manuscript, we came to appreciate that even though the quantification of components of genetic ancestries is a common first step to dissect population-specific genetic risk factors, it should not supplant current approaches (e.g., self-identification or genealogical records) to define community membership. As researchers, we are aware of the deficiency of research methods. We knew that estimated ancestry proportions can be sensitive to the choice of variants analyzed or reference panels used (Uren et al., 2020). We also understood the conceptual difference between genetic ancestry and genealogical ancestry. That is, an individual may not inherit any genetic material from a genealogical ancestor (Donnelly, 1983). But we did not necessarily appreciate how an estimated quantity for research use could detract from an individual's cultural identity or heritage. It is through communication with the NHCAB that we stressed and repeatedly clarified this concept in our eventual manuscript, and the reviewers noticed.

This is but the first step of active community engagement. A step toward the right direction, but the efforts need to be broadened and made consistent. The Aotearoa New Zealand genomic variome project (Caron et al., 2020) is an example of an inclusive framework in Polynesian populations that others can borrow. The Multiethnic Cohort has been entrusted by >5,000 self-identified Native Hawaiians who donated their biospecimen for research. These individuals have continued to show their support for research by responding to follow-up questionnaires, suggesting that the community is clearly open to partake in research. Now it is up to individual researchers, indigenous or non-indigenous alike, to continue to earn the trust from the indigenous community and be an ally.

DATA AVAILABILITY STATEMENT

Datasets analyzed in this can be found here in dbGAP with accession number phs000220.v2.p2.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Boards of the University

of Hawai'i and the University of Southern California. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CWKC conceived and designed the study, performed the analysis, and wrote the manuscript.

FUNDING

Research reported in this publication was supported by the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health under award number R35GM142783 (to CWKC).

ACKNOWLEDGMENTS

I would like to thank John Novembre, Vivian U, Philip Wilcox, Claradina Soto (Navajo/Jemez Pueblo), and members of the Native Hawaiian Community Advisory Board at the University of Hawai'i Cancer Center for their critical comments on earlier versions of this manuscript. I would also like to thank Xin Sheng, Victor Hom, and Bryan L. Dinh for assistance with imputation using the TOPMed reference panel. Computation for this work was supported by the Center for Advanced Research Computing (CARC) at the University of Southern California (https://carc.usc.edu).

REFERENCES

Asgari, S., Luo, Y., Akbari, A., Belbin, G. M., Li, X., Harris, D. N., et al. (2020). A positively selected FBN1 missense variant reduces height in Peruvian individuals. *Nature* 582, 234–239. doi: 10.1038/s41586-020-2302-0

Ayub, Q., Moutsianas, L., Chen, Y., Panoutsopoulou, K., Colonna, V., Pagani, L., et al. (2014). Revisiting the thrifty gene hypothesis via 65 loci associated with susceptibility to type 2 diabetes. *Am. J. Hum. Genet.* 94, 176–185. doi: 10.1016/j.ajhg.2013.12.010

Bellwood, P. (2011). Holocene Population History in the Pacific Region as a Model for Worldwide Food Producer Dispersals. *Curr. Anthropol.* 52, S363–S378. doi: 10.1086/658181

Bellwood, P. S. (1987). *The Polynesians: Prehistory of an Island People. Rev. Ed.* London: Thames and Hudson.

Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science* 367:eaay5012. doi: 10.1126/science.aay5012

Braden, K. W., and Nigg, C. R. (2016). Modifiable Determinants of Obesity in Native Hawaiian and Pacific Islander Youth. *Hawaii J. Med. Public Health* 75, 162–171.

Cadzow, M., Merriman, T. R., Boocock, J., Dalbeth, N., Stamp, L. K., Black, M. A., et al. (2016). Lack of direct evidence for natural selection at the candidate thrifty gene locus, PPARGC1A. *BMC Med. Genet.* 17:80. doi: 10.1186/s12881-016-0341-z

Caron, N. R., Chongo, M., Hudson, M., Arbour, L., Wasserman, W. W., Robertson, S., et al. (2020). Indigenous Genomic Databases: pragmatic Considerations and Cultural Contexts. *Front. Public Health* 8:111. doi: 10.3389/fpubh.2020.00111

Chiang, C. W. K., Mangul, S., Robles, C., and Sankararaman, S. A. (2018). Comprehensive Map of Genetic Variation in the World's Largest Ethnic Group-Han Chinese. *Mol. Biol. Evol.* 35, 2736–2750. doi: 10.1093/molbev/msy170

Claw, K. G., Anderson, M. Z., Begay, R. L., Tsosie, K. S., Fox, K., Garrison, N. A., et al. (2018). A framework for enhancing ethical genomic research with Indigenous communities. *Nat. Commun.* 9:2957. doi: 10.1038/s41467-018-05188-3

Donnelly, K. P. (1983). The probability that related individuals share some section of genome identical by descent. *Theor. Popul. Biol.* 23, 34–63. doi: 10.1016/0040-5809(83)90004-7

Easteal, S., Arkell, R. M., Balboa, R. F., Bellingham, S. A., Brown, A. D., Calma, T., et al. (2020). Equitable Expanded Carrier Screening Needs Indigenous Clinical and Population Genomic Data. *Am. J. Hum. Genet.* 107, 175–182. doi: 10.1016/j.ajhg.2020.06.005

Edge, M. D., and Coop, G. (2019). Reconstructing the History of Polygenic Scores Using Coalescent Trees. *Genetics* 211, 235–262. doi: 10.1534/genetics.118.301687

Fan, S., Hansen, M. E., Lo, Y., and Tishkoff, S. A. (2016). Going global by adapting local: a review of recent human adaptation. *Science* 354, 54–59. doi: 10.1126/science.aaf5098

Fay, J. C. (2013). Disease consequences of human adaptation. *Appl. Transl. Genom.* 2, 42–47. doi: 10.1016/j.atg.2013.08.001

Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., Golan, D., et al. (2016). Detection of human adaptation during the past 2000 years. *Science* 354, 760–764. doi: 10.1126/science.aag0776

Fox, K. (2020). The Illusion of Inclusion - The "All of Us" Research Program and Indigenous Peoples' DNA. *N. Engl. J. Med.* 383, 411–413. doi: 10.1056/NEJMp1915987

Garrison, N. A., Hudson, M., Ballantyne, L. L., Garba, I., Martinez, A., Taualii, M., et al. (2019). Genomic Research Through an Indigenous Lens: understanding the Expectations. *Annu. Rev. Genomics Hum. Genet.* 20, 495–517. doi: 10.1146/annurev-genom-083118-015434

GenomeAsia100K Consortium (2019). The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* 576, 106–111. doi: 10.1038/s41586-019-1793-z

Gosling, A. L., Buckley, H. R., Matisoo-Smith, E., and Merriman, T. R. (2015). Pacific Populations, Metabolic Disease and "Just-So Stories": a Critique of the "Thrifty Genotype" Hypothesis in Oceania. *Ann. Hum. Genet.* 79, 470–480. doi: 10.1111/ahg.12132

Gosling, A. L., and Matisoo-Smith, E. A. (2018). The evolutionary history and human settlement of Australia and the Pacific. *Curr. Opin. Genet. Dev.* 53, 53–59. doi: 10.1016/j.gde.2018.06.015

Grandinetti, A., Chen, R., Kaholokula, J. K., Yano, K., Rodriguez, B. L., Chang, H. K., et al. (2002). Relationship of blood pressure with degree of Hawaiian ancestry. *Ethn. Dis.* 12, 221–228.

Grarup, N., Moltke, I., Andersen, M. K., Bjerregaard, P., Larsen, C. V. L., Dahl-Petersen, I. K., et al. (2018a). Identification of novel high-impact recessively inherited type 2 diabetes risk variants in the Greenlandic population. *Diabetologia* 61, 2005–2015. doi: 10.1007/s00125-018-4659-2

Grarup, N., Moltke, I., Andersen, M. K., Dalby, M., Vitting-Seerup, K., Kern, T., et al. (2018b). Loss-of-function variants in ADCY3 increase risk of obesity and type 2 diabetes. *Nat. Genet.* 50, 172–174. doi: 10.1038/s41588-017-0022-7

Greaves, M. (2007). Darwinian medicine: a case for cancer. *Nat. Rev. Cancer* 7, 213–221. doi: 10.1038/nrc2071

Harris, D. N., Kessler, M. D., Shetty, A. C., Weeks, D. E., Minster, R. L., Browning, S., et al. (2020). Evolutionary history of modern Samoans. *Proc. Natl. Acad. Sci. U. S. A.* 117, 9458–9465. doi: 10.1073/pnas.1913157117

Hirschhorn, J. N., and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* 6, 95–108. doi: 10.1038/nrg1521

Hudjashov, G., Endicott, P., Post, H., Nagle, N., Ho, S. Y. W., Lawson, D. J., et al. (2018). Investigating the origins of eastern Polynesians using genome-wide data from the Leeward Society Isles. *Sci. Rep.* 8:1823. doi: 10.1038/s41598-018-20026-8

Hudson, M., Garrison, N. A., Sterling, R., Caron, N. R., Fox, K., Yracheta, J., et al. (2020). Rights, interests and expectations: indigenous perspectives on

unrestricted access to genomic data. *Nat. Rev. Genet.* 21, 377–384. doi: 10.1038/s41576-020-0228-x

Humes, K. R., Jones, N. A., and Ramirez, R. R. (2011). *Overview of Race and Hispanic Origin: 2010. 2011 [cited 29 Oct 2020]*. Available Online at: https://www.census.gov/library/publications/2011/dec/c2010br-02.html (accessed October 29, 2020).

Jewett, E. M., Zawistowski, M., Rosenberg, N. A., and Zöllner, S. (2012). A coalescent model for genotype imputation. *Genetics* 191, 1239–1255. doi: 10.1534/genetics.111.137984

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443. doi: 10.1038/s41586-020-2308-7

Kim, S. K., Gignoux, C. R., Wall, J. D., Lum-Jones, A., Wang, H., Haiman, C. A., et al. (2012). Population genetic structure and origins of Native Hawaiians in the multiethnic cohort study. *PLoS One* 7:e47881. doi: 10.1371/journal.pone.0047881

Kirch, P. V. (1985). *Feathered gods and fishhooks: an introduction to Hawaiian archaeology and prehistory*. Honolulu: University of Hawaii Press.

Koh, X.-H., Liu, X., and Teo, Y.-Y. (2014). Can evidence from genome-wide association studies and positive natural selection surveys be used to evaluate the thrifty gene hypothesis in East Asians? *PLoS One* 9:e110974. doi: 10.1371/journal.pone.0110974

Kolonel, L. N., Henderson, B. E., Hankin, J. H., Nomura, A. M., Wilkens, L. R., Pike, M. C., et al. (2000). A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am. J. Epidemiol.* 151, 346–357. doi: 10.1093/oxfordjournals.aje.a010213

Kwan, P., Briand, G., Lee, C., Lepule, J., Llave, K., Pang, K., et al. (2015). Reservations to Participate in Biospecimen Research among Pacific Islanders. *Calif. J. Health Promot.* 13, 27–33.

Lim, E. T., Wurtz, P., Havulinna, A. S., Palta, P., Tukiainen, T., Rehnstrom, K., et al. (2014). Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet.* 10:e1004494. doi: 10.1371/journal.pgen.1004494

Lin, M., Caberto, C., Wan, P., Li, Y., Lum-Jones, A., Tiirikainen, M., et al. (2020). Population-specific reference panels are crucial for genetic analyses: an example of the CREBRF locus in Native Hawaiians. *Hum. Mol. Genet.* 29, 2275–2284. doi: 10.1093/hmg/ddaa083

Lipson, M., Skoglund, P., Spriggs, M., Valentin, F., Bedford, S., Shing, R., et al. (2018). Population Turnover in Remote Oceania Shortly after Initial Settlement. *Curr. Biol.* 28, 1157–1165.e7. doi: 10.1016/j.cub.2018.02.051

Liu, S., Huang, S., Chen, F., Zhao, L., Yuan, Y., Francis, S. S., et al. (2018). Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History. *Cell* 175, 347–359.e14. doi: 10.1016/j.cell.2018.08.016

Locke, A. E., Steinberg, K. M., Chiang, C. W. K., Service, S. K., Havulinna, A. S., Stell, L., et al. (2019). Exome sequencing of Finnish isolates enhances rare-variant association power. *Nature* 572, 323–328. doi: 10.1038/s41586-019-1457-z

Lohmueller, K. E. (2014). The impact of population demography and selection on the genetic architecture of complex traits. *PLoS Genet.* 10:e1004379. doi: 10.1371/journal.pgen.1004379

Madan, A., Archambeau, O. G., Milsom, V. A., Goldman, R. L., Borckardt, J. J., Grubaugh, A. L., et al. (2012). More than black and white: differences in predictors of obesity among Native Hawaiian/Pacific Islanders and European Americans. *Obesity* 20, 1325–1328. doi: 10.1038/oby.2012.15

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., et al. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538, 201–206. doi: 10.1038/nature18964

Martin, A. R., Atkinson, E. G., Chapman, S. B., Stevenson, A., Stroud, R. E., Abebe, T., et al. (2021). Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. *Am. J. Hum. Genet.* 108, 656–668. doi: 10.1016/j.ajhg.2021.03.012

Maskarinec, G., Erber, E., Grandinetti, A., Verheus, M., Oum, R., Hopping, B. N., et al. (2009). Diabetes incidence based on linkages with health plans: the multiethnic cohort. *Diabetes* 58, 1732–1738. doi: 10.2337/db08-1685

Maskarinec, G., Morimoto, Y., Jacobs, S., Grandinetti, A., Mau, M. K., and Kolonel, L. N. (2016). Ethnic admixture affects diabetes risk in native Hawaiians: the Multiethnic Cohort. *Eur. J. Clin. Nutr.* 70, 1022–1027. doi: 10.1038/ejcn.2016.32

Mathieson, I. (2020). Human adaptation over the past 40,000 years. *Curr. Opin. Genet. Dev.* 62, 97–104. doi: 10.1016/j.gde.2020.06.003

Mau, M. K., Sinclair, K., Saito, E. P., Baumhofer, K. N., and Kaholokula, J. K. (2009). Cardiometabolic health disparities in native Hawaiians and other Pacific Islanders. *Epidemiol. Rev.* 31, 113–129. doi: 10.1093/ajerev/mxp004

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., et al. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9, 356–369. doi: 10.1038/nrg2344

Merriman, T. R., and Wilcox, P. L. (2018). Cardio-metabolic disease genetic risk factors among Māori and Pacific Island people in Aotearoa New Zealand: current state of knowledge and future directions. *Ann. Hum. Biol.* 45, 202–214. doi: 10.1080/03014460.2018.1461929

Minster, R. L., Hawley, N. L., Su, C. T., Sun, G., Kershaw, E. E., Cheng, H., et al. (2016). A thrifty variant in CREBRF strongly influences body mass index in Samoans. *Nat. Genet.* 48, 1049–1054. doi: 10.1038/ng.3620

Moltke, I., Grarup, N., Jorgensen, M. E., Bjerregaard, P., Treebak, J. T., Fumagalli, M., et al. (2014). A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. *Nature* 512, 190–193. doi: 10.1038/nature13425

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

Nordyke, E. C. (1989). *The Peopling of Hawaii*, 2nd Edn. Honolulu: University of Hawaii Press.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature* 246, 96–98.

Palamara, P. F., Terhorst, J., Song, Y. S., and Price, A. L. (2018). High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nat. Genet.* 50, 1311–1317. doi: 10.1038/s41588-018-0177-x

Pike, M. C., Kolonel, L. N., Henderson, B. E., Wilkens, L. R., Hankin, J. H., Feigelson, H. S., et al. (2002). Breast cancer in a multiethnic cohort in Hawaii and Los Angeles: risk factor-adjusted incidence in Japanese equals and in Hawaiians exceeds that in whites. *Cancer Epidemiol. Biomarkers Prev.* 11, 795–800.

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

Posth, C., Nägele, K., Colleran, H., Valentin, F., Bedford, S., Kami, K. W., et al. (2018). Language continuity despite population replacement in Remote Oceania. *Nat. Ecol. Evol.* 2, 731–740. doi: 10.1038/s41559-018-0498-2

Rees, J. S., Castellano, S., and Andrés, A. M. (2020). The Genomics of Human Local Adaptation. *Trends Genet.* 36, 415–428. doi: 10.1016/j.tig.2020.03.006

Shriner, D. (2017). Overview of Admixture Mapping. *Curr. Protoc. Hum. Genet.* 94, 1.23.1–1.23.8. doi: 10.1002/cphg.44

Sidore, C., Busonero, F., Maschio, A., Porcu, E., Naitza, S., Zoledziewska, M., et al. (2015). Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat. Genet.* 47, 1272–1281. doi: 10.1038/ng.3368

Singh, G. K., and Lin, S. C. (2013). Dramatic Increases in Obesity and Overweight Prevalence among Asian Subgroups in the United States, 1992-2011. *ISRN Prev. Med.* 2013:898691. doi: 10.5402/2013/898691

Skoglund, P., Posth, C., Sirak, K., Spriggs, M., Valentin, F., Bedford, S., et al. (2016). Genomic insights into the peopling of the Southwest Pacific. *Nature* 538, 510–513. doi: 10.1038/nature19844

Speidel, L., Forest, M., Shi, S., and Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* 51, 1321–1329. doi: 10.1038/s41588-019-0484-x

Spratt, D. E., Chan, T., Waldron, L., Speers, C., Feng, F. Y., Ogunwobi, O. O., et al. (2016). Racial/Ethnic Disparities in Genomic Sequencing. *JAMA Oncol.* 2, 1070–1074. doi: 10.1001/jamaoncol.2016.1854

Stearns, S. C., Nesse, R. M., Govindaraju, D. R., and Ellison, P. T. (2010). Evolution in health and medicine Sackler colloquium: evolutionary perspectives on health and medicine. *Proc. Natl. Acad. Sci. U. S. A.* 107, 1691–1695. doi: 10.1073/pnas.0914475107

Steri, M., Orru, V., Idda, M. L., Pitzalis, M., Pala, M., Zara, I., et al. (2017). Overexpression of the Cytokine BAFF and Autoimmunity Risk. *N. Engl. J. Med.* 376, 1615–1626. doi: 10.1056/NEJMoa1610528

Sun, H., Lin, M., Russell, E. M., Minster, R. L., Chan, T. F., Dinh, B. L., et al. (2021). The impact of global and local Polynesian genetic ancestry on complex traits in Native Hawaiians. Lachance J, editor. *PLoS Genet.* 17:e1009273. doi: 10.1371/journal.pgen.1009273

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 590, 290–299. doi: 10.1038/s41586-021-03205-y

Tung, W. C., and Barnes, M. (2014). Heart Diseases Among Native Hawaiians and Pacific Islanders. *Home Health Care Manag. Pract.* 26, 110–113. doi: 10.1177/1084822313516125

Uren, C., Hoal, E. G., and Möller, M. (2020). Putting RFMix and ADMIXTURE to the test in a complex admixed population. *BMC Genet.* 21:40. doi: 10.1186/s12863-020-00845-3

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

Winkler, C. A., Nelson, G. W., and Smith, M. W. (2010). Admixture mapping comes of age. *Annu. Rev. Genomics Hum. Genet.* 11, 65–89. doi: 10.1146/annurev-genom-082509-141523

Zhernakova, A., Elbers, C. C., Ferwerda, B., Romanos, J., Trynka, G., Dubois, P. C., et al. (2010). Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. *Am. J. Hum. Genet.* 86, 970–977. doi: 10.1016/j.ajhg.2010.05.004

Zoledziewska, M., Sidore, C., Chiang, C. W. K., Sanna, S., Mulas, A., Steri, M., et al. (2015). Height-reducing variants and selection for short stature in Sardinia. *Nat. Genet.* 47, 1352–1356. doi: 10.1038/ng.3403

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

REVIEW
published: 05 November 2021
doi: 10.3389/fgene.2021.703901

An Overview of Strategies for Detecting Genotype-Phenotype Associations Across Ancestrally Diverse Populations

Irving Simonin-Wilmer[1], Pedro Orozco-del-Pino[2], D. Timothy Bishop[3], Mark M. Iles[3] and Carla Daniela Robles-Espinoza[1,4]**

[1]Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Campus Juriquilla, Queretaro, Mexico, [2]Biostatistics Department, University of Michigan, Ann Arbor, MI, United States, [3]Leeds Institute for Data Analytics and Leeds Institute of Medical Research at St. James's, University of Leeds, Leeds, United Kingdom, [4]Wellcome Sanger Institute, Hinxton, Cambridge, United Kingdom

Genome-wide association studies (GWAS) have been very successful at identifying genetic variants influencing a large number of traits. Although the great majority of these studies have been performed in European-descent individuals, it has been recognised that including populations with differing ancestries enhances the potential for identifying causal SNPs due to their differing patterns of linkage disequilibrium. However, when individuals from distinct ethnicities are included in a GWAS, it is necessary to implement a number of control steps to ensure that the identified associations are real genotype-phenotype relationships. In this Review, we discuss the analyses that are required when performing multi-ethnic studies, including methods for determining ancestry at the global and local level for sample exclusion, controlling for ancestry in association testing, and post-GWAS interrogation methods such as genomic control and meta-analysis. We hope that this overview provides a primer for those researchers interested in including distinct populations in their studies.

Keywords: GWAS, admixture, ancestry, PCA, regression

OPEN ACCESS

Edited by:
Mashaal Sohail,
University of Chicago, United States

Reviewed by:
Arslan A. Zaidi,
University of Pennsylvania, United States
Arjun Biddanda,
University of Oxford, United Kingdom

****Correspondence:***
Irving Simonin-Wilmer
isimonin@liigh.unam.mx
Carla Daniela Robles-Espinoza
drobles@liigh.unam.mx

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *01 May 2021*
Accepted: *14 October 2021*
Published: *05 November 2021*

Citation:
Simonin-Wilmer I, Orozco-del-Pino P, Bishop DT, Iles MM and Robles-Espinoza CD (2021) An Overview of Strategies for Detecting Genotype-Phenotype Associations Across Ancestrally Diverse Populations. Front. Genet. 12:703901. doi: 10.3389/fgene.2021.703901

1 INTRODUCTION

Genome-wide association studies (GWAS) aim to identify genetic variants (usually single-nucleotide polymorphisms or SNPs) that are associated with a phenotype of interest. GWAS have been highly successful at identifying genetic variants influencing a large number of traits, with nearly 5,000 publications and more than 250,000 variant-phenotype associations included in the GWAS Catalog (Buniello et al., 2019). Not only have GWAS improved our understanding of the aetiology of complex traits, identifying potential new biological pathways influencing phenotypes, but they are also of potential clinical value in assessing an individual's risk of developing particular phenotypes (e.g., Manolio (2013); Khera et al. (2018); Lambert et al. (2019)).

However, focusing only on participants of European descent, a characteristic of many published studies, restricts extrapolation to those of non-European ancestry (most notably for individual risk prediction (Mills and Rahal, 2019)) and limits available samples for traits common to multiple ancestries. By including populations with differing ancestries, the potential is enhanced for identifying causal SNPs or haplotypes because of the differing patterns of linkage disequilibrium

(LD) across subpopulations. Driven by the need to identify SNPs with even more modest effect sizes to further elucidate genetic architecture, GWAS sample sizes have necessarily increased; therefore, studies of a wider range of populations are warranted. In recognition of this, the proportion of studies including individuals of non-European descent has increased in recent years (Gurdasani et al., 2019). Such adaptations of study design require re-assessment of analytical approaches; when individuals from multiple distinct genetic ancestries are included in a study, it is necessary to implement a number of control steps to ensure that the associations identified are not detecting ancestry-driven rather than trait-related genetic effects.

One of the challenges of performing association tests on genomic data is that demographic history influences the genomic structure of the population being analysed. If this is not properly controlled for, any genotype-phenotype association found in the study may be a consequence of this structure, rather than genuine trait association. The source of this potential bias is known as population stratification, where different trait distributions within genetically distinct subpopulations will result in those markers associated with the ancestry of the subpopulation to be also apparently associated with the trait. As an illustrative example, Choudhry et al. (2006) analysed the relationship between ancestry-informative markers (SNPs with considerably different allele frequencies between Native American, African, and European ancestral populations) and asthma. They found that three of the 44 tested markers appeared to be related to the disease in Mexicans, but none of these associations persisted when ancestry was controlled for suggesting that the association is driven at least in part by ancestry. Therefore, it is of utmost importance to ensure that either all the individuals in a study are from the same ancestry prior to performing a GWAS or that this ancestry is appropriately taken into account in the analysis.

Depending on the populations being studied, analysis may not be as simple as identifying subpopulations in the samples, since each individual may be descended from multiple subpopulations tracing back to a mixture event (or admix event) between them. One of the ways in which we can express this mixing in an individual is as a function of ancestral populations; that is, populations that have been isolated from each other in the past (e.g., European and African). If the combination of these ancestral populations has been recent, then we expect to observe longer LD tracts; but these will decay over time (Montana and Pritchard, 2004), thus adding to the complexity of finding significant relationships. However, the more diverse linkage disequilibrium structure also gives the possibility of finding more nuanced, ancestry-specific signals in a GWAS. The purpose of this review is to discuss the main approaches that are used in order to account for population structure in admixed individuals in a GWAS to select data to include, control for its influence on findings, and compare or aggregate results across populations.

In order to provide an understanding of the methods used for the analysis of admixed populations, we will first review the steps involved in performing a GWAS. Secondly, we will discuss some of the methods used in recent years to study admixed populations, and the way in which each methodology has been applied. Here, we will both explain the rationale behind each methodology and give some examples of applications in recent studies.

2 CONTROLLING FOR POPULATION STRUCTURE IN GENOME-WIDE ASSOCIATION STUDIES

For the purposes of this review, we will divide a GWAS into three steps:

1) Quality control. (QC). This first, critical step involves filtering poor quality germline DNA samples and inconsistently performing SNPs from further consideration. This consists on applying specific filtering criteria to samples and/or SNPs before proceeding.
2) Association testing. Once QC has been completed, a statistical test is performed with the aim of detecting association between variants in the genome and the trait under consideration.
3) Post-GWAS interrogation. Once candidate SNPs have been identified, other types of analyses are performed to ensure the integrity of the association testing including that the influence of genetic structure has been well controlled for and to explore the characteristics of the SNPs identified including for instance biological processes implicated.

In steps 2 and 3, there are ways in which population structure can be taken into account, but it is important to note that we can use more than one technique on a single GWAS; in fact, they are often combined to avoid spurious associations.

In order to illustrate the use of these methods, we sampled data using the 1,000 Genomes Project (Consortium, 2015) dataset. We decided to use this dataset because of the self-reported ancestry label of the samples; these are useful for visualizing and comparing different methods.

3 ESTIMATING POPULATION STRUCTURE

The next subsection will cover two methods that are helpful in investigating the ancestry for each of the individuals in our data. These methods will be present throughout the review and will become useful for both quality control and genotype-phenotype association testing. The first one is admixture analysis, which assumes the existence of discrete ancestral populations from which the current population is derived. The second is principal component analysis, which generates explanatory variables from the genotype data that summarise the sources of variation among the samples and helps visualise and interpret the genetic structure of the samples.

3.1 Ancestry Estimation

Ancestry estimation aims to divide an individual's genome between multiple ancestral populations from which it is

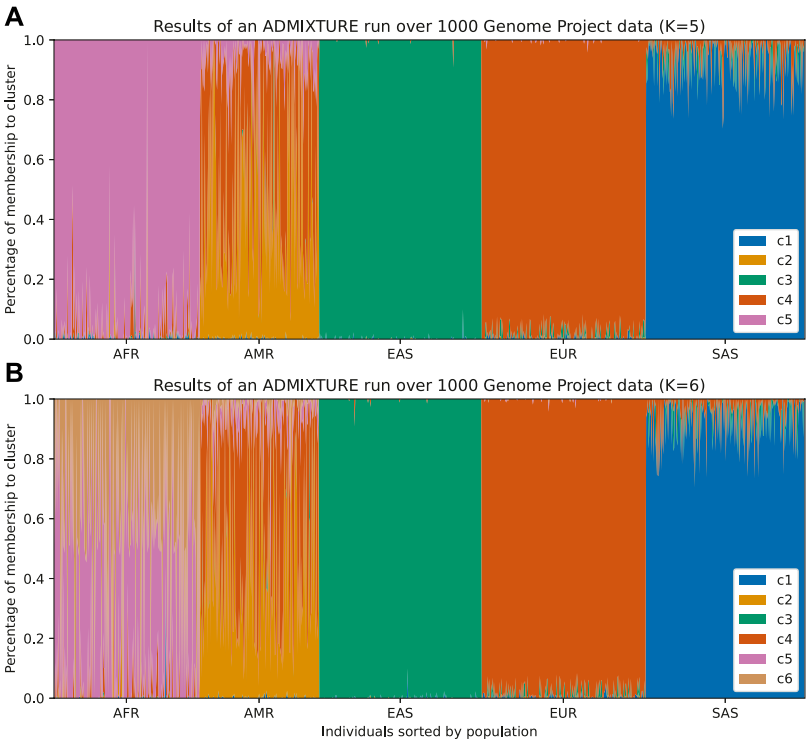

FIGURE 1 | Individuals from 1,000 Genomes Project are plotted according to their labeled self-reported ancestry (AFR, African; AMR, Ad Mixed American; EAS, East Asian; EUR, European; SAS, South Asian). **(A)** Results from an ADMIXTURE analysis with K = 5 (number of clusters). The colors represent the clusters inferred from the data. In this figure, we can infer that c1 corresponds to South Asian ancestry, c3 to East Asian ancestry, c4 to European ancestry, and c5 to African ancestry. The Admixed American population appears as the most varied across clusters and has an exclusive cluster (c2), which suggests that there is a mix of *native* ancestry and influx from Africa and Europe. **(B)** By running ADMIXTURE with K = 6 we can appreciate similar results. The extra cluster indicates further structure within the African population, which could be either from admixture or the existence of subpopulations in the African samples, but the rest remains unchanged.

hypothesised to have descended. Most methods used here follow a clustering approach, where each allele is assumed to have a probability of coming from one of the ancestral populations; these methods involve assessment of a large number of SNPs to estimate the contributions of each ancestral population. It is important to differentiate between two distinct forms of ancestry estimation: global and local (Thornton and Bermejo, 2014). Local ancestry is based on the fact that genetically adjacent regions form haplotypes whose ancestry can be probabilistically aligned to each population. There are local ancestry methods based on a model of recent admixture, and others that can infer gene flow from ancient hominids (Sankararaman et al., 2016; Durvasula and Sankararaman, 2019; Hubisz et al., 2020). The aim of global ancestry is to estimate the contribution, overall, of the genome from each ancestral population rather than each precise genomic region.

3.1.1 Global Ancestry

The main assumption for this estimation is that a given individual is descended from ancestors drawn from distinct ethnic groups. The result of an analysis of this kind is an estimation of the proportion of each individual's genome that comes from each of the ancestral populations.

The two most popular algorithms for global ancestry calculation are STRUCTURE (Pritchard et al. (2000); Falush et al. (2003); Porras-Hurtado et al. (2013)) and ADMIXTURE (Alexander et al., 2009). Both of these algorithms require choosing the number of ancestral populations a priori and modeling the probability of membership to each ancestral population. STRUCTURE assumes a Bayesian model that accounts for linkage disequilibrium within each ancestral population, whereas ADMIXTURE assumes linkage equilibrium and uses the unlinked SNPs to apportion ancestry; this is a practical observation since an extra step will be required to run ADMIXTURE by thinning the SNPs to create this set of "independent" SNPs. The results can be visualized in an admixture plot, which shows the percentage of each subpopulation (given by the cluster) that the model assigns to each individual in the sample (**Figure 1**, and **Figure 2A**). While these methods return "estimates" of ancestry, care must be taken not to overinterpret these results in terms of alignment with population history.

3.1.2 Local Ancestry

Although global ancestry uses unsupervised methods such as clustering, local ancestry is more restricted as it requires a locally recruited reference panel, enabling the estimation of the locus-

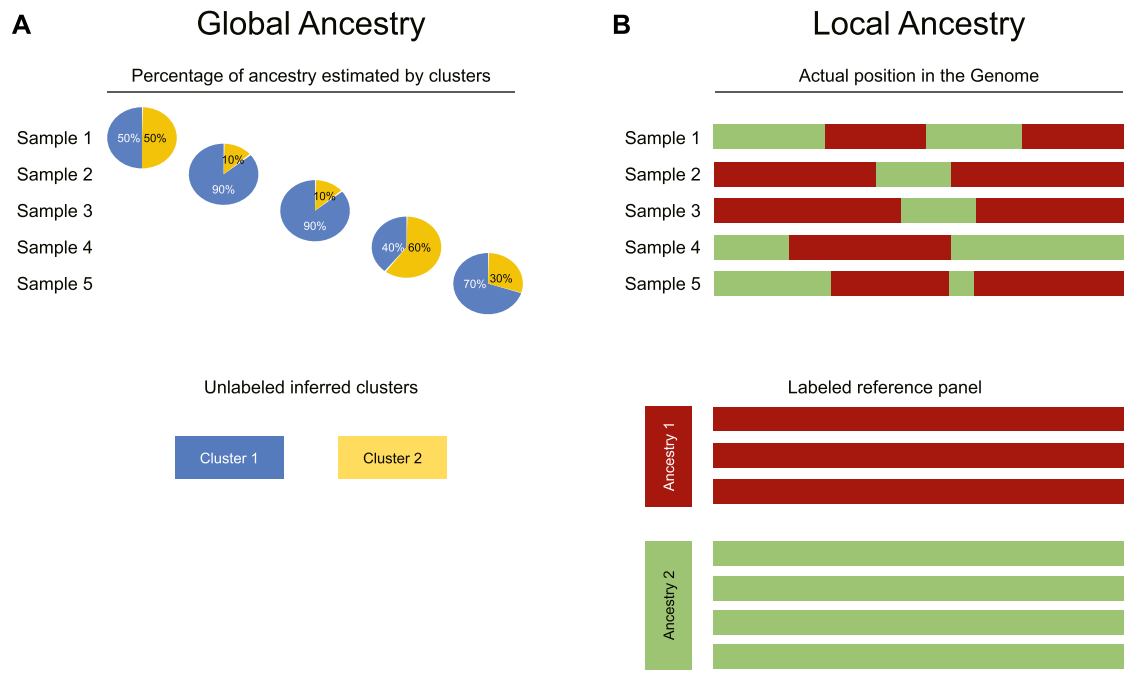

FIGURE 2 | Differences between global and local ancestry analysis. **(A)** Global ancestry analysis infers the proportion of each individual's genome that comes from each ancestral population (represented as clusters identified in an unsupervised manner). **(B)** In contrast to global ancestry analysis, local ancestry inference uses a reference panel to attribute each physical segment of the genome to a specific ancestry reported in the panel.

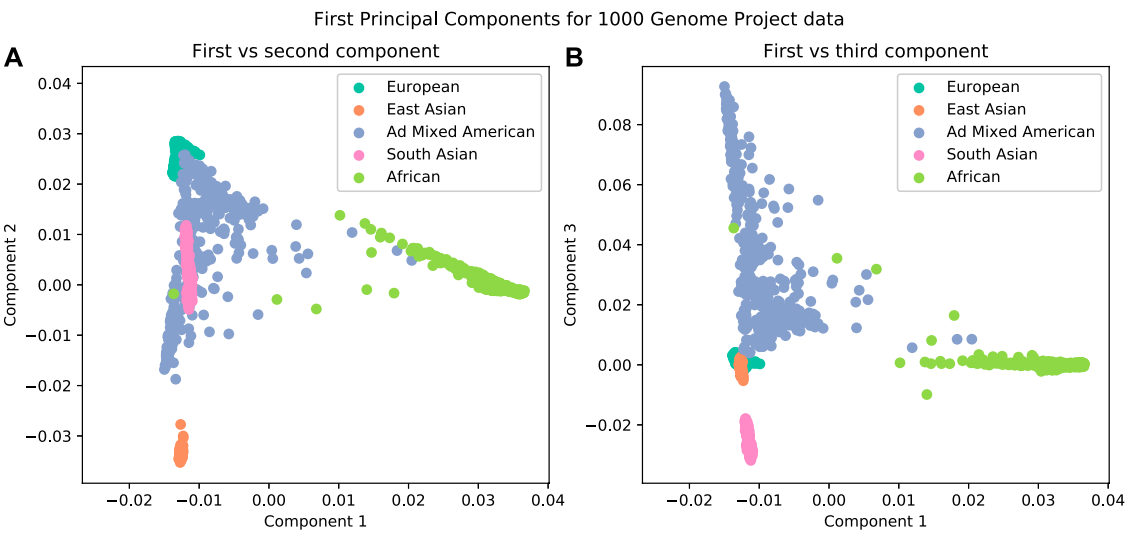

FIGURE 3 | The first three components from a Principal Component Analysis on data from the 1000 Genomes Project. A clear separation is observed when plotting the individuals by their components. **(A)** First vs second components. **(B)** First vs third components.

specific likelihood of ancestry. In other words, for each SNP, the ancestral population from which it has most probably been inherited is calculated (**Figure 2B**). If the estimation is correct, this analysis achieves global ancestry estimation too.

Although there are several packages to infer local ancestry, there are two that are most commonly used. The first one is RFMix (Maples et al., 2013), which adjusts samples to a reference panel of known ancestries through a random forest procedure. The second algorithm is implemented in the software LAMP-LD (Baran et al., 2012), which uses Hidden Markov Models to relate the linkage disequilibrium in the population to a set of reference haplotypes.

3.2 Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction method that finds the directions in the variable space under study that explain the most variance; these directions are called the components. In the case of genotype data, each SNP can be represented with values 0, 1 or 2 depending on the dosage of the alternative allele (aa, Aa, AA respectively, with "a" referring to the reference allele and "A" to the alternative). In this way, a data matrix can be created that has individuals in rows and SNPs in columns. From this matrix, we can compute the components. Each component is orthogonal to the others so they can be used, for example, to visualize the highly dimensional genotype data used in GWAS.

It has been observed that the first few principal components from genotype data are related to population structure (**Figure 3**). The advantage of using this method over admixture analysis is that PCA results in a more nuanced view of the genetic structure of the sample, given that there is no need to specify the number of ancestral populations. A number of distinguishing characteristics can be appreciated when 1,000 Genomes data are plotted in this way; for example, the admixed American population overlaps with other populations in the first two principal components; this illustrates the admixture in those individuals (**Figure 3A**). But if further components are examined (**Figure 3B**), there is a clear separation of the American population from others.

PCA is a widely used method in different disciplines, so its implementations are abundant. Some of the more popular software for genotype data are the PLINK (Purcell et al., 2007) --pca method, EIGENSOFT (Price et al., 2006), and the SNPRelate (Zheng et al., 2012) package for the R programming language. Results from different PCA implementations should not differ; however, given the complexity and size of genetic data, specialized bioinformatic software such as PLINK is usually preferable to more generic statistical software.

4 QUALITY CONTROL

In addition to estimating structure within the samples in our study, we also need to identify the individuals and genomic markers that are appropriate for our study. The first set of criteria that we can use to select our data corresponds to the task of spotting genotyping errors. These criteria are discussed in more depth in several reviews, as well as in original published research, and include missingness (applied to SNPs and samples), case-control differential missingness and tests for heterozygosity and Hardy-Weinberg equilibrium, and strand alignment checks when multiple datasets are involved (Turner et al., 2011; Medina-Gomez et al., 2015). Quality control is of particular importance when combining data from several sources in order to avoid confounding batch effects. However, there are some caveats that need to be considered when applying these criteria, because even though they are standard in homogeneous randomly mating populations they may not be appropriate in structured populations.

- Missingness. This includes removing SNPs that may give misleading results due to genotyping errors across many samples, or samples that have an excess of errors in the genotyping process and too few high-quality SNP.
- Strand alignment. Since DNA is double stranded, it is important to report (and compare) equivalent strands in the data; this can be a problem when merging data from different sources since there can be discrepancies in the reported strand. For example, the Illumina platform differs in definition on the concept of strand from the standard human genome reference (Zhao et al., 2018). It is important then to align the samples to the same strand. This becomes specially difficult in circumstances such as when the strands have complementary alleles (AT/CG). If these kind of uncertain SNPs are not too frequent in the data, it is probably better to remove them, since they can bias the results.
- Heterozygosity. In a homogeneous randomly mating population, very high or low levels of heterozygosity can indicate poor quality genotyping. However, this test is not appropriate in a non-randomly mating population, because population structure can lead to extremes of heterozygosity (Boca et al., 2020).
- Deviation from Hardy-Weinberg (HW) equilibrium. This test, standard in population studies, evaluates whether the expected relationship between allele frequency and genotype frequency exists. However, HW equilibrium assumes that there is random mating in the population under study; so if there are clear subpopulations (different ancestries) the conditions are not met and the test is not valid as a criteria for assessing quality. Therefore, this test is not generally recommended to use directly when studying structured populations. If the populations are labeled (e.g. we have data from different, clear sources) then it is better to apply HW tests separately.

The second set of criteria we can evaluate with genotype data can elucidate the ancestry of the individuals in the study. For this set we can use the methods we described above: admixture analysis and principal component analysis. There are two ways in which these are used as part of quality control:

- Firstly, individuals whose ancestry is not well represented in either cases or controls should be removed. In the case of a continuous trait this is equivalent to removing outliers. This avoids ancestry-specific biases in the association test, but is not expected to affect the variability of the data ancestry-wise.
- Secondly, if distinct populations (e.g. African, European, Asian) are represented across the phenotype, the study can be partitioned over these distinct populations. This would allow us to obtain multiple association tests, the results of which can subsequently be combined (see the post-GWAS Interrogation section). This method reduces ancestry-related variability and bias of each of the studies but decreases the amount of data in each of them, diminishing the statistical power of each test.

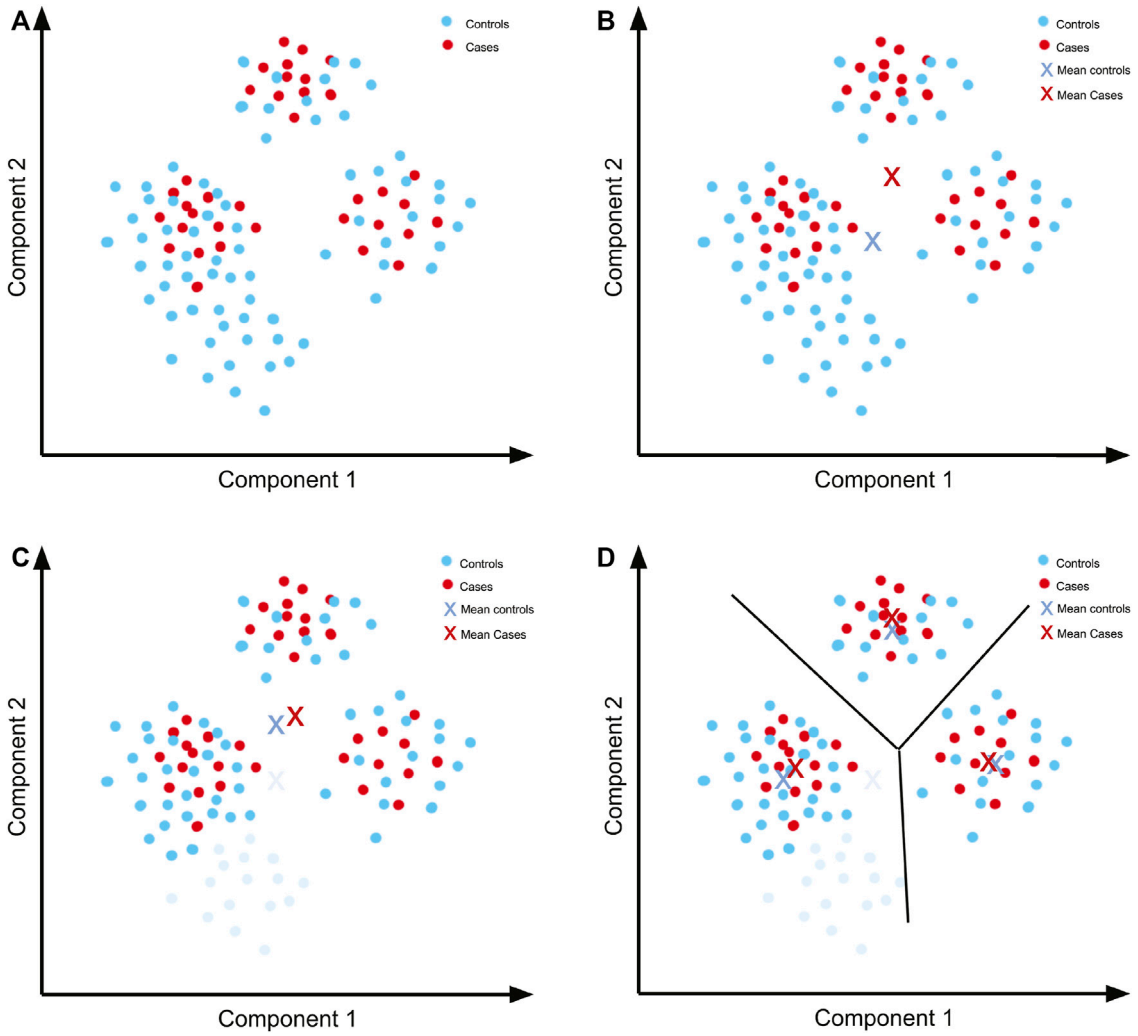

FIGURE 4 | Quality control previous to performing a GWAS. **(A)** Hypothetical PCA plot for a simple dataset. There are three separate populations including both cases and controls. **(B)** The means of the components for cases and controls are not close because there is a concentration of controls in an area with no representation from cases. **(C)** By removing the controls that are not well represented in the cases (light blue), we can get a mean that is closer to the mean of the cases. **(D)** The means are closer when separating by the observed populations, so these can then be analysed separately and the results subsequently combined.

In case-control studies in particular, the selection of controls is a crucial step. If there is a factor that can influence the outcome (in our case the phenotype) in some way other than the variable that we are measuring (the genotypes), then it must be accounted for in the experimental design. As an illustrative example, in a trial for testing a new drug, there may be covariates (such as sex or age) that should be controlled for in order to ensure that the effect of the drug versus a placebo is measurable; e.g. age and sex may influence the outcome variable due to, for example, metabolism changes and hormone differences. One option to control for these covariates is randomizing which patients will receive the drug. What this procedure does is ensure that the distribution of age and sex between cases and controls is effectively the same, so the influence of these variables does not influence our observation of the drug effects. In GWAS studies, the distribution we want to keep consistent between cases and controls (or across the continuous trait) is the ancestry. In the following example we will use the first two principal components to illustrate this.

4.1 A Motivating Example

In order to develop a feeling for what quality control means in a GWAS, imagine a simple dataset (**Figure 4A**) to which PCA was applied and for which only the first two components are relevant to account for population structure.

Since the principal components represent a factor that we want to control (ancestry/ethnicity), we need a similar distribution of the components in both the cases and controls. We can further simplify the example by summarising the distribution using the mean (**Figure 4B**). Even by using only the mean of the data, it is evident that the distribution of controls

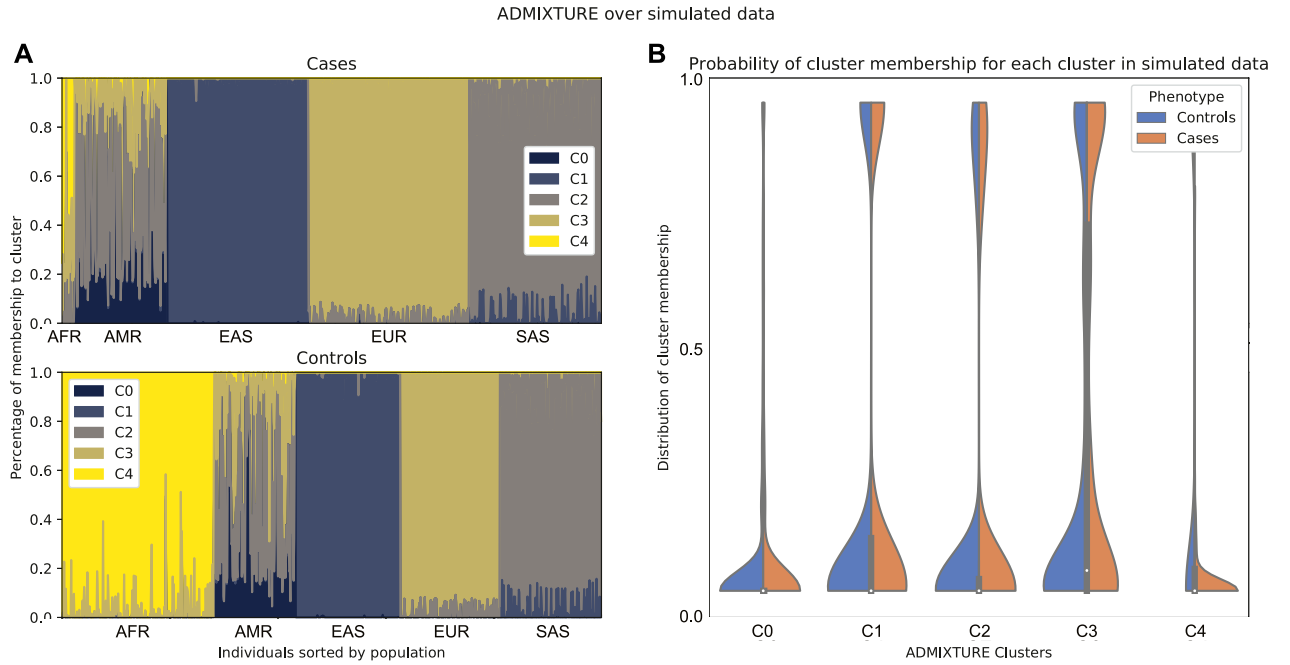

FIGURE 5 | Comparison between admixture of cases and controls of simulated data according to a run of ADMIXTURE with K = 5. In these data, there is a super population that is underrepresented in the cases. **(A)** Admixture plot sorted by population label on 1,000 Genomes. **(B)** Density plot for cases and controls on the probability of membership to each of the clusters of the ADMIXTURE run. Note that here, cluster c4 corresponds to the African population.

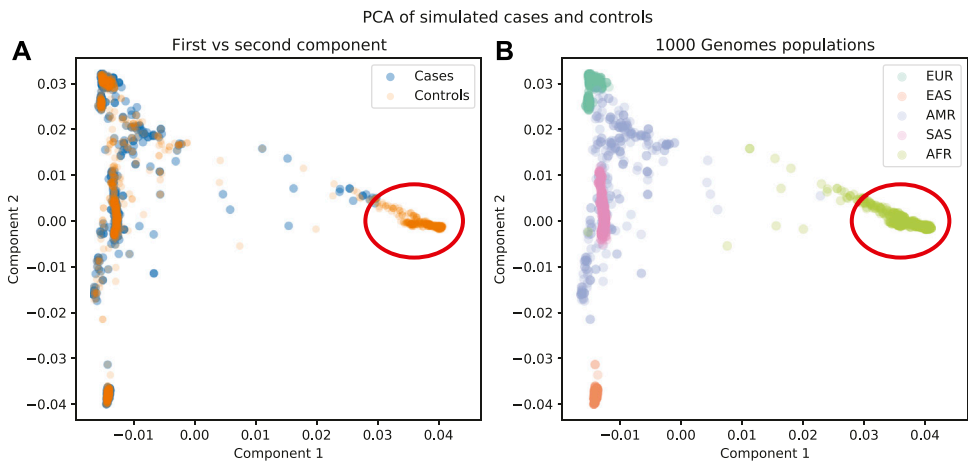

FIGURE 6 | (A) First components plotted for cases and controls. **(B)** First components plotted for 1000 Genomes population labels. There is a region in these data that is not well represented in the cases as well as in the controls (red circle).

does not follow the same distribution as the cases. A simple solution is to remove the controls with components that are unrepresented in cases (**Figure 4C**). The means of the cases and controls are now more similar, although not identical. We can further seek a better fit of the distributions by separating the populations according to the clusters that we can see in the plot.

Once this cluster separation has been done, there is a better fit in the distributions in each of the three sets of cases and controls (**Figure 4D**). Although for each of the association tests there will be less data to work with, and so less statistical power for each test, we can overcome this issue later via meta-analysis.

4.2 Comparability of Cases and Controls

In order to illustrate this approach, we up-sampled 2000 individuals from the 1000 Genomes Project dataset, removed a number of genotypically similar samples and assigned a fictitious

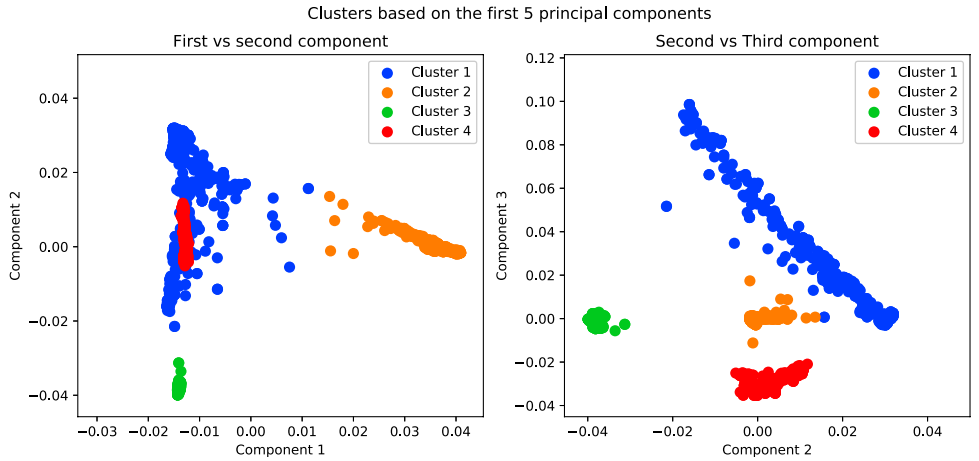

FIGURE 7 | Results of running a K-Means algorithm on the first 5 components of the simulated data with K = 4. The purpose of this analysis is to separate the data into subpopulations in order to conduct association tests within each.

case-control status to each of these in order to make the usefulness of the method more obvious. An ADMIXTURE run on the data shows a cluster that is underrepresented in the cases (**Figure 5**). This means that there is a super population from which almost no cases with the phenotype were sampled. In this case, it is a good idea to remove from the data the individuals from that population.

A PCA run on these data shows that there is a region of the plot where there are no cases, so the appropriate step would be to remove the individuals from that region (**Figure 6**). It is notable that if we just used the cluster results from the admixture analysis, cluster c4 would be a candidate for removal, but with PCA we find more nuanced criteria for the decision.

4.3 Separating the Data for Multiple Association Studies

If the clustering (in admixture analysis) or the separation (observed in PCA) is clear, such as in our sample data, it is preferable to analyse the populations in separate datasets as they may have different patterns of linkage disequilibrium. This can be useful to make statements about SNPs associated to the phenotype that are specific to subpopulations. However, if there are some population-specific signals for the tested trait, they may be lost in the subsequent meta analysis. If there are no distinct clusters, it is considered better to analyse the combined data in the association test (Begum et al., 2012).

In order to separate the data, in admixture analysis we can choose for each individual the cluster for which the probability of membership is maximised as its cluster. For PCA, we can use a clustering method on the first n components (**Figure 7**).

5 ASSOCIATION TEST

Once we have performed quality control of the samples and SNPs, and have chosen those to include in the analysis, as well as the number of separate population clusters we will be analysing, then we are ready to proceed to the identification of SNPs that are associated with the phenotype being tested. There are several ways to find candidate causal SNPs from genotype data, such as hypothesis testing and linear model-based approaches. In order to account for population structure, linear models are most widely used.

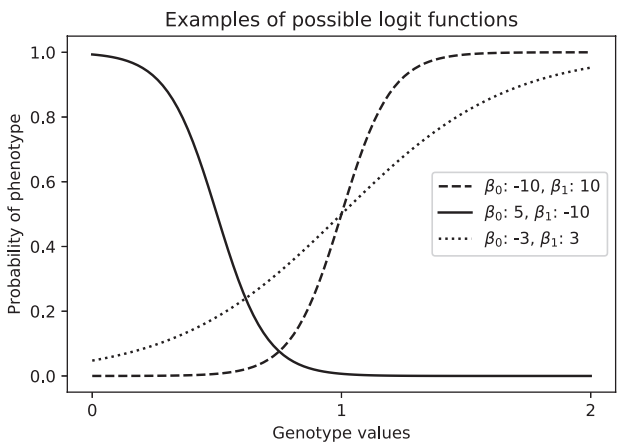

FIGURE 8 | Examples of estimated probability of phenotype given a genotype value, which is coded numerically as the number of *alternative* alleles (aa = 0, Aa = 1, AA = 2).

5.1 Methods to Perform Association Testing

5.1.1 Logistic Regression

In the case of case-control studies, phenotypes are binary, and so we can use logistic regression. This model consists on assuming a linear relationship between independent variables and the log-odds, which represents the logarithm of the ratio of the probability of being a case over the probability of being a control conditioned on the covariates. That is, for two independent variables x_1 and x_2,

$$\log\frac{p}{1-p} = \beta_0 + \beta_1 x_1 \quad (1)$$

Where p is the probability of being a case, β_0 is the intercept and β_1 is the effect size for the covariable x_1. If we have more than one covariable, we can add more terms $\beta_2 x_2$, $\beta_3 x_3$, ... The model in **Eq. 1** can yield results such as we see in **Figure 8**.

In any case, the logistic regression is performed on a locus-by-locus basis. This yields parameters with its respective p-value for each SNP. We will now discuss two methods to control for ancestry in the association test: via PCA and via admixture mapping. The difference between these methods lies in what independent variables are used in the logistic regression.

5.1.2 Mixed Models

Mixed models are an extension of linear models that allow us to include effects that account for dependency between data points. For example, in the case of genetic studies, the data points are the individuals, and the dependency can be thought of as being the ancestry.

The model for a mixed effects regression for association testing can be written as follows

$$f(x) = G\beta + \nu + X\gamma \quad (2)$$

Where the first term on the right side of the equation is the same as any linear model: the independent variables and the parameters; these are called the fixed effects and in the context of genetic association it is the genotype as described in the logistic regression section. The last term is the covariates (e.g., the first principal components, sex, etc). The second term represents the random effects, which model the error just like any other regression model, but in this case, the error is not equally distributed for every observation. Usually, we would say that the error follows a Normal distribution centered on zero with a fixed variance $N(0, \sigma^2)$; but in mixed models we say that $\nu \sim N(0, \tau Z)$, where τ is a parameter for Z, the matrix of random effects. Z is usually the genetic relationship matrix, which estimates the degree of sharing of identity by descent (IBD) between all pairs of individuals in the dataset, but it can also be a matrix of categories where each row (sample) is a vector of zeros everywhere except in the columns that represent the subpopulation to which it belongs (e.g. from Admixture analysis).

This is a general definition of mixed models, but there are several particular implementations based on variations of **Eq. 2** and in particular of matrix Z such as EMMA (Kang et al., 2008), FaST-LMM (Lippert et al., 2011), GCTA-LOCO (Yang et al., 2014) and some Bayesian modelling versions like BOLT-LMM (Loh et al., 2015).

5.2 Controlling the Association Model for Ancestry

In the association test, we can model each locus as an independent variable with values 0, 1 or 2 depending on the dosage of the alternative allele (aa, Aa, AA respectively) with the trait being measured as a dependent variable. This model allows us to add other covariables; in particular, we can use the first principal components from the genotype PCA. Since the components absorb information about the ancestry, the model will only give significance to the SNPs that are related to the trait without the confounding of the population structure captured by the PCs included in the model.

One way of determining how many components to use consists in plotting the components until no separation is found in the data. In our 1000 Genomes example, there is clear separation of the individuals in the scatterplot between components one through four, but the direction of the fifth component is reaching for a subset of less than 1% of the data (the few points with the component 5 greater than 0.4), so it is not accounting for a significant amount of ancestry-related variability (**Figure 9**). So in this case, we would probably be safe in controlling by using only the first 4 components. This is a simple example, but it is useful in practice to visually review the interaction between the components to get a grasp of the structure of the data. For a more automated and statistically sound procedure, the software EIGENSOFT provides methods to infer the statistically significant number of components for population structure by evaluating the significance of each component iteratively according to the variance explained by each (Patterson et al., 2006).

An example of this usage can be found in Nannini et al. (2017). The authors use PCA to compare their Latino population with a reference panel of Europeans and Africans. They also determine that using four principal components in their regression is enough to control for population stratification. Another interesting example can be found in Costa-Urrutia et al. (2019), where authors control not by the principal components, but for the proportion of Amerindian ancestry estimated via ADMIXTURE.

As for local ancestry, in Wang et al. (2011) the authors propose controlling each of the tests by their respective estimated local ancestry. However, this method is not widely used, as it has been argued that the bias introduced by using only global ancestry is small (Martin et al., 2018). The methods that we discuss below exploit the advantages of local ancestry more directly.

5.3 Admixture Mapping

Admixture mapping is motivated by the scenario of recent mixing of populations which occurs alongside discrepant incidence of the trait between two populations (i.e. a difference in the proportion of affected people between the ancient populations). Affected persons in the admixed population should therefore be expected to have preferentially inherited the risk locus from the higher incidence population (Patterson et al., 2004). The genome-wide approach is to examine each region of the genome systematically to identify regions where affected persons inherit a statistically higher proportion of their alleles from the high risk population than the overall pattern of inheritance for that person.

This method relies on the assumption that the phenotype-associated alleles have different frequencies across ancestral populations. This extra requirement helps specify a model with more statistical power to find these specific loci, so in this way fewer SNPs (and since this implies lower burden of tests, also fewer samples) are needed to find associations. However, this means that it will fail to identify all risk loci; since not all causal SNPs follow this pattern. Also, fewer loci means longer LD tracts and so a higher difficulty in identifying causal markers via fine mapping (Seldin, 2007).

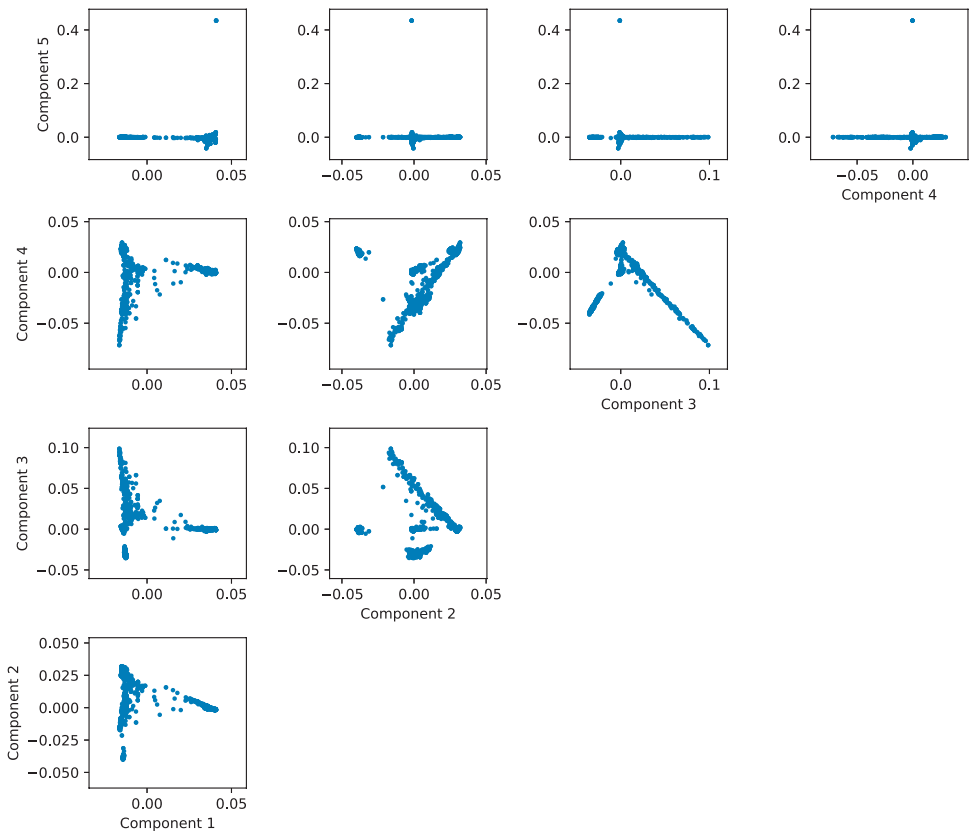

FIGURE 9 | Matrix of scatterplots between the first five principal components of simulated data based on 1,000 Genomes.

Admixture mapping has been successfully used to identify risk loci associated with specific ancestries across different traits; the tools and panels necessary for performing these kinds of analyses were developed in early 2000s. In 2005, the first applications of this seminal method were published, focusing on the study of African American individuals and finding a number of ancestry-specific associated loci (i.e., either European or African) to the traits: Zhu et al. (2005) found that excess African ancestry at 6q24 and 21q21 was associated with hypertension, and Reich et al. (2005) identified a European-derived locus in chromosome 1 associated to multiple sclerosis. Later, Freedman et al. (2006) identified that excess African ancestry at the 8q24 locus is associated to increased risk of prostate cancer.

More recently, Wang et al. (2019), used admixture mapping to find loci related to several traits used to measure sleep apnea; this study was performed on Latinos and found three novel regions associated with this condition. In another study in the Latino population, Burkart et al. (2018), identified genomic regions associated with lung function and chronic obstructive pulmonary disease, some of them previously undiscovered. In both of these studies, some of the risk loci found were replicated in Europeans, which illustrates the advantage of using samples from admixed populations.

As mentioned above, ADMIXTURE and STRUCTURE take different approaches to estimate a person's proportion of genome inherited from an ancestral population (global ancestry). If, as computed using either of these approaches, the average proportion of genome from the higher risk population is estimated as θ for a study participant, then the genome-wide analysis is conducted for each participant by examining their actual inheritance at each SNP from this average across the genome. The calculation of the actual number of alleles at this SNP that have ancestry from the high risk subpopulation requires some discussion (local ancestry). Analysis of a single SNP will often be uninformative in terms of identifying the ancestral origin of each allele so instead the approach required is to use SNPs in proximity to the SNP under consideration to estimate the actual number of alleles from the high risk subpopulation (McKeigue, 1998).

If x is the estimated number of alleles at an SNP that have ancestry from the high-risk subpopulation (0, 1, 2) for a person, then given θ and p, the prevalence of the disease (0.5 with equal number of cases and controls), we can fit the logistic regression model from **Eq. 3** (Hoggart et al., 2004):

$$\log\frac{p}{1-p} = log\frac{\pi}{1-\pi} + \left(\frac{x}{2} - \theta\right)\beta \tag{3}$$

Where β is the odds ratio for having 2 copies of the risk allele versus 0 in the high risk population. In the formula, the left

hand term is the log odds of the trait. The right hand term of the equation has two components: the first one reflects the prevalence of the disease in log odd terms, and the second models genetic risk considers deviation from the average genotypic contribution from the high risk population for that person.

One extra advantage of admixture mapping is that, since this model examines ancestry at each SNP with the average across the genome for that person, there is an alternative test that can be done without controls (the so called "case only study"). It involves testing whether there is an increased risk according to the local ancestry in a given SNP. However, in practice, power is usually greater for the case-control comparison.

One widely used software to run admixture mapping can be found in the GENESIS package for the R programming language via the admixMap function.

5.4 Local Ancestry Regression

A novel approach is using the inference of local ancestry directly in the association testing. The software Tractor (Atkinson et al., 2021) implements the following regression model for each locus:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k \qquad (4)$$

Where every β_i are the effect estimates, X_1 is the admixture proportion from the first ancestry, X_2, X_3 are the number of copies of the *alternative* allele coming from the first and second populations respectively (aa = 0, Aa = 1, AA = 2), and after that we can add any number of covariates such as age or some PCA components. This model allows for the inclusion of ancestry specific information, and in that way it results in relevant summary statistics related directly to each of the populations of the admixture. This model accounts only for two ancestries, however the model can be expanded to several ancestries.

Having a parameter associated to the ancestry in a given locus prevents the association model from attributing an effect to the allele count that is better explained by the ancient population from which the haplotype is descended. This avoids bias caused by local ancestry differences between populations that are not attributable to the trait (Atkinson et al., 2021).

In addition, although this method is analogous to controlling via PCA in the sense that we are controlling for ancestry, this type of regression achieves this by analyzing the ancestry of each specific locus at a time. This allows us to add samples without worrying about introducing population structure, which then translates into more statistical power.

These ancestry specific parameters provide information on ancestry predisposition to the trait. In contrast to admixture mapping, this method does not assume that the phenotype incidence differs across the ancestral populations. In this model, however, it is necessary to have data on both cases and controls.

6 POST-GENOME-WIDE ASSOCIATION STUDIES INTERROGATION

Association tests are performed on a SNP by SNP basis, so after the candidate SNPs have been identified, it is important to use techniques that help us validate the adequacy of our population adjustments in the previous steps. The technique of genomic control will allow us to evaluate whether the association test has a bias based on population structure. Performing a meta-analysis will allow us to combine the results of the different populations if we previously decided to separate by subpopulations in the quality control step.

6.1 Genomic Control

This method corrects the test statistics (p-values) obtained from the association analysis based on a single number, usually called the genomic inflation factor (Pritchard and Rosenberg, 1999) and denoted as λ. The inflation factor is calculated using the genetic markers that are not related to the disease, and it consists in testing whether there is a consistent difference between the allele frequencies in cases and controls across the genome.

This factor can be interpreted as follows: If $\lambda = 1$, there is no population stratification, and values greater than 1 indicate that there is structure unaccounted for in the study. However, in large well-powered studies, the inflation that this factor measures could be coming from a different source, such as polygenicity. For a more nuanced approach we can use LD score regression (Bulik-Sullivan et al., 2015), which leverages the relashionship between the SNP in question and those around it to discriminate the source of the inflation.

Even though the inflation factor can be used to correct for population stratification, it is not generally recommended to do so (Shmulewitz et al., 2004), as it is particularly ineffective in highly admixed data. It is however useful for identifying the presence of inflation in order to evaluate whether the methods in previous steps of the analysis were sufficient to account for population structure (Galanter et al., 2014; Conomos et al., 2016; Hodonsky et al., 2017; Jorgenson et al., 2017; Nannini et al., 2017).

6.2 Meta-Analysis

The meta-analysis is not in itself a method for correcting for population structure, but it is employed to analyse GWAS results from different populations. If we used the methods discussed in the Quality Control section to separate our individuals and performed one association test for each of those subpopulations, we can perform a meta-analysis to aggregate their results. This will help us regain statistical power lost by the reduced sample sizes of each study; the power is of course reduced if the effects are specific to some subpopulation, and this will be true no matter the analytical approach.

The results that we intend to aggregate from the studies are the effect sizes (γ) for the trait. However, since factors such as sample size can influence the existence of different levels of uncertainty on each study, we must have a measure available to assess uncertainty. For this purpose, having also the standard error will allow us to perform an inverse variance-weighted meta-analysis; which means that we are using the variance of the estimator to weigh in the uncertainty found in each of the studies before performing the meta-analysis.

The first model we can use is to use a fixed-effects-only model. This assumes that all of the effect sizes across all studies are the same, and the differences between them are the product of a normally distributed random error (ϵ).

$$\gamma = \beta + \epsilon \tag{5}$$

Another possible model would be to use a random-effects-only model. This is applied when we suspect that the underlying effect size varies between studies, for instance due to different patterns of linkage disequilibrium or gene-environment interactons.

$$\gamma = \theta + \mu_i + \epsilon \tag{6}$$

Where θ is the true effect size, and μ is the within study variance that will be estimated from the data (Kelley and Kelley, 2012).

The difference between the two models then, is that in the fixed effects model we are assuming that there is a single, true effect size across all the studies, and we are trying to find whether this true effect size is different from zero. In the random effects model we are assuming that there is a distribution of random effects, and we are trying to find whether the mean of the effect sizes is different from zero.

The fixed effects model assumes that there is no heterogeneity between the effects in the different studies being combined, this can be tested by referring to Higgins and Thompson (2002), where they propose a metric I^2 that measures the proportion of variation between studies that is due to heterogeneity. They propose as a rule of thumb that with an $I^2 > 30\%$ we should consider using random effects instead of fixed effects. The fixed effects model provides more power, but it is important to examine its appropriateness before enjoying its benefits.

Jorgenson et al. (2017) provide an example of a study with different ethnicities (Non-Hispanic Whites, Hispanic/Latinos, East Asians, and African Americans) where authors decided to separate the analysis into different studies and used meta-analysis to aggregate the results. They were successful in describing both genotype-phenotype associations that were unique to individual populations, and signals that reached significance when all populations were taken into account via a trans-ethnic meta-analysis.

If we have been careful in performing all steps above, including quality control, association testing and post-GWAS interrogation, we should have a list of SNPs that are enriched for real genotype-phenotype associations.

7 DISCUSSION

In this review, we have attempted to give an overview of the methods used for performing GWAS on admixed populations. The main objective was to shed some light on the intuition behind using each of them.

1) Quality Control. The objective in this step is to remove low-quality SNPs and samples and to ensure a comparable population structure across the phenotype (e.g. same distribution among cases and controls).
   - Comparability of cases and controls. Removing outliers from the data can be convenient to the analysis, but excluding whole subpopulations hurts the generalizability of the study. This strategy is used mostly when the control data has not been sampled according to the same protocol as the cases, like the case of using a generic database such as a biobank.
   - Separating the data for multiple association studies. If there is an overrepresentation of a subpopulation or if there is a need to report population specific related SNPs, it could be convenient to analyse the data separately. The main caveat of doing this is the possibility of having to perform an association test with few data.
2) Controlling for ancestry at the association test step. Here, we account for population structure in the actual modeling of the genotype-genotype relationship. This helps avoid spurious correlations. Methods that we can use for this purpose are:
   - PCA. There is no reason not to control for ancestry using PCA, but it is important to add the correct number of components to the model (Tian et al., 2008). The recommendation is to review the distribution of the data in several component plots and to examine the results of inflation by using the genomic factor, or use specialised software such as EIGENSTRAT.
   - Admixture mapping. If there are no clearly distinct subpopulations found in the sample, admixture mapping is an appropriate way to find regions where the admixture is related to the phenotype. Some methods such as Tractor can also find the specific effect sizes on each of the subpopulations.
3) Post-GWAS interrogation. As in many other cases of experimental studies, the results of a statistical procedure should be analysed and should be open to correction according to the data and data cleaning that has been used.
   - Genomic control. This tool is useful as a measure of the population structure that has been introduced to the study, and to suggest whether or not it is necessary to go back to previous steps in order to further account for the structure of the data. Although it is possible to use it to control for overall population structure by scaling the p-values of the association test, it is not recommended and should be used only as a sanity check.
   - Meta-analysis. This is necessary in order to aggregate the results in the case that we have separated the data into its subpopulations. It is possible to achieve the same power as a whole-data association test given some properties, but any population specific signal that may have appeared in the individual studies might be lost in the meta-analysis.

AUTHOR CONTRIBUTIONS

IS researched and wrote the majority of this review, with help from PO, DB and MI reviewed all sections and provided advice, and supervised the work together with CR.

FUNDING

IS is a PhD student from Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México. CR is supported by

the Medical Research Council (MR/S01473X/1), Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT UNAM) (IA202020), the Academy of Medical Sciences through a Newton Advanced Fellowship (NAF/R2/180782) and by the Wellcome Sanger Institute through an International Fellowship, and by CONACyT (Projects no. A1-S-30165 and A3-S-31603).

ACKNOWLEDGMENTS

We acknowledge help by Luis A. Aguilar, Alejandro de León and Carlos S. Flores of the Laboratorio Nacional de Visualización Científica Avanzada, and Jair S. García Sotelo, Abigayl Hernández, Eglee Lomelín, Alejandra Castillo and Carina Díaz from Laboratorio Internacional de Investigación sobre el genoma Humano, Universidad Nacional Autónoma de México.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.703901/full#supplementary-material

REFERENCES

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

Atkinson, E. G., Maihofer, A. X., Kanai, M., Martin, A. R., Karczewski, K. J., Santoro, M. L., et al. (2021). Tractor Uses Local Ancestry to Enable the Inclusion of Admixed Individuals in Gwas and to Boost Power. *Nat. Genet.* 53, 195–204. doi:10.1038/s41588-020-00766-y

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., et al.Consortium (2015). A Global Reference for Human Genetic Variation. *Nature* 526 (7571), 68–74. [Dataset]. doi:10.1038/nature15393

Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Eng, C., et al. (2012). Fast and Accurate Inference of Local Ancestry in Latino Populations. *Bioinformatics* 28, 1359–1367. doi:10.1093/bioinformatics/bts144

Begum, F., Ghosh, D., Tseng, G. C., and Feingold, E. (2012). Comprehensive Literature Review and Statistical Considerations for Gwas Meta-Analysis. *Nucleic Acids Res.* 40, 3777–3784. doi:10.1093/nar/gkr1255

Boca, S. M., Huang, L., and Rosenberg, N. A. (2020). On the Heterozygosity of an Admixed Population. *J. Math. Biol.* 81, 1217–1250. doi:10.1007/s00285-020-01531-9

Bulik-Sullivan, B. K., Loh, P.-R., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., et al. (2015). Ld Score Regression Distinguishes Confounding from Polygenicity in Genome-wide Association Studies. *Nat. Genet.* 47, 291–295. doi:10.1038/ng.3211

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2019). The Nhgri-Ebi Gwas Catalog of Published Genome-wide Association Studies, Targeted Arrays and Summary Statistics 2019. *Nucleic Acids Res.* 47, D1005–D1012. doi:10.1093/nar/gky1120

Burkart, K. M., Sofer, T., London, S. J., Manichaikul, A., Hartwig, F. P., Yan, Q., et al. (2018). A Genome-wide Association Study in Hispanics/latinos Identifies Novel Signals for Lung Function. The Hispanic Community Health Study/study of Latinos. *Am. J. Respir. Crit. Care Med.* 198, 208–219. doi:10.1164/rccm.201707-1493oc

Choudhry, S., Coyle, N. E., Tang, H., Salari, K., Lind, D., Clark, S. L., et al. (2006). Population Stratification Confounds Genetic Association Studies Among Latinos. *Hum. Genet.* 118, 652–664. doi:10.1007/s00439-005-0071-3

Conomos, M. P., Laurie, C. A., Stilp, A. M., Gogarten, S. M., McHugh, C. P., Nelson, S. C., et al. (2016). Genetic Diversity and Association Studies in Us Hispanic/latino Populations: Applications in the Hispanic Community Health Study/study of Latinos. *Am. J. Hum. Genet.* 98, 165–184. doi:10.1016/j.ajhg.2015.12.001

Costa-Urrutia, P., Colistro, V., Jiménez-Osorio, A. S., Cárdenas-Hernández, H., Solares-Tlapechco, J., Ramirez-Alcántara, M., et al. (2019). Genome-wide Association Study of Body Mass index and Body Fat in Mexican-Mestizo Children. *Genes* 10, 945. doi:10.3390/genes10110945

Durvasula, A., and Sankararaman, S. (2019). A Statistical Model for Reference-free Inference of Archaic Local Ancestry. *Plos Genet.* 15, e1008175. doi:10.1371/journal.pgen.1008175

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics* 164, 1567–1587. doi:10.1093/genetics/164.4.1567

Freedman, M. L., Haiman, C. A., Patterson, N., McDonald, G. J., Tandon, A., Waliszewska, A., et al. (2006). Admixture Mapping Identifies 8q24 as a Prostate Cancer Risk Locus in African-American Men. *Proc. Natl. Acad. Sci.* 103, 14068–14073. doi:10.1073/pnas.0605832103

Galanter, J. M., Gignoux, C. R., Torgerson, D. G., Roth, L. A., Eng, C., Oh, S. S., et al. (2014). Genome-wide Association Study and Admixture Mapping Identify Different Asthma-Associated Loci in Latinos: The Genes-Environments & Admixture in Latino Americans Study. *J. Allergy Clin. Immunol.* 134, 295–305. doi:10.1016/j.jaci.2013.08.055

Gurdasani, D., Barroso, I., Zeggini, E., and Sandhu, M. S. (2019). Genomics of Disease Risk in Globally Diverse Populations. *Nat. Rev. Genet.* 20, 520–535. doi:10.1038/s41576-019-0144-0

Higgins, J. P. T., and Thompson, S. G. (2002). Quantifying Heterogeneity in a Meta-Analysis. *Statist. Med.* 21, 1539–1558. doi:10.1002/sim.1186

Hodonsky, C. J., Jain, D., Schick, U. M., Morrison, J. V., Brown, L., McHugh, C. P., et al. (2017). Genome-wide Association Study of Red Blood Cell Traits in Hispanics/latinos: The Hispanic Community Health Study/study of Latinos. *Plos Genet.* 13, e1006760. doi:10.1371/journal.pgen.1006760

Hoggart, C. J., Shriver, M. D., Kittles, R. A., Clayton, D. G., and McKeigue, P. M. (2004). Design and Analysis of Admixture Mapping Studies. *Am. J. Hum. Genet.* 74, 965–978. doi:10.1086/420855

Hubisz, M. J., Williams, A. L., and Siepel, A. (2020). Mapping Gene Flow between Ancient Hominins through Demography-Aware Inference of the Ancestral Recombination Graph. *Plos Genet.* 16, e1008895. doi:10.1371/journal.pgen.1008895

Jorgenson, E., Thai, K. K., Hoffmann, T. J., Sakoda, L. C., Kvale, M. N., Banda, Y., et al. (2017). Genetic Contributors to Variation in Alcohol Consumption Vary by Race/ethnicity in a Large Multi-Ethnic Genome-wide Association Study. *Mol. Psychiatry* 22, 1359–1367. doi:10.1038/mp.2017.101

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., et al. (2008). Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics* 178, 1709–1723. doi:10.1534/genetics.107.080101

Kelley, G. A., and Kelley, K. S. (2012). Statistical Models for Meta-Analysis: A Brief Tutorial. *Wjm* 2, 27. doi:10.5662/wjm.v2.i4.27

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., et al. (2018). Genome-wide Polygenic Scores for Common Diseases Identify Individuals with Risk Equivalent to Monogenic Mutations. *Nat. Genet.* 50, 1219–1224. doi:10.1038/s41588-018-0183-z

Lambert, S. A., Abraham, G., and Inouye, M. (2019). Towards Clinical Utility of Polygenic Risk Scores. *Hum. Mol. Genet.* 28, R133–R142. doi:10.1093/hmg/ddz187

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). Fast Linear Mixed Models for Genome-wide Association Studies. *Nat. Methods* 8, 833–835. doi:10.1038/nmeth.1681

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., et al. (2015). Efficient Bayesian Mixed-Model Analysis Increases Association Power in Large Cohorts. *Nat. Genet.* 47, 284–290. doi:10.1038/ng.3190

Manolio, T. A. (2013). Bringing Genome-wide Association Findings into Clinical Use. *Nat. Rev. Genet.* 14, 549–558. doi:10.1038/nrg3523

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

Martin, E. R., Tunc, I., Liu, Z., Slifer, S. H., Beecham, A. H., and Beecham, G. W. (2018). Properties of Global- and Local-Ancestry Adjustments in Genetic Association Tests in Admixed Populations. *Genet. Epidemiol.* 42, 214–229. doi:10.1002/gepi.22103

McKeigue, P. M. (1998). Mapping Genes that Underlie Ethnic Differences in Disease Risk: Methods for Detecting Linkage in Admixed Populations, by Conditioning on Parental Admixture. *Am. J. Hum. Genet.* 63, 241–251. doi:10.1086/301908

Medina-Gomez, C., Felix, J. F., Estrada, K., Peters, M. J., Herrera, L., Kruithof, C. J., et al. (2015). Challenges in Conducting Genome-wide Association Studies in Highly Admixed Multi-Ethnic Populations: the Generation R Study. *Eur. J. Epidemiol.* 30, 317–330. doi:10.1007/s10654-015-9998-4

Mills, M. C., and Rahal, C. (2019). A Scientometric Review of Genome-wide Association Studies. *Commun. Biol.* 2, 9–11. doi:10.1038/s42003-018-0261-x

Montana, G., and Pritchard, J. K. (2004). Statistical Tests for Admixture Mapping with Case-Control and Cases-Only Data. *Am. J. Hum. Genet.* 75, 771–789. doi:10.1086/425281

Nannini, D. R., Torres, M., Chen, Y. D. I., Taylor, K. D., Rotter, J. I., Varma, R., et al. (2017). A Genome-wide Association Study of Vertical Cup-Disc Ratio in a Latino Population. *Invest. Ophthalmol. Vis. Sci.* 58, 87–95. doi:10.1167/iovs.16-19891

Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K. E., Hafler, D. A., Oksenberg, J. R., et al. (2004). Methods for High-Density Admixture Mapping of Disease Genes. *Am. J. Hum. Genet.* 74, 979–1000. doi:10.1086/420871

Patterson, N., Price, A. L., and Reich, D. (2006). Population Structure and Eigenanalysis. *Plos Genet.* 2, e190. doi:10.1371/journal.pgen.0020190

Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, Á., and Lareu, M. V. (2013). An Overview of Structure: Applications, Parameter Settings, and Supporting Software. *Front. Genet.* 4, 98. doi:10.3389/fgene.2013.00098

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal Components Analysis Corrects for Stratification in Genome-wide Association Studies. *Nat. Genet.* 38, 904–909. doi:10.1038/ng1847

Pritchard, J. K., and Rosenberg, N. A. (1999). Use of Unlinked Genetic Markers to Detect Population Stratification in Association Studies. *Am. J. Hum. Genet.* 65, 220–228. doi:10.1086/302449

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155, 945–959. doi:10.1093/genetics/155.2.945

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., 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

Reich, D., Patterson, N., Jager, P. L. D., McDonald, G. J., Waliszewska, A., Tandon, A., et al. (2005). A Whole-Genome Admixture Scan Finds a Candidate Locus for Multiple Sclerosis Susceptibility. *Nat. Genet.* 37, 1113–1118. doi:10.1038/ng1646

Sankararaman, S., Mallick, S., Patterson, N., and Reich, D. (2016). The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Curr. Biol.* 26, 1241–1247. doi:10.1016/j.cub.2016.03.037

Seldin, M. F. (2007). Admixture Mapping as a Tool in Gene Discovery. *Curr. Opin. Genet. Development* 17, 177–181. doi:10.1016/j.gde.2007.03.002

Shmulewitz, D., Zhang, J., and Greenberg, D. A. (2004). Case-control Association Studies in Mixed Populations: Correcting Using Genomic Control. *Hum. Hered.* 58, 145–153. doi:10.1159/000083541

Thornton, T. A., and Bermejo, J. L. (2014). Local and Global Ancestry Inference and Applications to Genetic Association Analysis for Admixed Populations. *Genet. Epidemiol.* 38, S5–S12. doi:10.1002/gepi.21819

Tian, C., Gregersen, P. K., and Seldin, M. F. (2008). Accounting for Ancestry: Population Substructure and Genome-wide Association Studies. *Hum. Mol. Genet.* 17, R143–R150. doi:10.1093/hmg/ddn268

Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., et al. (2011). Quality Control Procedures for Genome-wide Association Studies. *Curr. Protoc. Hum. Genet.* Chapter 1, Unit1–19. doi:10.1002/0471142905.hg0119s68

Wang, H., Cade, B. E., Sofer, T., Sands, S. A., Chen, H., Browning, S. R., et al. (2019). Admixture Mapping Identifies Novel Loci for Obstructive Sleep Apnea in Hispanic/latino Americans. *Hum. Mol. Genet.* 28, 675–687. doi:10.1093/hmg/ddy387

Wang, X., Zhu, X., Qin, H., Cooper, R. S., Ewens, W. J., Li, C., et al. (2011). Adjustment for Local Ancestry in Genetic Association Analysis of Admixed Populations. *Bioinformatics* 27, 670–677. doi:10.1093/bioinformatics/btq709

Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., and Price, A. L. (2014). Advantages and Pitfalls in the Application of Mixed-Model Association Methods. *Nat. Genet.* 46, 100–106. doi:10.1038/ng.2876

Zhao, S., Jing, W., Samuels, D. C., Sheng, Q., Shyr, Y., and Guo, Y. (2018). Strategies for Processing and Quality Control of Illumina Genotyping Arrays. *Brief. Bioinformatics* 19, 765–775. doi:10.1093/bib/bbx012

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., and Weir, B. S. (2012). A High-Performance Computing Toolset for Relatedness and Principal Component Analysis of Snp Data. *Bioinformatics* 28, 3326–3328. doi:10.1093/bioinformatics/bts606

Zhu, X., Luke, A., Cooper, R. S., Quertermous, T., Hanis, C., Mosley, T., et al. (2005). Admixture Mapping for Hypertension Loci with Genome-Scan Markers. *Nat. Genet.* 37, 177–181. doi:10.1038/ng1510

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

REVIEW
published: 12 November 2021
doi: 10.3389/fgene.2021.763363

Maintenance of Complex Trait Variation: Classic Theory and Modern Data

Evan M. Koch[1,2] *and Shamil R. Sunyaev*[1,2]*

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, United States, [2]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States

Numerous studies have found evidence that GWAS loci experience negative selection, which increases in intensity with the effect size of identified variants. However, there is also accumulating evidence that this selection is not entirely mediated by the focal trait and contains a substantial pleiotropic component. Understanding how selective constraint shapes phenotypic variation requires advancing models capable of balancing these and other components of selection, as well as empirical analyses capable of inferring this balance and how it is generated by the underlying biology. We first review the classic theory connecting phenotypic selection to selection at individual loci as well as approaches and findings from recent analyses of negative selection in GWAS data. We then discuss geometric theories of pleiotropic selection with the potential to guide future modeling efforts. Recent findings revealing the nature of pleiotropic genetic variation provide clues to which genetic relationships are important and should be incorporated into analyses of selection, while findings that effect sizes vary between populations indicate that GWAS measurements could be misleading if effect sizes have also changed throughout human history.

Keywords: population genetics, genome-wide association study, statistical genetics, evolution, quantitative genetics

OPEN ACCESS

Edited by:
Mashaal Sohail,
University of Chicago, United States

Reviewed by:
Jeremy Berg,
University of Chicago, United States
Diego Ortega-Del Vecchyo,
National Autonomous University of Mexico, Mexico

***Correspondence:**
Shamil R. Sunyaev
ssunyaev@hms.harvard.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *23 August 2021*
Accepted: *19 October 2021*
Published: *12 November 2021*

Citation:
Koch EM and Sunyaev SR (2021) Maintenance of Complex Trait Variation: Classic Theory and Modern Data. Front. Genet. 12:763363. doi: 10.3389/fgene.2021.763363

1 INTRODUCTION

Attempts to understand genetic architecture preceded the discovery of DNA as the model of heredity (Fisher, 1918), and much theoretical work on selection, the maintenance of variation, and the adaptation of complex traits began before the ability to record genotypes on a scale sufficient to meaningfully contribute to these questions (Walsh and Lynch, 2018). The modern genetic era has provided an opportunity to test classic theories and to expand models—both long-standing and relatively recent—based on new understandings of genetic architecture and mechanisms. Genome-wide association studies (GWAS) and other data-driven tools have raised additional questions, including how so much heritability for many traits is contributed by relatively common alleles when natural selection is often expected to remove deleterious variation from the population. The flood of methods and data has sharpened and revised our understanding of many components that fashion the structure of the genome—polygenicity, selection, the distribution of mutational effects, pleiotropy—but has left us wanting for models capable of reconciling these elements (Sella and Barton, 2019).

The analysis of GWAS data revealed an extraordinary degree of polygenicity, and showed that most heritability is explained by relatively common, mostly noncoding alleles of small effect. At first

glance, this observation is surprising. Natural selection is expected to maintain the population near an optimum value for quantitative traits and to reduce the prevalence of potentially maladaptive phenotypes such as diseases. Such optimums and maladaptive phenotypes are defined within a given environmental context (Harpak and Przeworski, 2021). Selection generally acts by reducing the frequency of phenotypically relevant alleles, though it may drive allele frequency increases when shifts in the optimum phenotype occur. This basic logic led to the question whether the effect of natural selection is evident from GWAS data at all. Recent studies have reached a strong consensus that phenotypic effect sizes are negatively correlated with allele frequency (Gazal et al., 2018; Zeng et al., 2018; Schoech et al., 2019; Speed et al., 2020; Zeng et al., 2021). These findings are inconsistent with purely neutral models, but various models of natural selection influencing trait variation remain plausible (Walsh and Lynch, 2018). Uncertainty largely surrounds whether the focal trait is causally important for fitness compared to pleiotropically related ones, and whether selection is primarily stabilizing or has important directional components. In spite of many unresolved details, the emerging picture is that a vast supply of mutations with weak effects, coupled with generally inefficient selection against such alleles, is the basis of phenotypic variation.

Empirical results from GWAS on the distribution of effect sizes and allele frequencies still pose the challenge of which classic and emerging models from theoretical population genetics are able to best explain the emerging observations. Existing theories range from models of selection acting directly on the focal trait to models where selection on genetic variation is driven by simultaneous effects on other traits (pleiotropy), to even fully "apparent" selection, which assumes the focal trait is not subject to any selective constraint. In this review, we discuss a relevant subset of these models and how their predictions look in light of recent studies of selection in GWAS. We identify pleiotropy and variable effect sizes of genetic variants across time and space as important factors that have yet to be satisfactorily included into statistical methods and theoretical models.

2 THEORETICAL MODELS OF MAINTENANCE OF COMPLEX TRAITS AND PREDICTIONS THEY GENERATE

Evolutionary quantitative genetics has subsisted for most of its existence on a limited set of possible measurements. Estimates of the genetic and mutational variance, as well as selection gradients, are informative, especially with respect to contemporary patterns of selection. However, most progress in explaining maintenance of genetic variation in phenotypic traits was theoretical. Now that GWAS have generated an abundance of matched phenotypic and genetic measurements we live in a much more data-rich world. If we turn our attention to a single, focal trait, what sort of data would we ideally wish for? We would probably include the impact of genetic variants (estimated as their effect size) on the trait in a range of environments, the fitness effects of these alleles, as well as their frequencies and linkage patterns (Johnson and Barton, 2005). These would yield a satisfying and useful description of the genetic architecture and the process of its development, but there are fundamental details not immediately obvious from this description. We would like to know whether fitness effects arise primarily through selection on the focal trait, and if so what form it takes. If there are substantial fitness effects unrelated to the focal trait, what other traits are involved and how does selection act on them? Is the population in equilibrium? Has the genetic architecture changed in the past and will it do so in the future? Questions like these can be addressed by modeling how selection acts on traits, the mutational distributions underlying them, and how these generate the genetic architectures we observe.

Textbook introductions to population genetics begin by assigning fitnesses to genotypes and examine the consequences for allele frequencies and overall patterns of genetic variation. Connecting trait values, such as those measured in GWAS, to selection on individual causative alleles requires the additional step of specifying how selection on phenotypes leads to fitness differences among genotypes. While slightly less familiar than other selection results, this task was also taken up by many of the authors of classical population genetics and has grown into a large branch of evolutionary theory.

The simplest and most obvious model predicts the selection on individual causative loci arising from stabilizing selection on a single polygenic trait with purely additive genetic variance (Wright, 1935; Robertson, 1956; Bulmer, 1972) (**Figure 1A**). In this model an individual's trait value (z) is determined by the sum of effects from L independent loci: $z = \sum_{l=1}^{L} (\beta_l g_m + \beta_l g_p) + e$, where β_l is the effect size of the allele at locus l, $g_{l,m}$ and $g_{l,p}$ are the maternal and paternal genotypes at each locus, and e is a normally distributed environmental effect centered at zero. If an individual's fitness is a Gaussian function centered at the population mean M and with width V_S ($w(z) = \exp(-(M-z)^2/2V_S)$), then selection will change the average frequency of a causative allele at locus l with effect size $\beta_l = \beta$ as follows:

$$\mathrm{E}[\Delta x] \approx -\frac{\beta^2}{2V_S} x(1-x)\left(\frac{1}{2} - x\right). \quad (1)$$

Stabilizing selection tends to remove genetic variation in this trait from the population. A balance between mutation, selection, and drift generates the trait's genetic variance in the population (Bulmer, 1972; Keightley and Hill, 1988). Such direct stabilizing selection leads to a negative correlation between minor allele frequencies and the effect size magnitude.

We can write the selection coefficient for this model as $s_{ud} = -\frac{\beta^2}{2V_S}$ to acknowledge that stabilizing selection takes the underdominant form shown in **Eq. 1** ($\mathrm{E}[\Delta x] = sx(1-x)(\frac{1}{2} - x)$) rather than the more familiar additive one ($\mathrm{E}[\Delta x] = sx(1-x)$). The $(\frac{1}{2} - x)$ term appearing in the underdominant formula means that selection against the derived allele actually decreases as it approaches 50% frequency and actually switches signs after that point. The minor allele is therefore always disfavored. However, when selection is strong or the allele frequency is low, the differences are minor as x is small compared to 1/2. We generally omit the subscript in s_{ud} for ease of

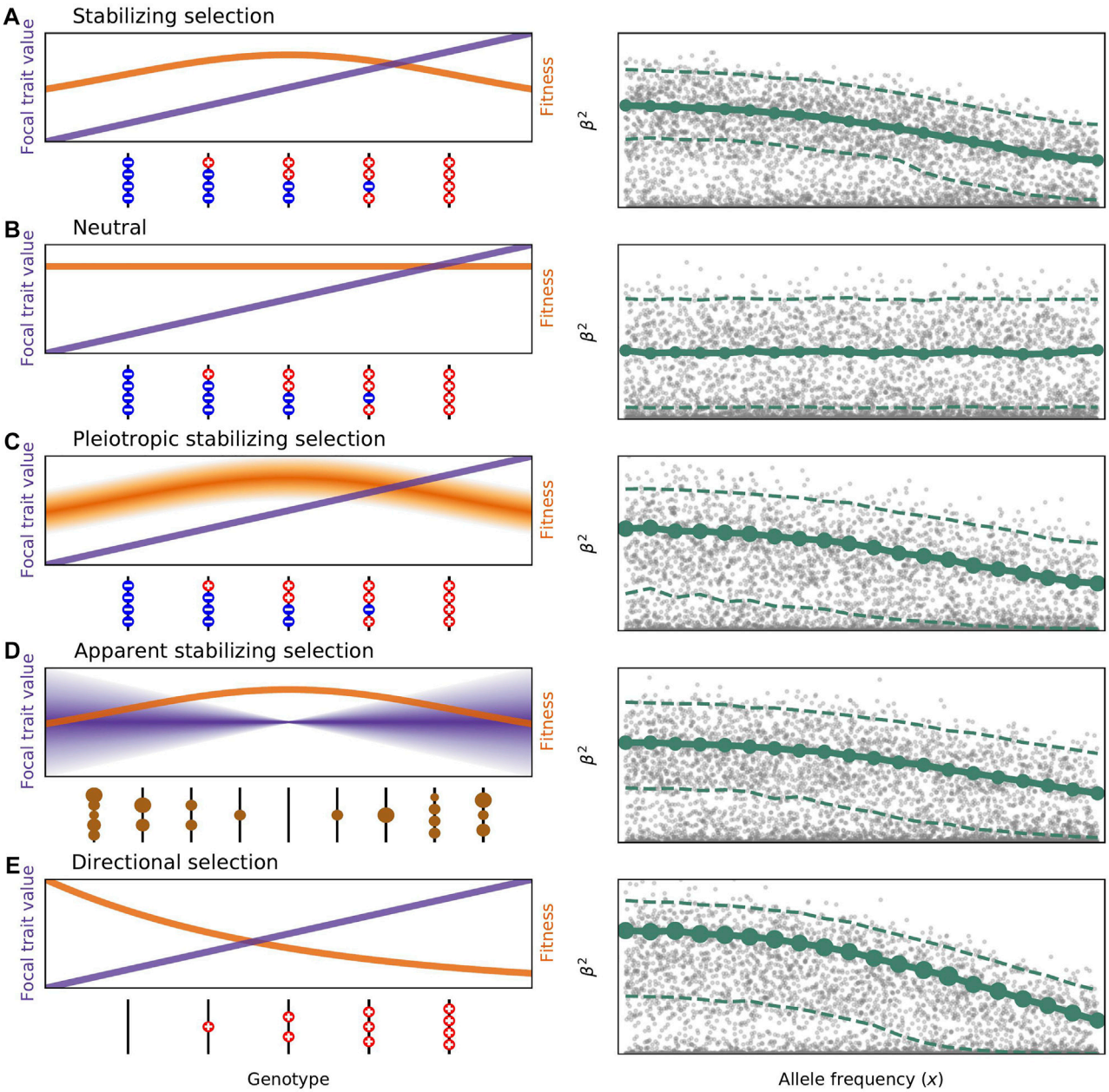

FIGURE 1 | Models of selection on the genetic variation influencing complex traits. Panels on the left show how different genotypes affect trait values (in purple) and fitness (in orange). Panels on the right illustrate how squared trait values change with frequency in each model of selection. Simulated values are shown in grey, the mean $E[\beta^2|x]$ is represented by the solid green line, and the standard deviation of $\beta^2|x$ is represented by the size of the green circles. The median and 97.5% quantile are shown as dashed lines to give a better sense of the full leptokurtic distribution of effect sizes. The DFE used in all plots (shape = 0.25, scale = 40) was chosen to be within the range fit by Schoech et al. (2019). Effect sizes were simulated uniformly on log frequency, and both axes are on a log scale. **(A)** Classic stabilizing selection as described by **Eq. 1**. Genotypes containing more trait-increasing alleles than decreasing, and vice versa, have lower fitness as a result of selection on the focal trait. Large effect alleles are prevented from reaching high frequencies due to the variance-reducing property of stabilizing selection. **(B)** In the neutral model no genotype is more fit than any other and the distribution of effect sizes at any frequency reflects only the distribution of mutational effects. **(C)** In pleiotropic stabilizing selection as studied by Simons et al. (2018), there is variation in fitness for each genotypic values depending on the effects mutations have on pleiotropic traits. This leads to the same average $E[\beta^2|x]$ but a greater variance and therefore different genetic architecture. **(D)** Models of apparent stabilizing selection first specify the deleterious fitness effects of mutations, represented here by the size of the brown circles. Genotypes with more and stronger deleterious mutations have a greater variance in phenotypic outcomes. This too leads to a negative relationship between β^2 and x. Here we use the Eyre-Walker (2010) model with $\tau = 0.4$ as fit by Schoech et al. (2019), and $\sigma^2 = 1$. Altering these would change the mean and variance of the (β^2, x) relationship. **(E)** Directional selection is shown here for a scenario where trait-decreasing mutations are unlikely or impossible. Selection therefore acts to reduce the frequency of trait-increasing alleles. All new mutations are disfavored with $s \propto \beta$. $E[\beta^2|x]$ again decreases with x.

reading, but it is important to note that the interpretation of selection coefficients differs depending on whether stabilizing selection is explicitly modeled or not.

The variance-reducing property of stabilizing selection motivated the development of other models with variance-promoting features like overdominant side-effects of causative alleles (Robertson, 1956; Gillespie, 1984) and strong mutational pressure (Lande, 1976b). As always in evolution, we must also at least consider the possibility that a trait of interest has a negligible impact on organismal fitness. The population mean value of a trait controlled by strictly neutral mutation will drift in Brownian motion and have a genetic variance that depends on the mutation rate, the second moment of the distribution of mutation effect sizes, and the average pairwise coalescent time between randomly sampled loci (Lande, 1976a; Lynch and Hill, 1986; Koch, 2019). $\mathrm{Var}[z] = \mathrm{E}[T_2]\theta\mu_2$, where T_2 is the average number of generations it takes for a pair of sites to coalesce, θ is the mutation rate per generation, and μ_2 is the second moment of the distribution of mutational effects. Crucially, there would be no relationship between the effect size and frequency of alleles (**Figure 1B**).

Of course, both intuitively and empirically, traits in natural and contemporary human populations at least appear to be under some selection (Kingsolver et al., 2001; Corbett et al., 2018; Sanjak et al., 2018), and involve some level of pleiotropy (Stearns, 2010). Models of apparent selection begin with the assumption that the focal trait is not itself under any selection but add pleiotropic fitness effects of the causative alleles. Individuals in the tails of a phenotypic distribution will carry more mutations overall, and if trait-affecting mutations have deleterious pleiotropic effects those individuals will also have lover fitness on average (Barton, 1990; Kondrashov and Turelli, 1992). Fitness that decreases away from the mean is reminiscent of stabilizing selection, but the strict deleterious model of apparent selection does not induce the negative correlation between allele frequencies and effect size magnitudes expected when the focal trait itself is actively selected. The negative correlation between β^2 and x may yet be rescued if the deleterious pleiotropic effects of variants affecting the focal trait arise from genetic covariance (Lande and Arnold, 1983) or correlated effect size magnitudes (Keightley and Hill, 1990). In this scenario, alleles with larger effects (or absolute magnitudes) on the focal, neutral trait are more likely to have larger effects on a second, selected trait. Allele frequencies are suppressed through selection on the second. In the extreme where the focal and selected trait are so closely biologically related that the effect sizes of mutations are deterministically linked, it is indistinguishable which trait causally impacts fitness, although a strong genetic covariance would be measurable. One can also imagine a model where each mutation has such a relationship with a unique pleiotropic trait, for instance, molecular effects in different pathways. Assuming that large-effect alleles for the focal trait induce stronger molecular effects, there may be strong selection without measurable genetic covariance between the focal trait and any individual pleiotropic trait.

The differences between models come down to how the statistical relationship between selection coefficients and effect sizes is specified: how s scales on average with β and what the random variation around this looks like. In multivariate stabilizing selection, s scales with β^2 as in direct stabilizing selection, but apparent selection models don't have this restriction. Apparent selection models were extended, as described above, to include increasing selection with greater β in addition to the negative pleiotropic consequences (Keightley and Hill, 1990; Zhang and Hill, 2002; Eyre-Walker, 2010) (**Figure 1D**). Models of multivariate stabilizing selection paint a similar picture, but the focal and pleiotropic traits are explicitly under stabilizing selection (Zhang and Hill, 2003; Simons et al., 2018) (**Figure 1C**). All lead to a negative (β^2, x) relationship, so differences between models come down to the shape and variance of that relationship, along with impacts on the genetic architecture.

Directional selection on complex disease susceptibility is also a viable hypothesis. In this view, the disease phenotype is itself deleterious and alleles that increase susceptibility will be selected against (Charlesworth, 2001; Wright et al., 2003) (**Figure 1E**). This also implies a mutational bias towards susceptibility-increasing alleles. It is plausible that there is a fitness cost associated with carrying such alleles, even for late-onset diseases (Pavard and Coste, 2021). All of the pleiotropy arguments made for stabilizing selection would apply equally well here.

There is an emerging consensus that models of mutation-selection-drift balance are likely to explain the genetic architecture of many, if not most, complex traits (Sella and Barton, 2019). The models of apparent, stabilizing, and directional selection described above, with varying possible degrees of pleiotropic selection, all remain possibilities within this consensus and are not mutually exclusive. Progress in statistical genetics methodology and increasing GWAS sample sizes are starting to clarify these details.

3 DETECTING NEGATIVE SELECTION IN GENOME-WIDE ASSOCIATION STUDIES DATA

As sample sizes increased and GWAS became sufficiently powered to detect larger numbers of loci for different traits, attention started shifting from the speculative question of how study design should be informed by selection and its effect on genetic architecture (Pritchard, 2001; Reich and Lander, 2001), to what the genetic architecture, as revealed through these studies, might say about selection. A transitional form was contributed by (Agarwala et al., 2013), who investigated how selection may have shaped the genetic architecture of Type 2 Diabetes, which had recently gone from 2 to 39 genome-wide significant loci. Using primarily the number of associations, conditional on the heritability and prevalence of the disease, they ruled out both neutrality of the focal trait and a model where selection is proportional to effect size: $\beta \propto |s|$.

Following this, methods were developed that do not explicitly model natural selection on causative variants, but ask whether lower frequency variants contribute disproportionately to

heritability. This heritability bias should only occur if rare variants have larger effect sizes on average, the most plausible explanation being negative selection correlated with the magnitude of effect sizes. A simple approach is to divide variants into MAF bins and estimate the heritability contribution of each in a mixed model framework (Yang et al., 2015). applied this approach to height and body mass index (BMI) and (Mancuso et al., 2016) to prostate cancer risk. Both found increased heritability in rare variants compared to common, and Mancuso et al. (2016) performed simulations to demonstrate that, conditional on disease heritability and prevalence, they could also rule out focal trait neutrality and directly proportional selection.

More sophisticated analyses using the same general idea as partitioning heritability by allele frequency have been developed and applied to a wide variety of human traits. Extensions of LD score regression (LDSC), a useful tool for partitioning heritability among large numbers of annotations (Finucane et al., 2015), were developed for features of negative selection. These analyses found that younger genetic variants contribute more heritability than older genetic variants at the same frequency (Gazal et al., 2017), a key feature of negative selection (Maruyama, 1974; Kiezun et al., 2013). They also confirmed earlier findings that rare variants have greater effect sizes than common ones for a larger number of traits (Gazal et al., 2018). Another popular and tractable approach, termed the alpha model, explicitly sets the MAF dependence of heritability contributions through a single parameter α: $E[\beta^2|x] \propto (x(1-x))^{\alpha}$ (Zeng et al., 2018; Schoech et al., 2019; Speed et al., 2020; Zeng et al., 2021). An $\alpha < 0$ indicates a heritability bias towards rare variants and some amount of negative selection. Applications of this model have been remarkable in their consistently negative α estimates across all analyzed traits. While some differences between traits are inferred, e.g. height has a smaller α than BMI, estimates are consistently within the range $[-0.5, -0.2]$.

The negative relationships between effect size magnitude and minor allele frequency inferred for so many traits are informative about the model of selection. In particular, they allow us to rule out neutral models where the focal trait and all underlying variation are unaffected by selection, as well as strict models of apparent selection where the causative variants are deleterious, but the strength of this selection is uncorrelated with effect sizes. However, many other models of selection may still be compatible with these findings (**Figure 1**, **Figure 2**). The model of direct stabilizing selection on a single trait first proposed by Wright and others (**Equation 1**) is plausible for some traits. On the other hand, a genetic correlation between the focal trait and another (or many) under stabilizing selection could produce the observed negative correlations even if the focal traits were completely neutral. There is also a lot of space in between with varying contributions to selection from the focal and pleiotropic traits. While seemingly semantic, the question is really about the extent to which variation in the underlying biology of the focal trait causes variation in fitness. This can apply even when the focal trait is something seemingly benign, an arbitrary bone for instance, whose size is governed chiefly by the biology of overall body size.

While the alpha model does not explicitly incorporate a population genetics model in any statistical analysis, it is possible to further interpret results using simulations and theory (**Figure 2**). In simulations, the idea is to use a model of choice to generate allele frequencies and effect sizes and then use the inference procedure to estimate what α corresponds to those model parameters. For theory, one derives $E[\beta^2|x]$ under the selection model and compare this to the approximate alpha model expectation of $E[\beta^2|x] \propto x^{\alpha}$. Using this approach, Schoech et al. (2019) showed that the inferred α depends both on the distribution of fitness effects of new alleles affecting the trait (DFE), and on the average scaling of effect sizes and selection $E[\beta^2|s] \propto s^{2\tau}$, where the parameter τ determines the scaling through the relationship $E[\beta] \propto s^{\tau}$ (Eyre-Walker, 2010). The DFE dependence enters primarily through a frequency-threshold effect: alleles below the threshold have effect sizes roughly uncorrelated with frequency because, while above the dependence scales approximately like $E[\beta^2|x] \propto x^{-2\tau}$. The threshold is the frequency below which most new mutations from the DFE are still mostly affected by drift rather than selection, and is therefore lower for a heavy-tailed DFE with a high average s. Zeng et al. (2021) used population genetic simulations to fit the DFE for various traits by conditioning on the values they had estimated for α, polygenicity, and SNP heritability. They found greater variation in the DFE among trait categories than variation in α estimates. α estimates that are relatively insensitive to the DFE are consistent with the predictions of Schoech et al. (2019) if most SNPs included in the analysis are above the frequency threshold where selection is detectable. These simulations therefore also demonstrate that polygenicity and heritability are informative about the DFE.

Simons et al. (2018) developed a model for the relationship between effect sizes and selection coefficients based on isotropic stabilizing selection and Fisher's geometric model (the specifics of the model is discussed in a subsequent section). The number of trait dimensions in this model corresponds to the effective number of independent axes of genetic variation, a value that can be interpreted as the degree of pleiotropy. With a single dimension the selection coefficient is the same as in the classical model of one dimensional stabilizing selection: $|\beta| = \sqrt{2\ sV_S}$. When the number of traits becomes large the relationship becomes $\beta \sim N\ (0,(V_S/n_e)s)$, where n_e is the effective number of traits, and expressions for moderate pleiotropy interpolate between these extremes. Rather than fit this model to the heritability explained by different minor allele frequencies, Simons et al. (2018) analyzed the distribution of variance contributions, $v = 2\beta^2 x(1-x)$, among genome-wide significant SNPs. For a given mean among discovered loci, the variance of v is higher with greater pleiotropy (n_e), with a parametric likelihood derived by the authors. The high-pleiotropy model was found to fit the distribution of GWAS hits for standing height and BMI better than the no- and low-pleiotropy alternatives.

Zeng et al. (2021) also simulated varying degrees of pleiotropy using the Simons et al. (2018) model and found that α estimates were insensitive to changes in the degree of pleiotropy (n_e). This

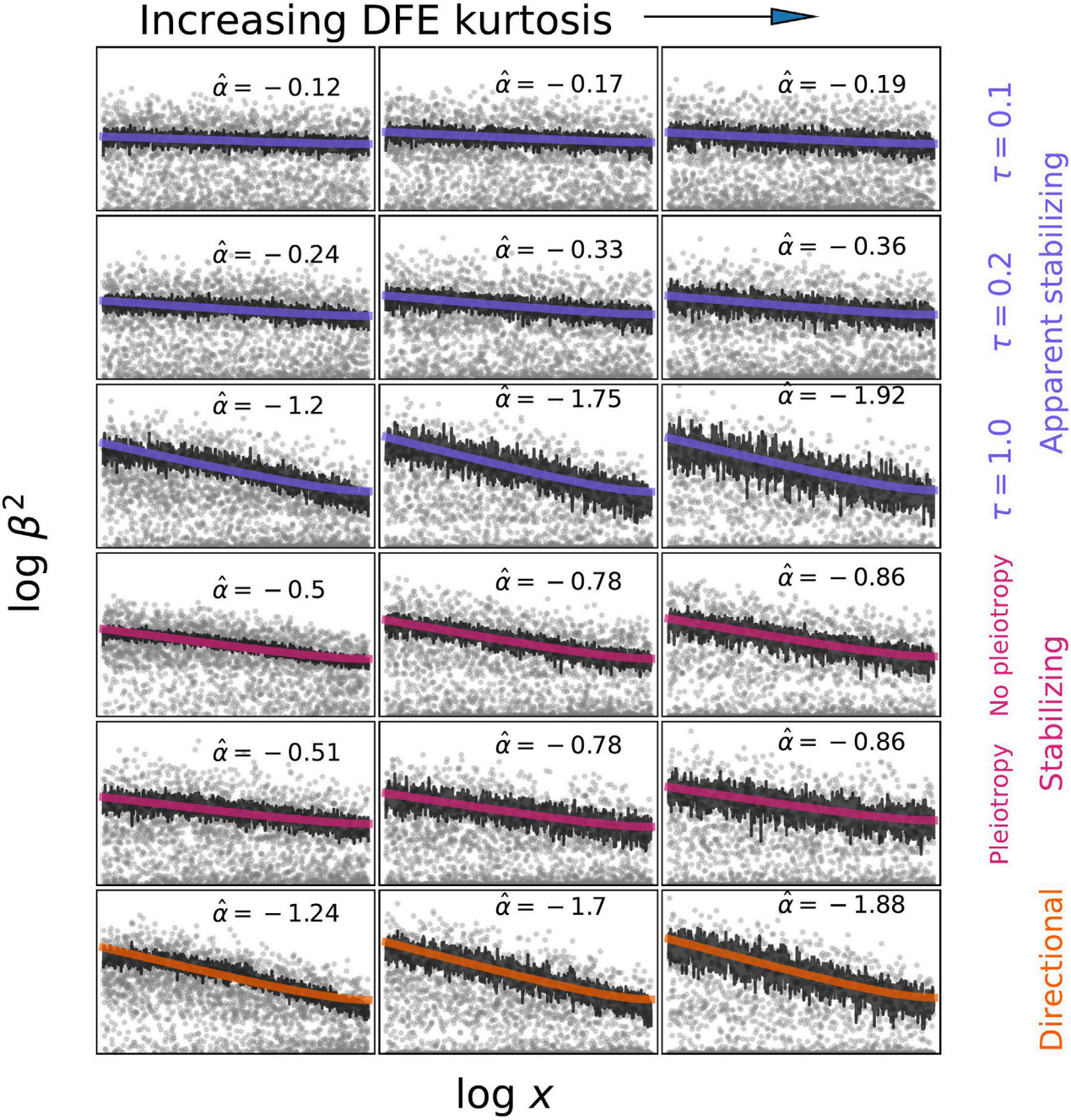

FIGURE 2 | Examples of what alpha models may infer under different models of selection and different distributions of fitness effects (DFE). Effect sizes were simulated by sampling from $p(s|x)$ and then from $p(\beta|s)$ under the different models described in the text. Derived allele frequencies are uniform between 0.01 and 0.5. Estimates of α were obtained by fitting $\log \beta^2 = \alpha \log x(1-x) + c$ to the average $\beta^2|x$ values calculated from simulations. The DFE was varied by decreasing the shape parameter from 1 to 0.25 to 0.125 while keeping the mean constant. It is important to recognize that $\hat{\alpha}$ values reported here would not necessarily correspond to those obtained by real statistical genetics methods (Zeng et al., 2018; Schoech et al., 2019; Speed et al., 2020; Zeng et al., 2021). Those methods employ particular likelihoods and are applied to real genetic data where frequencies are not uniform and effect sizes are estimated with error. The frequencies of analyzed variants may be particularly important since the slope of the (β^2, x) relationship (local α) varies with frequency (Schoech et al., 2019). Estimated α values increase with increasing DFE kurtosis, reflecting the proportion of variants that are strongly selected. For high kurtosis, estimates approach the theoretical expectation of $\alpha = -2\tau$ for the Eyre-Walker (2010) model as derived by Schoech et al. (2019). As expected, in a model of stabilizing selection (Simons et al., 2018), the degree of pleiotropy does not affect the α estimate. Directional selection is associated with higher α estimates than stabilizing selection.

makes sense, given that the alpha model only attempts to fit the average effect size - frequency relationship and suggests that new approaches will be needed to investigate the nature of pleiotropy and the relative importance of the focal trait.

4 MODEL BUILDING USING GEOMETRY AND PLEIOTROPY

Using the distribution of causative allele frequencies and their effects solely on the focal trait, what could be done to further interpret the results of GWAS studies? One advance would be to explicitly include selection in the next generation of models that build upon LDSC or α models (Sella and Barton, 2019). We may start by imagining what class of models could fit the joint distribution of (x, β). Assume a set of parameters Θ that describes the selection model. An analysis would use the likelihood $p(x, \beta|\Theta)$, which decomposes into $p(x, \beta|\Theta) = p(x|\beta, \Theta)p(\beta|\Theta)$. Since the distribution of effect sizes is not a major concern for selection, inference would focus on $p(x|\beta, \Theta)$. The distribution of frequencies for a given effect size is determined by integrating over the possible fitness effects of a mutation with effect size β: $p(x|\beta, \Theta) = \int p(x|s, \Theta)p(s|\beta, \Theta)ds$. The effect of selection, s, could be either additive or underdominant (stabilizing selection), but could also represent other models beyond these two such as overdominant or fluctuating selection. x can be replaced by the age or historical frequency path of the allele (Stern et al., 2021). $p(x|s)$ can be tackled with standard population genetics, so the trickier problem is to provide $p(s|\beta, \Theta)$ in cases of pleiotropic selection.

In an early attempt to do this explicitly, Keightley and Hill (1990) proposed $p(s|\beta, \Theta)$ as the conditional distribution of a two-dimensional Wishart distribution. In this formulation both the mean and variance of s, conditional on β, are proportional to $|\beta|$ plus a constant, and a correlation parameter determines how the variance scales with the mean. This contrasts with the model of direct stabilizing selection where s is proportional to β^2. Another approach decomposes this distribution as $p(s|\beta, \Theta) \propto p(\beta|s, \Theta)$ $p(s|\Theta)$. This has the appealing property of separating the link between fitness and trait effects from the distribution of fitness effects (DFE). Eyre-Walker (2010) proposed a form for $p(\beta|s, \Theta)$ where $E[\beta] \propto s^{\tau}$ with multiplicative noise. Both of these models try to capture a space of potential relationships between effect size and selection without being over-parameterized. However, it is not actually clear how one should interpret results from either. A weak correlation parameter from the Keightley and Hill (1990) model would perhaps indicate the importance of pleiotropy, but the linear scaling between $|\beta|$ and s would not make sense with direct stabilizing selection. Eyre-Walker's τ doesn't necessarily mean stronger or weaker selection. Would it mean anything for the relative importance of directional or stabilizing selection? Moreover, these two models make divergent predictions for the contribution of rare versus common alleles to the genetic variance (Eyre-Walker, 2010; Caballero et al., 2015; Sella and Barton, 2019).

Simons et al. (2018) made a strong argument for interpretability when deriving their distribution for $p(\beta|s, \Theta)$. The framework they used was multivariate stabilizing selection in a geometric model (Fisher, 1930). Models within this framework generally posit a multidimensional phenotypic space with a selection function that describes the fitness of each possible phenotypic combination. Typically, the fitness function is Gaussian and centered at some optimum phenotype. A mutation is a vector that moves an individual to a different point in phenotype space, thereby altering fitness. Assuming a population centered at its optimum value, with each phenotypic direction under equal stabilizing selection and mutational pressure, $p(\beta|s, \Theta)$ takes a simple parametric form depending only on the number of dimensions n_e and the strength of selection V_s.

Previous work using Fisher's geometric model had used it to derive the DFE of new mutations (Martin and Lenormand, 2006; Lourenço et al., 2011) or the expected genetic variance and correlation of the focal trait with fitness (Zhang and Hill, 2003) rather than $p(\beta|s)$. A major assumption of these studies was that the phenotypic effects of new mutations were drawn from a multivariate normal distribution with different dimensions representing different phenotypes. While a seemingly natural starting place, the assumption of normally distributed mutations is far from realistic and mathematically troublesome. There is accumulating evidence that the mutational effect distribution is substantially leptokurtic for many traits (Zhang et al., 2018; O'Connor et al., 2019; O'Connor, 2021). It is has also been shown that, for a single normal distribution of mutations, the DFE concentrates around a point value of s as the number of traits becomes large, an obviously unrealistic scenario (Waxman and Peck, 1998; Wingreen et al., 2003; Zhang and Hill, 2003). Thankfully, one may still rescue the utility of geometric models by using a mixture of normals.

For example, the Simons et al. (2018) likelihood can be derived from the geometric model with normal mutation proposed by Martin and Lenormand (2006) by integrating out a variance parameter: $p(s|\beta) = \int \frac{p(\beta|s,\sigma^2)p(s|\sigma^2)}{p(\beta|\sigma^2)} p(\sigma^2)d\sigma^2$. If mutations are uncorrelated, equally affected by stabilizing selection, and drawn from a mixture of normal distributions, then the distribution of variances, $p(\sigma^2)$, fully describes the mutational distribution. It is straightforward using Bayes' theorem to show that $p(s|\beta) \propto p(\beta|s)\int p(s|\sigma^2)p(\sigma^2)d\sigma^2$ and contains two components. The first component, $p(\beta|s)$, has the form derived by Simons et al. (2018), a normal distribution with variance proportional to s when the number of traits is large. This part is independent of σ^2. The second component, $\int p(s|\sigma^2)p(\sigma^2)d\sigma^2$, is the DFE itself. In this example, the DFE is generated by the distribution of mutational effects. The variance of what normal distribution in the mixture a mutation comes from determines how strongly selected it is.

The above approach suggests that a fruitful way to propose future models would be to propose that there exist different mutational modes. Modes might represent different biological pathways and could be parameterized by which traits are involved, the correlation of mutational effects among these, and the distribution of mutational effect sizes. If summarized in Θ_M, we might then integrate over the distribution of modes. If β is conditionally independent of s given Θ_M, then the form of the

DFE will be separable from the link between selection and effect sizes, though it is not always clear that this will be the case. Directional selection as well as antagonistic pleiotropy may be possible to model this way, at least for a population at equilibrium in its fitness landscape. To more directly analyze selection and the pleiotropic relationships among traits, a vector of effect sizes could replace the effect β on a single, focal trait.

5 EMPIRICAL DEMONSTRATIONS OF THE EXISTENCE AND NATURE OF PLEIOTROPY

Since models of the evolution and maintenance of complex trait variation strongly depend on assumptions regarding the degree of pleiotropy. Modeling and measurement of pleiotropy is key to the empirical questions of whether the focal trait is under meaningful direct selection and how selection coefficients depend on the phenotypic effects of individual variants.

Current estimates of polygenicity indirectly but strongly suggest highly pleiotropic genetic architecture for most complex traits (Zeng et al., 2018; O'Connor et al., 2019; Zeng et al., 2021). Indeed, it was estimated that 2% of genetic variation is involved in height and a similar proportion (1%) is involved in risk of Type 2 Diabetes (Zeng et al., 2021). It is clear that a model where every quantitative trait locus (QTL) affects just a single trait is, due to the finite nature of the human genome, inconsistent with high polygenicity (defined here as the probability that a variant has a non-zero effect on the focal trait). We do not know exactly what fraction of the genome plays any functional role; comparative and functional genomics produce a range of estimates generally on the order of 0.1 (Rands et al., 2014; Gulko et al., 2015). If 10% of the genome is of any functional importance, and trait-affecting mutations originate from this functional fraction, it clearly cannot harbor independent QTLs for a vast number of complex traits each with a polygenicity of 2% (Jordan et al., 2019).

With the abundance of GWAS data, many aspects of pleiotropy can be empirically estimated using corresponding well-developed statistical approaches. The specific relationships between causative QTL effect sizes on different traits that these approaches investigate are illustrated in **Figure 3**. The relationship between two phenotypes is most commonly expressed as global genetic covariance, which reflects the overall degree of pleiotropy in the form of correlation of QTL effects across all loci ($\mathrm{Cov}[\beta_1, \beta_2]$, where β_1 and β_2 are allelic effect sizes for phenotypes 1 and 2), scaled by the heterozygosity at each causative locus. A significant genetic covariance below one would indicate that individual QTL effects are correlated but not identical. Global genetic covariance is estimated using statistical approaches related to those used to estimate heritability including random effect models implemented into the GCTA software or LD-score regression (Lee et al., 2012; Bulik-Sullivan et al., 2015). Estimates of genetic covariance come with the same caveats as must apply to heritability estimates (Visscher et al., 2008). Measurements apply to the particular environment in which the different traits are measured and offer no guarantee of a fundamental relationship between traits. Under different conditions, gene-by-environment interactions can change which genetic variants contribute to heritability and different pleiotropic traits may associate with the new regime.

Using these and related statistical techniques, highly significant genetic covariances were estimated among various autoimmune diseases and among various psychiatric diseases and related phenotypes (Cotsapas et al., 2011; Lee et al., 2013; Watanabe et al., 2019; Lincoln et al., 2021). The analysis of genetic covariances between two traits has some limitations. The genetic covariance alone is not informative about biological mechanisms and per locus patterns. For example, the same genetic covariance may indicate either pleiotropy limited to just a few loci with very similar effects on both traits on the background of other non-pleiotropic loci or the broad pleiotropy of all loci but with non-identical effects (**Figures 3A** vs **3B**). For autoimmune traits, it is possible that some loci impact immune function while others determine tissue or organ specificity.

The question of contribution of individual loci into global genetic correlation must be, therefore, addressed at the local level by studying individual loci. When studying individual loci, one of the challenges is that linkage disequilibrium confounds the analysis. Genetic covariance may imply real pleiotropy, meaning that the same genetic variants causally affect both traits. Alternatively, some variants may exclusively impact the first trait and other variants exclusively impact the second trait, but local genetic correlation can still be induced by linkage disequilibrium between the two sets. Consequently, the field has developed two different classes of methods to address this issue. Methods that estimate local genetic covariance (Shi et al., 2017) do not distinguish between functional pleiotropy versus non-independence induced by linkage disequilibrium. A different class of methods called "colocalization" (Giambartolomei et al., 2014; Hormozdiari et al., 2016; Chun et al., 2017) relies on linkage disequilibrium patterns to specifically test the hypothesis that the same causative variant (or variants) in the locus impacts both traits (**Figure 3B**). Multiple examples of local genetic correlations and individually colocalized loci have been described (van Rheenen et al., 2019; Aguet et al., 2020; Vuckovic et al., 2020). However, some QTLs involved in genetically correlated traits do not show obvious signals of colocalization, suggesting that genetic correlation does not necessarily imply pleiotropic effects of all variants (Lincoln et al., 2021).

A separate aspect of pleiotropy that statistical genetics addresses is the causal relationship between phenotypes (van Rheenen et al., 2019). There is an important distinction between "horizontal" pleiotropy with genetic variants exerting independent effects on both traits and a causal path or "vertical" pleiotropy, where one trait directly contributes to the other (Jordan et al., 2019). Examples of the latter include LDL cholesterol being a causative risk factor of heart disease (Zhu et al., 2018), the genetic component of smoking being a causative risk of lung cancer (McKay et al., 2017), and all molecular effects (considered as "molecular" phenotypes) leading to changes in a phenotype of the organism. If one trait is a cause of the other trait, every variant inducing an effect on the first trait also affects the second trait (**Figure 3C**). Moreover, these effect sizes are

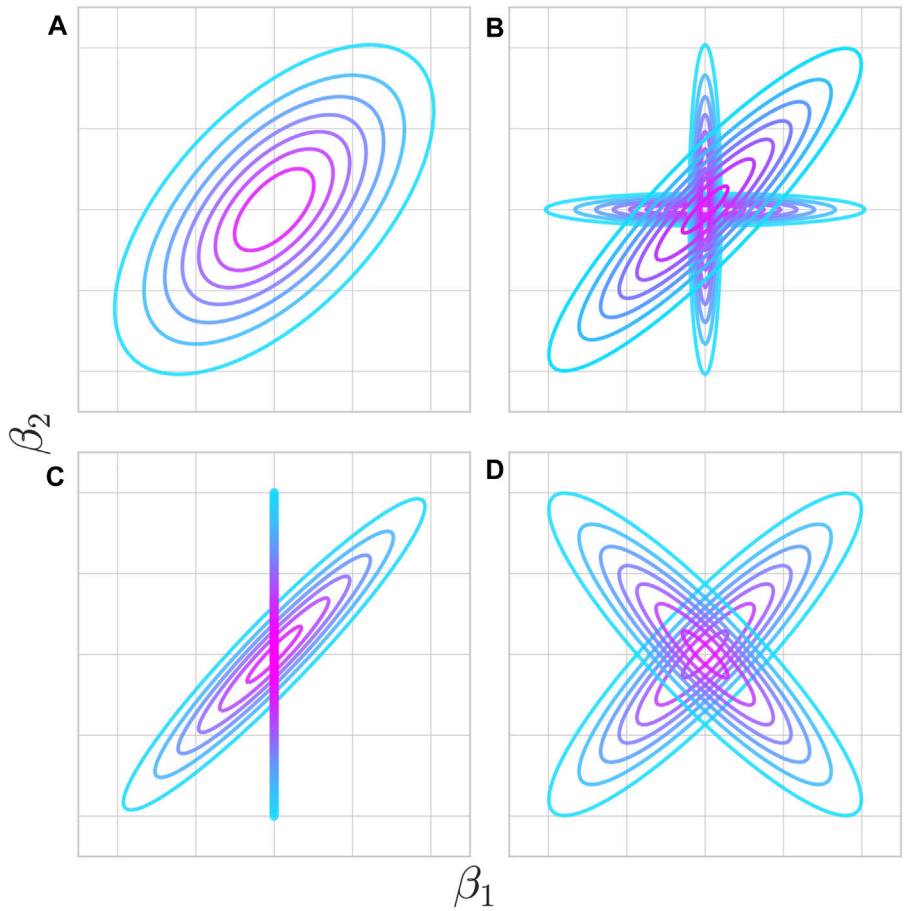

FIGURE 3 | Various potential pleiotropic relationships at individual loci underlie genetic correlations between traits. **(A)** Mutations affecting trait 1 have a tendency to impact trait 2 in a particular direction, although a variety of outcomes are possible through the functional particulars of that change. **(B)** Mutations fall either into a shared or unshared functional pathway between the two traits. Colocalization analysis aims to test which distribution a given QTL comes from. Even though not every mutation is pleiotropic, the two traits are genetically correlated. The proportion of mutations falling into either pathway determines the strength of genetic correlation. **(C)** Trait 1 has a causal impact on trait 2 such that every mutation with a non-zero effect on trait one has a strongly correlated effect on trait 2, but not vice-versa. Mendelian randomization aims to test for the existence and direction of this effect. This also manifests as a genetic correlation at the phenotypic level. **(D)** Individual variants may be pleiotropic, but can result in low or zero genetic correlation if different pathways have opposing effects.

proportional and correspond to the causal effect of the first trait on the second trait. Because the first trait is usually just one of many causes, most variants affecting the second trait would not be expected to have any effect on the first trait. These considerations are a foundation of Mendelian Randomization methods that attempt to infer causal relationships even if genetic associations for the two phenotypes are measured separately in independent datasets (Pingault et al., 2018). This approach relies on a large number of QTLs and does not translate to individual loci.

Many recent studies of pleiotropy, colocalization and causality have focused on molecular phenotypes such as gene expression, chromatin accessibility or DNA methylation (Umans et al., 2020; Vuckovic et al., 2020; Ye et al., 2020; Morabito et al., 2021). Numerous QTLs for various molecular phenotypes have been identified for these classes of traits (most prominently expression QTLs or eQTLs). The main motivation of these studies is to identify the primary molecular events underlying genetic associations with human traits and diseases. However, it is not guaranteed that genetic covariance or colocalization of a molecular trait with a focal trait is indicative of an underlying causal impact of variation in the molecular trait on the focal trait. One example is that changes in BMI actually induce changes in DNA methylation rather than DNA methylation acting as a molecular mechanism mediating genetic effects on BMI (Wahl et al., 2017). The direction of causality was demonstrated by showing that SNPs which predict methylation levels at individual loci did not predict BMI levels, while a genetic risk score for BMI levels did predict methylation levels.

Even in the absence of genetic covariance, molecular effects may induce pleiotropic relationships between two traits. Imagine a scenario where the two traits are both mediated by a large number of molecular phenotypes (activities of many individual genes or other latent factors), but these molecular phenotypes do not exhibit correlated effects on the two traits (**Figure 3D**). In this case, genetic covariance might not exist or be very weak on

aggregate but covariance between absolute (or squared) genetic effects $\mathrm{Cov}[\beta_1^2, \beta_2^2]$ may be substantial. The popular "omnigenic" model offers one version of such a scenario (Boyle et al., 2017). Genetic covariance may also be close to zero if the pleiotropic effect is limited to one or a small number of loci (in other words, with a substantial local genetic covariance and even colocalization in individual loci) (Liu et al., 2019).

These methodological developments and empirical results related to pleiotropy are important in light of the main subject of this review. They motivate consideration of evolutionary models that take into account groups of correlated traits. For causally related traits, selection effects would probably differ depending on whether selection primarily acts on the upstream or downstream trait along the causal chain. An interesting perspective is also brought by the consideration of molecular phenotype. If each molecular phenotype is pleiotropically involved with many downstream organismal phenotypes, and the focal trait is merely one of these, selection coefficients can depend on effect sizes even if the focal trait is neutral. Variants with larger effect sizes on molecular function would be under stronger selection because this molecular function impacts multiple other selected downstream traits in addition to the neutral focal trait.

Few studies have analyzed the effects of pleiotropy on selection by actually incorporating the measured effects of variants on multiple traits. Some mutation accumulation studies have tried to demonstrate whether pleiotropic mutations are under stronger selection. McGuigan et al. (2014) provide some evidence that mutations underlying combinations of correlated gene expression traits in Drosophila serrata are under stronger selection than the average mutation affecting a given trait. In humans, Sanjak et al. (2018) regressed lifetime reproductive success on genetic scores for multiple traits simultaneously in United Kingdom Biobank participants. Compared to the univariate, this full analysis lacked power, but quadratic and linear selection terms did change in both magnitude and sign for some traits, indicating the importance of accounting for pleiotropy. Stern et al. (2021) took a similar approach, but used the shape of genealogies at GWAS loci to look at historical rather than contemporary patterns of directional selection. Again, the authors found that many estimates of selection changed substantially, and were largely attenuated, when accounting for the correlated response in other traits. At the time of writing, no attempt has been made to account for pleiotropy in the alpha model approaches discussed above that have demonstrated negative selection on many human traits.

6 CONSTANCY OF EFFECT SIZES ACROSS TIME AND SPACE

Everything discussed so far has assumed that genetic variants have well-defined additive effects on traits of interest, and that these effects are measurable in contemporary human populations. Although convenient, and the correct starting place for most analyses, recent research has demonstrated that causative variants for many traits and diseases have population-specific effect sizes. Such studies are possible when GWAS for the same traits have been performed in different populations (De Candia et al., 2013; Mancuso et al., 2016). One approach has been to estimate the cross-population genetic correlation, the correlation in causal effect sizes between the different samples, and these estimates are often less than one (Brown et al., 2016; Galinsky et al., 2019). This is most likely due to gene-by-environment (GxE) and gene-by-gene (GxG) interactions, with some effect driven by different measurement practices and diagnosis criteria.

Shi et al. (2021) estimated the impact of different functional annotations on the degree of cross-population effect size correlation of variants within those genomic regions. They found that the squared genetic correlation was depleted most strongly in regions under strong background selection as well as in and around functional elements such as exons, promoters, and enhancers. These regions are also enriched for heritability and, as previous research reviewed here has shown, variants residing there are likely under stronger selection. If a variant has different effect sizes in different contemporary populations, we should be more uncertain about its effect size in the ancestral population where the majority of its existence may have taken place. Cross-population genetic correlation could therefore be used as a measure of the temporal stability of allelic effects. Alternatively, the aggregate pleiotropic effects of an allele may stay roughly constant even as the effects on individual traits vary due to GxE or other factors.

7 CONCLUSION

Direct data on genotype-phenotype associations for numerous human traits have provided an opportunity to investigate which, if any, of the current theoretical models for the maintenance of complex trait variation fit observed genetic architectures. Depending on the degree and nature of pleiotropy, as well as the importance of the focal trait for selection, these models predict the relationship between β and s (Johnson and Barton, 2005). Selection analyses of human GWAS data have consistently demonstrated a negative relationship between effect size magnitudes and allele frequencies, implying that larger effect sizes are associated with stronger selection on average (Zeng et al., 2018; Schoech et al., 2019; Speed et al., 2020; Zeng et al., 2021). Models where the focal trait is neutral, or largely biologically unrelated to any aspect of fitness, are therefore ruled out. Within the class of alpha models, the scaling between β^2 and frequency varies across traits, likely reflecting differences in the DFE and the scaling between β and s. These estimates are difficult to interpret in terms of classical stabilizing selection models, and work is needed to reconcile tractable statistical models of how effect sizes change with frequency with realistic models of selection at the phenotypic level. Studies have also largely been limited to analyzing the average relationship of effect size to frequency. This limits the ability to capture the importance of pleiotropy which should create variance around that average. By directly modeling the variation in genome-wide significant variance contributions, Simons et al. (2018) were able to infer a high degree of pleiotropy for height and BMI.

All the approaches reviewed here infer the nature of selection on GWAS loci by analyzing the distribution of allele frequencies and effect sizes (x, β), with the overall trait heritability sometimes included. Future work along these lines may utilize fine-mapping (Weissbrod et al., 2020) or other techniques to better capture this distribution (O'Connor, 2021). An interesting approach was developed by O'Connor et al. (2019) who estimated the kurtosis of heritability contributions using the LDSC framework to measure trait polygenicity at different allele frequencies and functional genomic annotations. The kurtosis depends on the fourth moment of the distribution of effect sizes, and therefore contains additional information beyond that contained in the alpha model. A low kurtosis, and therefore high polygenicity, of common variants indicated a "flattening" due to selection strongly preventing any large-effect variants from reaching high frequencies. This indicates a high importance of the focal trait for selection, but more thought is needed to tell what degree of pleiotropy is consistent with these results.

The greatest advances in our ability to make sense of the maintenance of complex trait variation will likely come from analyses that utilize variant-level pleiotropy and account for effect sizes that vary across time and space. Methods to investigate pleiotropy in statistical genetics are already well-developed (van Rheenen et al., 2019) but have yet to intersect with analyses of stabilizing or negative selection. Effect size differences between populations are also well-documented (Brown et al., 2016; Shi et al., 2021), but have received less attention, likely in part due to the lack of large GWAS from diverse populations and the difficulty of standardizing measurement for some phenotypes. The portability of polygenic scores is also potentially more strongly impacted by differences in allele frequencies and linkage disequilibrium than effect size variation (Wang et al., 2020), and allele frequencies will differentiate more rapidly under stabilizing or negative selection (Yair and Coop, 2021). However, understanding effect size variation in space and time may ultimately end up being more important for modeling the maintenance of variation in complex traits as well as detecting selection on them (Mathieson, 2021).

AUTHOR CONTRIBUTIONS

EK and SS jointly wrote this review.

FUNDING

This work was supported by NIH grants R35GM127131, RO1MH101244, and R01HG010372 to SRS.

ACKNOWLEDGMENTS

We thank Noah Connally for helpful comments on this manuscript. We are grateful for the suggestions of the editor and two reviewers which greatly improved this review.

REFERENCES

Agarwala, V., Flannick, J., Flannick, J., Sunyaev, S., and Altshuler, D. (2013). Evaluating Empirical Bounds on Complex Disease Genetic Architecture. *Nat. Genet.* 45, 1418–1427. doi:10.1038/ng.2804

Aguet, F., Barbeira, A. N., Bonazzola, R., Jo, B., Kasela, S., Liang, Y., et al. (2020). The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues. *Science* 369, 1318–1330. doi:10.1126/science.aaz1776

Barton, N. H. (1990). Pleiotropic Models of Quantitative Variation. *Genetics* 124, 773–782. doi:10.1093/genetics/124.3.773

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* 169, 1177–1186. doi:10.1016/j.cell.2017.05.038

Brown, B. C., 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

Bulik-Sullivan, B., Finucane, H. K., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., et al. (2015). An Atlas of Genetic Correlations across Human Diseases and Traits. *Nat. Genet.* 47, 1236–1241. doi:10.1038/ng.3406

Bulmer, M. G. (1972). The Genetic Variability of Polygenic Characters under Optimizing Selection, Mutation and Drift. *Genet. Res.* 19, 17–25. doi:10.1017/s0016672300014221

Caballero, A., Tenesa, A., and Keightley, P. D. (2015). The Nature of Genetic Variation for Complex Traits Revealed by GWAS and Regional Heritability Mapping Analyses. *Genetics* 201, 1601–1613. doi:10.1534/genetics.115.177220

Charlesworth, B. (2001). Patterns of Age-specific Means and Genetic Variances of Mortality Rates Predicted by the Mutation-Accumulation Theory of Ageing. *J. Theor. Biol.* 210, 47–65. doi:10.1006/jtbi.2001.2296

Chun, S., Casparino, A., Patsopoulos, N. A., Croteau-Chonka, D. C., Raby, B. A., De Jager, P. L., et al. (2017). Limited Statistical Evidence for Shared Genetic Effects of eQTLs and Autoimmune-Disease-Associated Loci in Three Major Immune-Cell Types. *Nat. Genet.* 49, 600–605. doi:10.1038/ng.3795

Corbett, S., Courtiol, A., Lummaa, V., Moorad, J., and Stearns, S. (2018). The Transition to Modernity and Chronic Disease: Mismatch and Natural Selection. *Nat. Rev. Genet.* 19, 419–430. doi:10.1038/s41576-018-0012-3

Cotsapas, C., Voight, B. F., Rossin, E., Lage, K., Neale, B. M., Wallace, C., et al. (2011). Pervasive Sharing of Genetic Effects in Autoimmune Disease. *Plos Genet.* 7, e1002254. doi:10.1371/journal.pgen.1002254

de Candia, T. R., Lee, S. H., Yang, J., Browning, B. L., Gejman, P. V., Levinson, D. F., et al. (2013). Additive Genetic Variation in Schizophrenia Risk Is Shared by Populations of African and European Descent. *Am. J. Hum. Genet.* 93, 463–470. doi:10.1016/j.ajhg.2013.07.007

Eyre-Walker, A. (2010). Genetic Architecture of a Complex Trait and its Implications for Fitness and Genome-wide Association Studies. *Proc. Natl. Acad. Sci.* 107, 1752–1756. doi:10.1073/pnas.0906182107

Finucane, H. K., Bulik-Sullivan, B., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., et al. (2015). Partitioning Heritability by Functional Annotation Using Genome-wide Association Summary Statistics. *Nat. Genet.* 47, 1228–1235. doi:10.1038/ng.3404

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. (UK: Clarendon.

Fisher, R. (1918). The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Trans. Roy. Soc. Edinb.* 52, 399–433.

Galinsky, K. J., Reshef, Y. A., Finucane, H. K., Loh, P.-R., Zaitlen, N., Patterson, N. J., et al. (2019). Estimating Cross-Population Genetic Correlations of Causal Effect Sizes. *Genet. Epidemiol.* 43, 180–188. doi:10.1002/gepi.22173

Gazal, S., Finucane, H. K., Furlotte, N. A., Loh, P.-R., Palamara, P. F., Liu, X., et al. (2017). Linkage Disequilibrium-dependent Architecture of Human Complex Traits Shows Action of Negative Selection. *Nat. Genet.* 49, 1421–1427. doi:10.1038/ng.3954

Gazal, S., Loh, P.-R., Finucane, H. K., Ganna, A., Schoech, A., Sunyaev, S., et al. (2018). Functional Architecture of Low-Frequency Variants Highlights

Strength of Negative Selection across Coding and Non-coding Annotations. *Nat. Genet.* 50, 1600–1607. doi:10.1038/s41588-018-0231-8

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., et al. (2014). Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *Plos Genet.* 10, e1004383. doi:10.1371/journal.pgen.1004383

Gillespie, J. H. (1984). Pleiotropic Overdominance and the Maintenance of Genetic Variation in Polygenic Characters. *Genetics* 107, 321–330. doi:10.1093/genetics/107.2.321

Gulko, B., Hubisz, M. J., Gronau, I., and Siepel, A. (2015). A Method for Calculating Probabilities of Fitness Consequences for point Mutations across the Human Genome. *Nat. Genet.* 47, 276–283. doi:10.1038/ng.3196

Harpak, A., and Przeworski, M. (2021). The Evolution of Group Differences in Changing Environments. *Plos Biol.* 19, e3001072–14. doi:10.1371/journal.pbio.3001072

Hormozdiari, F., van de Bunt, M., Segrè, A. V., Li, X., Joo, J. W. J., Bilow, M., et al. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* 99, 1245–1260. doi:10.1016/j.ajhg.2016.10.003

Johnson, T., and Barton, N. (2005). Theoretical Models of Selection and Mutation on Quantitative Traits. *Phil. Trans. R. Soc. B* 360, 1411–1425. doi:10.1098/rstb.2005.1667

Jordan, D. M., Verbanck, M., and Do, R. (2019). HOPS: a Quantitative Score Reveals Pervasive Horizontal Pleiotropy in Human Genetic Variation Is Driven by Extreme Polygenicity of Human Traits and Diseases. *Genome Biol.* 20, 222. doi:10.1186/s13059-019-1844-7

Keightley, P. D., and Hill, W. G. (1988). Quantitative Genetic Variability Maintained by Mutation-Stabilizing Selection Balance in Finite Populations. *Genet. Res.* 52, 33–43. doi:10.1017/S0016672300027282

Keightley, P. D., and Hill, W. G. (1990). Variation Maintained in Quantitative Traits with Mutation-Selection Balance: Pleiotropic Side-Effects on Fitness Traits. *Proc. R. Soc. B: Biol. Sci.* 242, 95–100. doi:10.1098/rspb.1990.0110

Kiezun, A., Pulit, S. L., Francioli, L. C., van Dijk, F., Swertz, M., Boomsma, D. I., et al. (2013). Deleterious Alleles in the Human Genome Are on Average Younger Than Neutral Alleles of the Same Frequency. *Plos Genet.* 9, e1003301–12. doi:10.1371/journal.pgen.1003301

Kingsolver, J. G., Hoekstra, H. E., Hoekstra, J. M., Berrigan, D., Vignieri, S. N., Hill, C. E., et al. (2001). The Strength of Phenotypic Selection in Natural Populations. *The Am. Naturalist* 157, 245–261. doi:10.1086/319193

Koch, E. M. (2019). The Effects of Demography and Genetics on the Neutral Distribution of Quantitative Traits. *Genetics* 211, 1371–1394. doi:10.1534/genetics.118.301839

Kondrashov, A. S., and Turelli, M. (1992). Deleterious Mutations, Apparent Stabilizing Selection and the Maintenance of Quantitative Variation. *Genetics* 132, 603–618. doi:10.1093/genetics/132.2.603

Lande, R. (1976b). The Maintenance of Genetic Variability by Mutation in a Polygenic Character with Linked Loci. *Genet. Res.* 26, 221–235. doi:10.1017/s0016672300016037

Lande, R., and Arnold, S. (1983). The Measurement of Selection on Correlated Characters. *Evolution* 37, 1210. doi:10.2307/2408842

Lande, R. (1976a). Natural Selection and Random Genetic Drift in Phenotypic Evolution. *Evolution* 30, 314. doi:10.2307/2407703

Lee, S. H., Ripke, S., Neale, B. M., Faraone, S. V., Purcell, S. M., Perlis, R. H., et al. (2013). Genetic Relationship between Five Psychiatric Disorders Estimated from Genome-wide SNPs. *Nat. Genet.* 45, 984–994. doi:10.1038/ng.2711

Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M., and Wray, N. R. (2012). 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

Lincoln, M. R., Connally, N., Axisa, P.-P., Gasperi, C., Mitrovic, M., van Heel, D., et al. (2021). Joint Analysis Reveals Shared Autoimmune Disease Associations and Identifies Common Mechanisms. *medRxiv*, 1–15. doi:10.1101/2021.05.13.21257044

Liu, X., Li, Y. I., and Pritchard, J. K. (2019). Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* 177, 1022–1034.e6. doi:10.1016/j.cell.2019.04.014

Lourenço, J., Galtier, N., and Glémin, S. (2011). Complexity, Pleiotropy, and the Fitness Effect of Mutations. *Evolution* 65, 1559–1571. doi:10.1111/j.1558-5646.2011.01237.x

Lynch, M., and Hill, W. G. (1986). Phenotypic Evolution by Neutral Mutation. *Evolution* 40, 915–935. doi:10.1111/j.1558-5646.1986.tb00561.x

Mancuso, N., Rohland, N., Rohland, N., Rand, K. A., Tandon, A., Allen, A., et al. (2016). The Contribution of Rare Variation to Prostate Cancer Heritability. *Nat. Genet.* 48, 30–35. doi:10.1038/ng.3446

Martin, G., and Lenormand, T. (2006). A General Multivariate Extension of Fisher's Geometrical Model and the Distribution of Mutation Fitness Effects across Species. *Evolution* 60, 893–907. doi:10.1111/j.0014-3820.2006.tb01169.x

Maruyama, T. (1974). The Age of a Rare Mutant Gene in a Large Population. *Am. J. Hum. Genet.* 26, 669–673.

Mathieson, I. (2021). The Omnigenic Model and Polygenic Prediction of Complex Traits. *Am. J. Hum. Genet.* 108, 1558–1563. doi:10.1016/j.ajhg.2021.07.003

McGuigan, K., Collet, J. M., Allen, S. L., Chenoweth, S. F., and Blows, M. W. (2014). Pleiotropic Mutations Are Subject to strong Stabilizing Selection. *Genetics* 197, 1051–1062. doi:10.1534/genetics.114.165720

McKay, J. D., Hung, R. J., Han, Y., Zong, X., Carreras-Torres, R., Christiani, D. C., et al. (2017). Large-scale Association Analysis Identifies New Lung Cancer Susceptibility Loci and Heterogeneity in Genetic Susceptibility across Histological Subtypes. *Nat. Genet.* 49, 1126–1132. doi:10.1038/ng.3892

Morabito, S., Miyoshi, E., Michael, N., Shahin, S., Martini, A. C., Head, E., et al. (2021). Single-nucleus Chromatin Accessibility and Transcriptomic Characterization of Alzheimer's Disease. *Nat. Genet.* 53, 1143–1155. doi:10.1038/s41588-021-00894-z

O'Connor, L. J., Schoech, A. P., Hormozdiari, F., Gazal, S., Patterson, N., and Price, A. L. (2019). Extreme Polygenicity of Complex Traits Is Explained by Negative Selection. *Am. J. Hum. Genet.* 105, 456–476. doi:10.1016/j.ajhg.2019.07.003

O'Connor, L. J. (2021). The Distribution of Common-Variant Effect Sizes. *Nat. Genet.* 53, 1243–1249. doi:10.1038/s41588-021-00901-3

Pavard, S., and Coste, C. F. D. (2021). Evolutionary Demographic Models Reveal the Strength of Purifying Selection on Susceptibility Alleles to Late-Onset Diseases. *Nat. Ecol. Evol.* 5, 392–400. doi:10.1038/s41559-020-01355-2

Pingault, J.-B., O'Reilly, P. F., Schoeler, T., Ploubidis, G. B., Rijsdijk, F., and Dudbridge, F. (2018). Using Genetic Data to Strengthen Causal Inference in Observational Research. *Nat. Rev. Genet.* 19, 566–580. doi:10.1038/s41576-018-0020-3

Pritchard, J. K. (2001). Are Rare Variants Responsible for Susceptibility to Complex Diseases? *Am. J. Hum. Genet.* 69, 124–137. doi:10.1086/321272

Rands, C. M., Meader, S., Ponting, C. P., and Lunter, G. (2014). 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. *Plos Genet.* 10, e1004525. doi:10.1371/journal.pgen.1004525

Reich, D. E., and Lander, E. S. (2001). On the Allelic Spectrum of Human Disease. *Trends Genet.* 17, 502–510. doi:10.1016/S0168-9525(01)02410-6

Robertson, A. (1956). The Effect of Selection against Extreme Deviants Based on Deviation or on Homozygosis. *J. Genet.* 54, 236–248. doi:10.1007/bf02982779

Sanjak, J. S., Sidorenko, J., Robinson, M. R., Thornton, K. R., and Visscher, P. M. (2018). Evidence of Directional and Stabilizing Selection in Contemporary Humans. *Proc. Natl. Acad. Sci. USA* 115, 151–156. doi:10.1073/pnas.1707227114

Schoech, A. P., Jordan, D. M., Loh, P.-R., Gazal, S., O'Connor, L. J., Balick, D. J., et al. (2019). Quantification of Frequency-dependent Genetic Architectures in 25 UK Biobank Traits Reveals Action of Negative Selection. *Nat. Commun.* 10, 790. doi:10.1038/s41467-019-08424-6

Sella, G., and Barton, N. H. (2019). Thinking about the Evolution of Complex Traits in the Era of Genome-wide Association Studies. *Annu. Rev. Genom. Hum. Genet.* 20, 461–493. doi:10.1146/annurev-genom-083115-022316

Shi, H., Gazal, S., Kanai, M., Koch, E. M., Schoech, A. P., Siewert, K. M., et al. (2021). Population-specific Causal Disease Effect Sizes in Functionally Important Regions Impacted by Selection. *Nat. Commun.* 12, 1098. doi:10.1038/s41467-021-21286-1

Shi, H., Mancuso, N., Spendlove, S., and Pasaniuc, B. (2017). Local Genetic Correlation Gives Insights into the Shared Genetic Architecture of Complex Traits. *Am. J. Hum. Genet.* 101, 737–751. doi:10.1016/j.ajhg.2017.09.022

Simons, Y. B., Bullaughey, K., Hudson, R. R., and Sella, G. (2018). A Population Genetic Interpretation of GWAS Findings for Human Quantitative Traits. *Plos Biol.* 16, e2002985. doi:10.1371/journal.pbio.2002985

Speed, D., Holmes, J., and Balding, D. J. (2020). Evaluating and Improving Heritability Models Using Summary Statistics. *Nat. Genet.* 52, 458–462. doi:10.1038/s41588-020-0600-y

Stearns, F. W. (2010). One Hundred Years of Pleiotropy: A Retrospective. *Genetics* 186, 767–773. doi:10.1534/genetics.110.122549

Stern, A. J., Speidel, L., Zaitlen, N. A., and Nielsen, R. (2021). Disentangling Selection on Genetically Correlated Polygenic Traits via Whole-Genome Genealogies. *Am. J. Hum. Genet.* 108, 219–239. doi:10.1016/j.ajhg.2020.12.005

Umans, B. D., Battle, A., and Gilad, Y. (2021). Where Are the Disease-Associated eQTLs? *Trends Genet.* 37, 109–124. doi:10.1016/j.tig.2020.08.009

van Rheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H., and Wray, N. R. (2019). Genetic Correlations of Polygenic Disease Traits: from Theory to Practice. *Nat. Rev. Genet.* 20, 567–581. doi:10.1038/s41576-019-0137-z

Visscher, P. M., Hill, W. G., and Wray, N. R. (2008). Heritability in the Genomics Era - Concepts and Misconceptions. *Nat. Rev. Genet.* 9, 255–266. doi:10.1038/nrg2322

Vuckovic, D., Bao, E. L., Akbari, P., Lareau, C. A., Mousas, A., Jiang, T., et al. (2020). The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell* 182, 1214–e11. doi:10.1016/j.cell.2020.08.008

Wahl, S., Drong, A., Lehne, B., Loh, M., Scott, W. R., Kunze, S., et al. (2017). Epigenome-wide Association Study of Body Mass index, and the Adverse Outcomes of Adiposity. *Nature* 541, 81–86. doi:10.1038/nature20784

Walsh, B., and Lynch, M. (2018). *Evolution and Selection of Quantitative Traits*. Oxford: OUP Oxford.

Wang, Y., Guo, J., Ni, G., Yang, J., Visscher, P. M., and Yengo, L. (2020). Theoretical and Empirical Quantification of the Accuracy of Polygenic Scores in Ancestry Divergent Populations. *Nat. Commun.* 11, 3865–3869. doi:10.1038/s41467-020-17719-y

Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., de Leeuw, C., Polderman, T. J. C., et al. (2019). A Global Overview of Pleiotropy and Genetic Architecture in Complex Traits. *Nat. Genet.* 51, 1339–1348. doi:10.1038/s41588-019-0481-0

Waxman, D., and Peck, J. R. (1998). Pleiotropy and the Preservation of Perfection. *Science* 279, 1210–1213. doi:10.1126/science.279.5354.1210

Weissbrod, O., Hormozdiari, F., Benner, C., Cui, R., Ulirsch, J., Gazal, S., et al. (2020). Functionally Informed fine-mapping and Polygenic Localization of Complex Trait Heritability. *Nat. Genet.* 52, 1355–1363. doi:10.1038/s41588-020-00735-5

Wingreen, N. S., Miller, J., and Cox, E. C. (2003). Scaling of Mutational Effects in Models for Pleiotropy. *Genetics* 164, 1221–1228. doi:10.1093/genetics/164.3.1221

Wright, A., Charlesworth, B., Rudan, I., Carothers, A., and Campbell, H. (2003). A Polygenic Basis for Late-Onset Disease. *Trends Genet.* 19, 97–106. doi:10.1016/s0168-9525(02)00033-1

Wright, S. (1935). The Analysis of Variance and the Correlations between Relatives with Respect to Deviations from an Optimum. *Journ Genet.* 30, 243–256. doi:10.1007/BF02982239

Yair, S., and Coop, G. (2021). Population Differentiation of Polygenic Score Predictions under Stabilizing Selection. *bioRxiv*. 2021.09.10.459833.

Yang, J., Bakshi, A., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A. E., et al. (2015). Genetic Variance Estimation with Imputed Variants Finds Negligible Missing Heritability for Human Height and Body Mass index. *Nat. Genet.* 47, 1114–1120. doi:10.1038/ng.3390

Ye, Y., Zhang, Z., Liu, Y., Diao, L., and Han, L. (2020). A Multi-Omics Perspective of Quantitative Trait Loci in Precision Medicine. *Trends Genet.* 36, 318–336. doi:10.1016/j.tig.2020.01.009

Zeng, J., De Vlaming, R., Wu, Y., Robinson, M. R., Lloyd-Jones, L. R., Yengo, L., et al. (2018). Signatures of Negative Selection in the Genetic Architecture of Human Complex Traits. *Nat. Genet.* 50, 746–753. doi:10.1038/s41588-018-0101-4

Zeng, J., Xue, A., Jiang, L., Lloyd-Jones, L. R., Wu, Y., Wang, H., et al. (2021). Widespread Signatures of Natural Selection across Human Complex Traits and Functional Genomic Categories. *Nat. Commun.* 12, 1164. doi:10.1038/s41467-021-21446-3

Zhang, X.-S., and Hill, W. G. (2002). Joint Effects of Pleiotropic Selection and Stabilizing Selection on the Maintenance of Quantitative Genetic Variation at Mutation-Selection Balance. *Genetics* 162, 459–471. doi:10.1093/genetics/162.1.459

Zhang, X.-S., and Hill, W. G. (2003). Multivariate Stabilizing Selection and Pleiotropy in the Maintenance of Quantitative Genetic Variation. *Evol* 57, 1761–1775. doi:10.1554/02-587

Zhang, Y., Qi, G., Park, J.-H., and Chatterjee, N. (2018). Estimation of Complex Effect-Size Distributions Using Summary-Level Statistics from Genome-wide Association Studies across 32 Complex Traits. *Nat. Genet.* 50, 1318–1326. doi:10.1038/s41588-018-0193-x

Zhu, Z., Zheng, Z., Zhang, F., Wu, Y., Trzaskowski, M., Maier, R., et al. (2018). Causal Associations between Risk Factors and Common Diseases Inferred from GWAS Summary Data. *Nat. Commun.* 9, 224. doi:10.1038/s41467-017-02317-2

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 handling editor declared a past co-authorship with one of the authors (SS).

GLOSSARY

Causative Allele The allele which causally affects the focal trait. Due to linkage disequilibrium, many alleles at a GWAS-identified locus are highly correlated. The causative allele refers only to the one which causally affects the trait.

Distribution of Mutational Effects The distribution from which the phenotypic effects of new mutations are drawn. This can include the focal trait as well as related pleiotropic ones.

Effect Size Magnitude The absolute value of the effect size of an allele. It is often useful to ignore the direction of effect that an allele has on the trait.

Focal Trait All studies in quantitative genetics must choose some measurable aspects of biology to focus on. This can be something of obvious importance like diabetic status, or could be simply something easily queried in a biobank. Analyzed one at a time, we call the current trait the focal trait.

Genetic Architecture The joint distribution of allele frequencies and effect sizes in a population or sample. This determines how much different frequency and effect size ranges contribute to heritability, and answers questions surrounding the importance of rare versus common variants;

genetic covariance, The covariance between the effects different genotypes have on two traits. This measures the propensity for an individual with a high genetic value for one trait to also have a high (or low) genetic value for the second. It averages over all alleles and their effects on both traits, scaled by their contributions to the genetic variance.

Genetic Risk Score A phenotypic prediction calculated for an individual using a weighted sum of the estimated effect sizes of variants found in that individuals genome.

Molecular Phenotype Phenotypes such as gene expression, methylation levels, or metabolite concentration that are measured at the molecular level. These are hoped to represent "low-level" traits that mediate the effects of genetic variants on other phenotypes.

Overdominant Selection where the heterozygous genotype has higher average fitness than either of the two homozygotes.

Pairwise Coalescent Time The amount of time it takes two sampled loci to find a common ancestor going backwards in time. The longer this time, the more likely it is that mutations occur to differentiate the two loci.

QTL Quantitative trait locus. A region in the genome that has been statistically associated with a quantitative trait.

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 January 2022
doi: 10.3389/fgene.2021.719791

Imputation Performance in Latin American Populations: Improving Rare Variants Representation With the Inclusion of Native American Genomes

Andrés Jiménez-Kaufmann[1], Amanda Y. Chong[2], Adrián Cortés[2], Consuelo D. Quinto-Cortés[1], Selene L. Fernandez-Valverde[1], Leticia Ferreyra-Reyes[3], Luis Pablo Cruz-Hervert[3], Santiago G. Medina-Muñoz[1], Mashaal Sohail[1,4], María J. Palma-Martinez[1], Gudalupe Delgado-Sánchez[3], Norma Mongua-Rodríguez[3], Alexander J. Mentzer[2], Adrian V. S. Hill[2,5], Hortensia Moreno-Macías[6,7], Alicia Huerta-Chagoya[6], Carlos A. Aguilar-Salinas[8,9], Michael Torres[1], Hie Lim Kim[10,11,12], Namrata Kalsi[10,11], Stephan C. Schuster[10,11,12], Teresa Tusié-Luna[6,13], Diego Ortega Del-Vecchyo[14], Lourdes García-García[3] and Andrés Moreno-Estrada[1]**

[1]Laboratorio Nacional de Genómica para la Biodiversidad (UGA-LANGEBIO), Unidad de Genómica Avanzada, Irapuato, Mexico, [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, [3]Instituto Nacional de Salud Pública, Cuernavaca, Mexico, [4]Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Mexico, [5]Nuffield Department of Medicine, The Jenner Institute, University of Oxford, Oxford, United Kingdom, [6]Unidad de Biología Molecular y Medicina Genómica, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán (INCMNSZ), Mexico City, Mexico, [7]Departamento de Economía, Universidad Autónoma Metropolitana, Mexico City, Mexico, [8]Departamento de Endocrinología y Metabolismo, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Unidad de Investigación de Enfermedades Metabólicas, Mexico City, Mexico, [9]Tecnológico de Monterrey, Escuela de Medicina y Ciencias de la Salud, Monterrey, Mexico, [10]Singapore Centre on Environmental Life Sciences Engineering, Nanyang Technological University, Singapore, [11]GenomeAsia 100K (GA100K) Consortium, Singapore, [12]School of Biological Science, Nanyang Technological University, Singapore, [13]Instituto de Investigaciones Biomédicas de la UNAM, Mexico City, Mexico, [14]Laboratorio Internacional de Investigación sobre el Genoma Humano (LIIGH), UNAM, Juriquilla, Mexico

Current Genome-Wide Association Studies (GWAS) rely on genotype imputation to increase statistical power, improve fine-mapping of association signals, and facilitate meta-analyses. Due to the complex demographic history of Latin America and the lack of balanced representation of Native American genomes in current imputation panels, the discovery of locally relevant disease variants is likely to be missed, limiting the scope and impact of biomedical research in these populations. Therefore, the necessity of better diversity representation in genomic databases is a scientific imperative. Here, we expand the 1,000 Genomes reference panel (1KGP) with 134 Native American genomes (1KGP + NAT) to assess imputation performance in Latin American individuals of mixed ancestry. Our panel increased the number of SNPs above the GWAS quality threshold, thus improving statistical power for association studies in the region. It also increased imputation accuracy, particularly in low-frequency variants segregating in Native American ancestry tracts. The improvement is subtle but consistent across countries and proportional to the number of genomes added from local source populations. To project the potential improvement with a higher number of reference genomes, we performed simulations and found that at least 3,000 Native American genomes are

OPEN ACCESS

Edited by:
Tony Merriman, University of Otago, New Zealand

Reviewed by:
Inaho Dnjoh, Tohoku University, Japan
Mohamad Saad, Qatar Computing Research Institute, Qatar

***Correspondence:**
Andrés Moreno-Estrada
andres.moreno@cinvestav.mx
Lourdes García-García
garcigarml@gmail.com

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *03 June 2021*
Accepted: *01 November 2021*
Published: *03 January 2022*

Citation:
Jiménez-Kaufmann A, Chong AY, Cortés A, Quinto-Cortés CD, Fernandez-Valverde SL, Ferreyra-Reyes L, Cruz-Hervert LP, Medina-Muñoz SG, Sohail M, Palma-Martinez MDJ, Delgado-Sánchez G, Mongua-Rodríguez N, Mentzer AJ, Hill AVS, Moreno-Macías H, Huerta-Chagoya A, Aguilar-Salinas CA, Torres M, Kim HL, Kalsi N, Schuster SC, Tusié-Luna T, Del-Vecchyo DO, García-García L and Moreno-Estrada A (2022) Imputation Performance in Latin American Populations: Improving Rare Variants Representation With the Inclusion of Native American Genomes. Front. Genet. 12:719791. doi: 10.3389/fgene.2021.719791

needed to equal the imputation performance of variants in European ancestry tracts. This reflects the concerning imbalance of diversity in current references and highlights the contribution of our work to reducing it while complementing efforts to improve global equity in genomic research.

Keywords: Imputation, reference panels, GWAS, Native American ancestry, Latin Americans, underrepresented populations

INTRODUCTION

Over the past years, GWAS have identified thousands of genetic associations to multiple phenotypes (MacArthur et al., 2017; Visscher et al., 2017), targets for potential new drugs (Agrawal and Brown 2014; Flannick et al., 2014; Nelson et al., 2015), and facilitated disease stratification (Chatterjee, Shi, and García-Closas 2016). However, most GWAS have been performed in populations with European ancestry (Popejoy and Fullerton 2016). Unfortunately, the findings of large-scale GWAS performed in populations of European descent have limited portability to other ancestry groups (Duncan et al., 2019; Sirugo, Williams, and Tishkoff 2019) due to population substructure. This represents a major limitation in the case of Latin American populations as they are the result of recent admixture primarily between Native American, European, and African populations, and only 1.3% of both discovery and replication studies have been performed in these populations (Mills and Rahal 2019). Furthermore, the genetic composition of Latin American populations is heterogeneous between countries (Chacón-Duque et al., 2018; Soares-Souza et al., 2018) and within countries (Moreno-Estrada et al., 2014; Harris et al., 2018; Kehdy et al., 2015). Different demographic histories often lead to different associated variants to a given phenotype (Martin et al., 2017). For example, variants in the *SLC16A11* gene have been associated with an increased risk of diabetes in Mexicans and appear to be segregating at low frequency in Latin American populations specifically (SIGMA Type 2 Diabetes Consortium et al., 2014). Likewise, risk variants of renal disease in *APOL1* associated with renal disease in west African populations are also found in the Americas as a result of the Transatlantic slave trade, differentially shaping the frequency spectrum of disease variants among Afro-descendent Latino populations (Nadkarni et al., 2018). If the current bias in catalogs of human variation persists, many population-specific variants will be overlooked, and precision medicine strategies will not benefit all populations equally (Martin et al., 2019).

A critical step when performing a GWAS is genotype imputation, which leverages linkage disequilibrium (LD) structure and haplotype sharing to estimate untyped variation in a SNP array based on a reference panel (Marchini et al., 2007). Genotype imputation increases statistical power, improves fine-mapping of association signals, and facilitates meta-analysis (Marchini and Howie 2010). Currently, available imputation panels do not have an explicit representation of Native American genomes. A previous study showed that in Latin American populations, SNPs in chromosomal segments with Native American ancestry have reduced imputation quality compared to those in chromosomal segments of European ancestry (Martin et al., 2017). Therefore, association signals coming from chromosomal segments with Native American ancestry will be harder to detect. This limits the scope and impact of biomedical research in the region.

Several projects and initiatives around the world are contributing to revert this trend (GenomeAsia100K Consortium 2019; Mulder et al., 2018; Gurdasani et al., 2015; Magalhães et al., 2018). For example, the Ugandan Genome Resource (Gurdasani et al., 2019) comprises genome-wide data for 6,400 individuals, including a subset of 1,978 whole genomes, which is enabling researchers to explore the genetic substructure of the region, improve imputation in African populations, and foster the discovery of novel association signals. In Latin America, recent sequencing efforts have generated whole-genome data from dozens of Native American genomes, including the Peruvian Genome Project (Harris et al., 2018) and the 12G and 100G-MX Projects (Romero-Hidalgo et al., 2017; Aguilar-Ordoñez et al., 2021) from the National Institute of Genomic Medicine (INMEGEN) in Mexico. However, only a subset of the existing generated data is available to the scientific community given the data sharing mechanisms implemented in each country. An ongoing multi-institutional effort in Mexico, the MX Biobank Project, is generating genome-wide data for more than 6,000 individuals nationwide, including 50 whole genomes of Native American ancestry representing the genetic variation of indigenous diversity within Mexico (http://www.mxbiobankproject.org). At a global scale, the inclusion of diverse populations in disease association research has been well demonstrated by the PAGE study (Wojcik et al., 2019), which combines genome-wide data for 49,839 individuals with diverse ancestries, enabling the discovery of novel association signals to well-studied phenotypes. Here, we combine novel and publicly available data from multiple sources to build a population-specific reference panel of Native American variation aimed at improving imputation performance in Latin American populations by expanding the current and widely used reference of the 1,000 Genomes Project (1KGP) (The 1000 Genomes Project Consortium et al., 2015) with 134 Native American genomes. Using a demographic simulation framework, we also explore the number of additional reference genomes that should be sequenced to bridge the gap in imputation quality between different ancestries. Strengthening these efforts in diverse populations is not only a question of equality in genomics, but it also entails the scientific advantage of furthering our understanding of complex phenotypes in biomedical research.

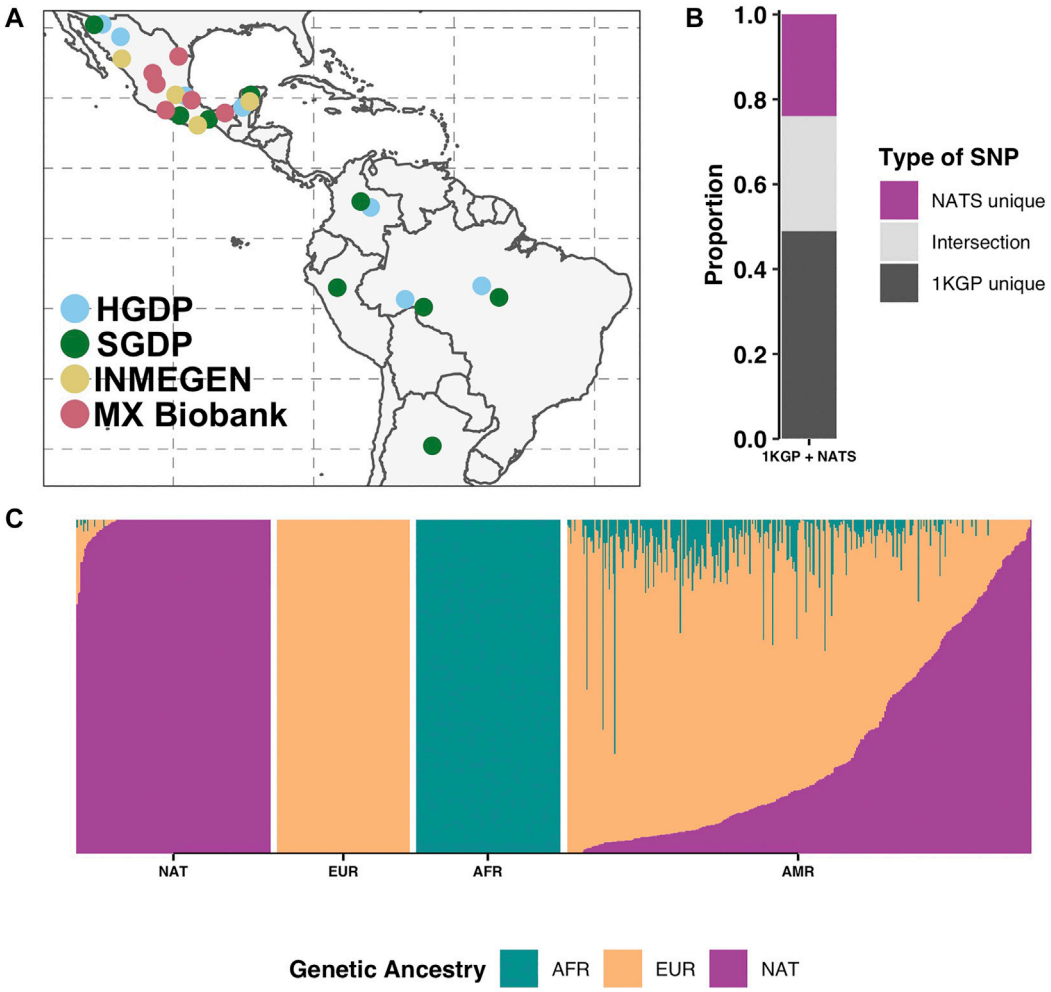

FIGURE 1 | Native American reference panel (NATS). **(A)** Geographical sampling locations of the NATS reference panel. Colors represent the four data sources: HGDP (61) (Bergström et al., 2020), SGDP (11) (Mallick et al., 2016), INMEGEN (12) (Romero-Hidalgo et al., 2017), and MX Biobank (50) totaling 134 genomes. **(B)** SNP proportions of the union of 1KGP and NATS (1KGP + NATS) by SNP sharing categories. We show the proportion of SNPs unique to 1KGP, SNPs unique to the NATS panel, and the intersection. **(C)** Unsupervised ADMIXTURE analysis at $K = 3$ of the NATS reference panel (far left, $N = 134$) together with 104 European (CEU), 113 African (YRI), and 347 admixed Latin American (AMR) samples from 1KGP. Genetic ancestry abbreviations: AFR—African, EUR—European, NAT—Native American.

MATERIALS AND METHODS

Building a Native American Reference Panel

Our panel consists of 134 Native American individuals broadly distributed across the continent (**Figure 1**; **Supplementary Tables S1, S2**). We gathered publicly available whole-genome sequencing (WGS) data from HGDP (Bergström et al., 2020) (61 individuals), SGDP (Mallick et al., 2016) (11 individuals), and INMEGEN (Romero-Hidalgo et al., 2017) (12 individuals). Additionally, we whole-genome sequenced the genome of 50 Mexican individuals with the highest Native American ancestry (99% on average) from the MX Biobank Project (http://www.mxbiobankproject.org). These were selected to maximize indigenous ancestry and geographical representation across Mexico. Individual genetic ancestry proportions were estimated using ADMIXTURE (Alexander, Novembre, and Lange 2009) at $K = 3$ using Utah residents with Northern and Western European ancestry (CEU), Yoruba in Ibadan, Nigeria (YRI), and the Latin Americans (AMR) of 1KGP as references.

To construct the panel, we restricted the datasets to biallelic SNPs with no missing data in any individual across each data source. This was done for all four data sources (**Supplementary Table S3**). The data processing was done using *VCFtools v0.1.17* (Danecek et al., 2011). Then, we merged the data using *bcftools v1.9* (Danecek et al., 2021) using the flag --missing-to-ref that fills the missing positions in one panel but present in another with homozygous reference. To minimize any potential bias introduced with this strategy, we made sure that any previously removed position in any of the sources was not present in the final freeze. The final dataset consists of a total of 10,981,451 SNPs.

Finally, we phased the data using *SHAPEIT2 v2. r837* (Delaneau et al., 2014) using the following flags: --window 0.5 --states 500 --burn 10 --prune 10 --main 50. Then, we converted the data to the reference format used by *IMPUTE2* (Howie et al., 2012). We named this panel NATS.

Whole-Genome Sequencing and Variant Calling

Fifty individuals from the MX Biobank Project were sequenced at 40X on Illumina HiSeqX instruments using dual indexed barcodes. The raw reads were aligned to the human genome assembly GRCh37 using *BWA v.0.7.17-r1198-dirty* (Li and Durbin 2009). We added the mate tags with *Samblaster v0.1.24* (Faust and Hall 2014) and used *Sambamba v0.7.1* (Tarasov et al., 2015) for file conversion and sorting. To generate the alignment statistics, we used *Samtools v1.10* (Li 2011) with the option *depth -a*. Finally, we performed variant calling and generated the final gvcf files with *GATK v4.1.9.0* (McKenna et al., 2010) using the human genome assembly GRCh37 as the reference genome. Details are available as part of the Supplementary Material (**Supplementary Table S2**; **Supplementary Figure S9**).

Creating a SNP Array Subset From WGS Data for Imputation Performance Evaluation

To evaluate the performance of our panel, we used WGS data from the 347 AMR individuals in 1KGP as target individuals for imputation. Namely, Puerto Ricans in Puerto Rico (PUR), Peruvians in Lima (PEL), Colombian in Medellin (CLM), and Mexican ancestry in Los Angeles (MXL). We generated an array dataset by subsetting the AMR individual genomes to the existing positions in the Multi-Ethnic Global Array (MEGA) using *VCFtools v.0.1.17* and saved the removed positions from the WGS data to use for imputation validation. Illumina's MEGA array includes nearly 1.8 M markers (1,779,819) genome-wide distributed and was designed to leverage SNP content from various global sequencing efforts, mostly Phase 3 of the 1,000 Genomes Project. To better approximate a real scenario, we unphased the array dataset with *Plink v1.9* (Chang et al., 2015) by transforming the data to bed format. Finally, we phased the dataset again with *SHAPEIT2 v2. r837* using 1KGP as a phasing reference.

Local Ancestry Inference

To evaluate the performance by ancestry, we deconvoluted local ancestry for the Latin American individuals from 1,000 Genomes. We used 70 YRI individuals in 1KGP as the African reference, 70 CEU individuals from 1KGP as the European reference, and 70 Native American individuals from (Moreno-Estrada et al., 2014) as the Native American reference. The selected individuals had the highest African, European, and Native American genetic components, respectively. We used the PopPhased version of *RFMix v.1.5.4* (Maples et al., 2013) with the following flags: -w 0.2 -e 0 --forward-backward.

Imputation and Imputation Performance

We implemented a leave-one-out strategy for imputation. Namely, the target individual was removed from the 1KGP reference. We performed imputation with *IMPUTE2* for chromosomes 2 and 9. These chromosomes, being the largest and of intermediate size, respectively, were selected to ensure a representative subset of variants across the genome while keeping the project within the available computational capacity. We used 1KGP and 1KGP + NATS as reference panels. When using 1KGP as a reference, we used the flag --k_haps 1,000, and when using 1KGP + NATS, we used the flags --merge-ref-panels and --k_haps 1,250.

We obtained the imputed dosages with the formula: P(Aa) + 2P(aa). We computed the Pearson squared correlation (r^2) between the imputed dosages and the real dosages for each individual using R software. Overall imputation accuracy was stratified by minor allele frequency and local ancestry diplotype (AFR_AFR, AFR_EUR, AFR_NAT, EUR_EUR, EUR_NAT, NAT_NAT). We also compared the number of SNPs above the GWAS quality threshold (MAF >=0.01 and INFO >0.3) for both reference panels stratified by local ancestry diplotype in the target individuals.

Demographic Simulation

We simulated neutral genetic sequence data under a coalescent model. We used the *msprime* (Kelleher, Etheridge, and McVean 2016) option of *stdpopsim* (Adrion et al., 2020) to simulate a previously defined American admixture model for Latinos (Browning et al., 2018). It models African, European, and Asian (as Native American proxy) demographic history and an admixture event taking place 12 generations ago. In the absence of realistic admixture models that use Native American instead of East Asian genomes as proxy in the simulations and based on the framework described by Browning et al. (2018), we will now refer to the simulated Asian population as Native American for the purpose of predicting imputation performance at incremental numbers of reference genomes in a similar scenario to the Latin American admixture. The simulated admixed population ancestral proportions are 1/6 African, ⅓ European, and ½ Native American. In total, we simulated chromosome 9 for 661 Africans, 503 Europeans, 3,000 Native Americans, and 657 admixed individuals. We selected all the Africans, Europeans, and the first 347 admixed individuals to serve as the base reference panel (note that these numbers mirror the sample sizes of 1KGP for each ancestry). The remaining 300 admixed individuals were used as imputation targets, and incremental subsets of the 3,000 Native American genomes were added sequentially to the base reference panel.

To simulate genotype array data for the target individuals, we downsampled the simulated neutral sequence to match the allele frequency spectrum in European populations of 1KGP and the average distance between SNPs of the MEGA array. We used the European populations in 1KGP to mirror the ascertainment bias towards European ancestry in current array designs. We estimated local ancestry using *RFMix* for the 300 admixed individuals used as imputation targets. We randomly selected

TABLE 1 | SNPs above the standard quality threshold using both panels after imputing missing variants. We show the average number of SNPs with MAF >= 0.01 and INFO >= 0.3 using both reference panels and the overall proportion of Native American ancestry of the population. *p*-value was calculated with a two-tailed paired *t*-test. The average number of SNPs with MAF <0.01 and INFO >0.3 for both panels is shown in **Supplementary Table S4**.

Population	SNPs above quality threshold (1KGP)	SNPs above quality threshold (1KGP + NATS)	Increase of SNPs using 1KGP + NATS	Average proportion of Nat. American ancestry
Peru (PEL)	244,818	248,087	3,269 (*p*-value = 2.03e-49)	0.70
Mexico (MXL)	265,619	268,254	2,635 (*p*-value = 6.5e-31)	0.42
Colombia (CLM)	279,828	281,911	2,163 (*p*-value = 8.3e-47)	0.18
Puerto Rico (PUR)	291,035	292,734	1,699 (*p*-value = 2.9e-67)	0.06

100 simulated individuals from each ancestral population (African, European, and Asian) as references for the local ancestry inference. Here, again, we used Asians as the closest proxy for Native Americans in the available simulation model.

We conducted imputation with the base reference panel plus a varying number of additional reference genomes (0, 100, 134, 200, 400, 600, 800, 1,000, 1,500, 2000, and 3,000). Finally, we compared imputation r^2 of using different reference panels stratified by local ancestry and allele frequency in the target individual genomes.

RESULTS

The Native American Reference Panel NATS

We built a Native American reference panel (NATS) representing indigenous populations across Latin America. The panel consists of publicly available data [HGDP (Bergström et al., 2020), SGDP (Mallick et al., 2016), and INMEGEN (Romero-Hidalgo et al., 2017)] and 50 new genomes from the MX Biobank Project (*Materials and Methods*, and **Supplementary Table S2**). While most of the genomes in the panel are from indigenous groups in Mexico (103 of 134; 76.8%) (**Figure 1A**; **Supplementary Table S1**), our panel also encompasses native groups from Colombia, Brazil, and Peru. When merging NATS with 1KGP, the total number of SNPs is 102,336,497, of which 24,518,242 (24%) are unique to our panel (**Figure 1B**). The amount of non-indigenous admixture in our panel is less than 1.5% overall (**Figure 1C**). Only some Mayan individuals from HGDP show between 0.8 and 23% of European admixture (on average 6%) (**Supplementary Table S1**). Overall, our panel has 98.5% of Native American genetic ancestry. We acknowledge that, while this panel includes as many genomes as possible from those publicly available at the time of publication, it does not fully capture the genetic variation of the vast ethnic diversity in the continent. It is intended to serve as a first approximation to evaluate the impact of ancestry representation in imputation performance.

Imputation Performance of the NATS Reference Panel

To assess the impact of our panel on imputation performance, we imputed the AMR individuals (from Colombia, Peru, Puerto Rico, and Mexico in 1KGP) at SNPs not found on the MEGA array using a leave-one-out strategy, with either 1KGP or 1KGP + NATS as reference panels (Materials and Methods). We chose the MEGA array because it was specifically designed to capture global variation better. We compared the mean number of SNPs above the standard quality threshold for human genetic studies (MAF >= 1% and INFO >= 0.3) using the two reference panels. We were able to increase the number of SNPs above the quality threshold across the four populations using our NATS panel (**Table 1**). The magnitude of the increase is correlated with the individual's proportion of native ancestry (**Supplementary Figure S1**). Furthermore, the majority of these SNPs fall into diploid European tracts of the genome (**Supplementary Figure S2**) regardless of the ancestry composition of each population, and which reference panel was used for imputation. This is because even though the 1KGP has as many African individuals as Europeans, European ancestry is more predominant in AMR individuals.

To determine imputation accuracy, we computed the correlation between the real allele dosages and the imputed dosages (*Materials and Methods*). We checked imputation accuracy in 1KGP admixed individuals trimmed down to SNP array positions stratified by diploid ancestry (**Figure 2A**). Overall, imputation accuracy is worse in AMR populations with the highest proportion of Native American ancestry (**Supplementary Figure S3**). As previously reported (Martin et al., 2017), the ancestry tracts that perform the worst are the ones that are underrepresented in the reference panel, specifically African and Native American. Next, we evaluated imputation accuracy using our panel (1KGP + NATS). We were able to increase imputation accuracy particularly in rare alleles (frequency >0.003 and <0.008) with diploid Native ancestry of the Mexican population (*p*-value < 0.05 two-tailed paired *t*-test) (**Figure 2B**) but not for the other populations (**Supplementary Figure S3**) or in common frequencies (**Supplementary Figure S4**). Interestingly, we do not see the same increase in the Peruvian population, which has the highest proportion of Native American ancestry overall. This could be explained by the fact that the majority of our reference data comes from native Mexicans (**Figure 1A**; **Supplementary Table S1**). Since rare variants tend to be more private to each population (Biddanda, Rice, and Novembre 2020), we could better impute rare alleles in admixed Mexicans. This suggests that, to see a similar improvement in accuracy in the other populations, we would need to include more native individuals from each local region.

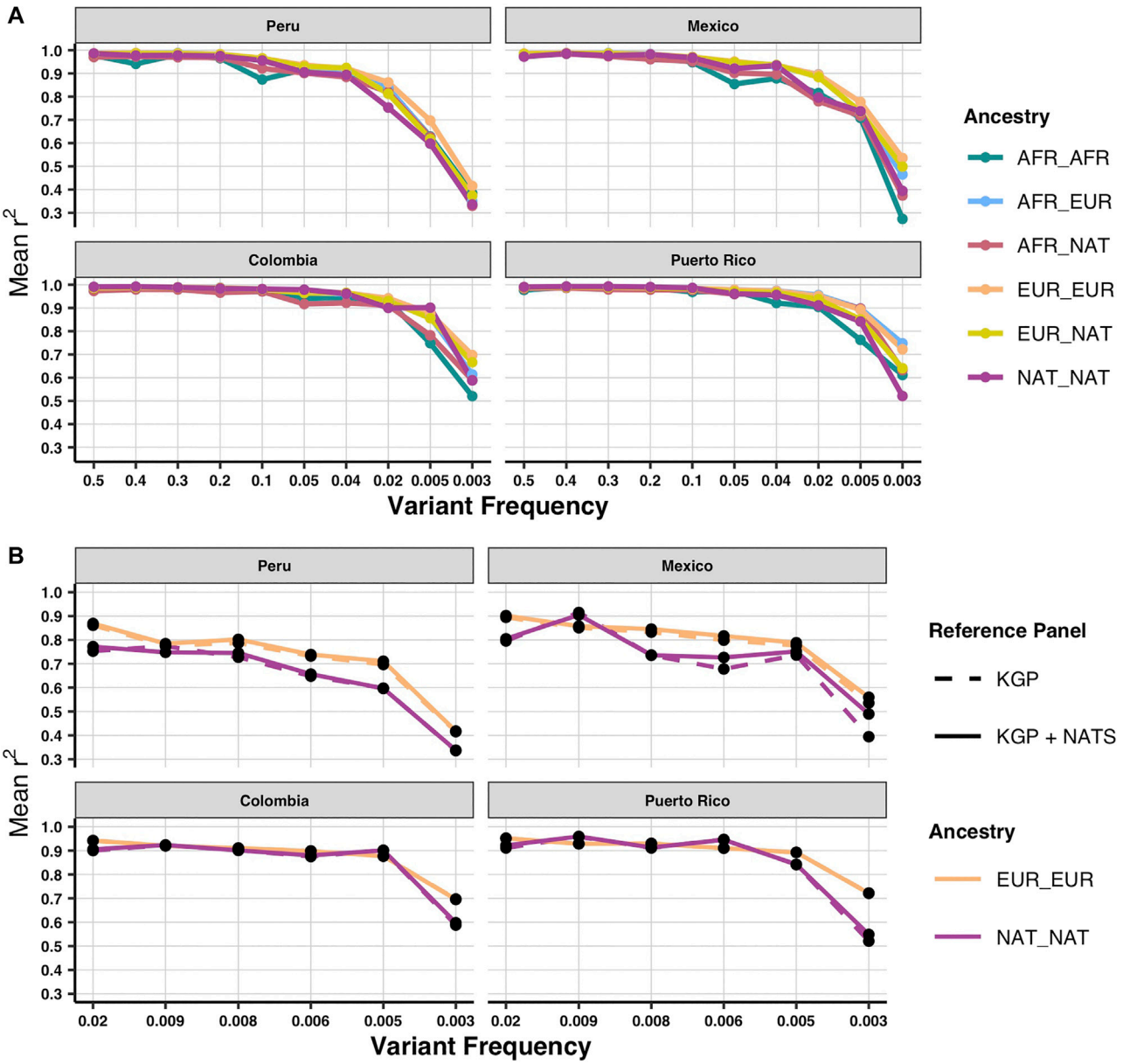

FIGURE 2 | Imputation accuracy by local ancestry and population using both reference panels. **(A)** Imputation accuracy of the four AMR populations stratified by diploid local ancestry for the MEGA array using 1KGP as reference panel. **(B)** Imputation accuracy for the Native and European diploid ancestries using 1KGP and 1KGP + NATS as reference panel focusing on rare alleles. Imputation accuracy was calculated with the Pearson squared correlation between imputed and real allele dosages.

Surprisingly, we could also see an improvement in diploid European ancestry tracts in the Mexican population (p-value < 0.05 two-tailed paired t-test for SNPs with frequency >0.003) (**Figure 2B**). One possible explanation is that because our NATS reference panel still keeps a minor fraction of European ancestry, some European haplotypes at higher frequency in Mexico could be better captured by reference genomes with such a genetic mixture. In some cases, like variants of frequency <0.02 and >0.009 with diploid Native ancestry in PEL, we could also observe a slight decrease in imputation accuracy using NATS. This could result from the uncertainty added to the data in the cross-imputation step that *IMPUTE2* performs when merging two reference panels (Howie, Marchini, and Stephens 2011).

Predicting Imputation Improvement From Additional Native American Genomes Using Simulations

Our results show that after adding 134 Native American genomes to the most widely used reference panel of global variation, we observe a promising trend of improvement. Still,

we do not come close enough to equal the imputation performance for other better represented ancestries. The question remains of how many additional genomes are still needed to close the gap. To explore this, we employed demographic simulations using *stdpopsim* (Adrion et al., 2020) and *msprime* (Kelleher, Etheridge, and McVean 2016) to generate data for a previously defined American admixture model (Browning et al., 2018). This approach allows us to explore a simulated scenario where three divergent populations intermingle to form a new admixed population (like it occurred in Latin America). With this, we can replicate the current situation where reference data are mostly available for two of the three source populations. By being able to simulate any amount of data, we can assess how many genomes of the underrepresented population (in our case, Native Americans) are necessary to equal imputation performance across ancestries. Briefly, the model simulates African, European, and Asian source populations. In the context of this analysis, the Asian population serves as a proxy for a Native American reference. We do not directly simulate a Native American population due to the lack of realistic admixture models that incorporate Native American instead of East Asian genomes as proxy in the inference of demographic parameters, which are needed to properly run the simulations. Building such demographic model is beyond the scope of this study, so given the available model and since this project focuses on Latin American populations, we will refer to the simulated Asian population as Native American. The model also simulates an admixed population that consists of 1/6 African, ⅓ European, and ½ Native American. We generated a base reference panel consisting of 661 Africans, 503 Europeans, and 347 admixed individuals (matching 1KGP sample sizes for those ancestries), as well as 3,000 Native American individuals to add sequentially to the base reference, and 300 additional admixed individuals as imputation targets (*Materials and Methods*).

We confirmed the ancestry proportions of our simulated data using *ADMIXTURE* (**Supplementary Figure S5**). To replicate the imputation pipeline, we created a genotype array dataset for the simulated target individuals by matching mean distance between markers and frequency in the European population of SNPs in the MEGA array to the simulated array, to mirror the bias in standard arrays (*Materials and Methods* and **Supplementary Figure S6**). Then, we imputed the 300 target individuals with the base reference plus either 0, 100, 134 (to mirror the sample size in NATS), 200, 400, 600, 800, 1,000, 1,500, 2000, or 3,000 Native Americans. We were able to recover roughly the same pattern of imputation accuracy (**Supplementary Figure S7**). Namely, accuracy decreased the less represented the ancestry was in the base reference with the Native American as the worst-performing ancestry. One caveat is that the best-performing ancestry is African contrary to what we see in the real data (**Figure 2A**). This is likely because the 661 African individuals are from the population that contributed to the admixed population in the simulation, which is not the case for real data. Different African ancestries contributed more or less to different Latin American populations (Micheletti et al., 2020) and not all are present in 1KGP.

When incorporating additional Native American genomes, imputation accuracy only increased in those tracts with any Native ancestry (**Supplementary Figure S8**). Furthermore, for imputation accuracy in Native American diploid ancestry tracts to equal that in European diploid ancestry tracts, 3,000 Native genomes were needed for variants with frequency>=2%, while 1,500 were enough for variants with frequency <2% (**Figure 3A**). To ask whether we reach a saturation point in the increase of imputation accuracy in the Native diploid ancestry, we compared the difference between accuracy in the base reference versus each additional reference. As expected, the behavior is different for common (frequency >0.05), low (frequency <0.05 and >0.01), and rare (frequency <0.01) variants (**Figure 3B**). Neither of them seems to show a saturation point at 3,000 newly added Native genomes. The steepest increase is achieved for the rare alleles, whereas for the common alleles, the increase is slower. This agrees with the previous result where more genomes were needed to match the Native imputation accuracy to the European one for common variants. It is also evident that the variants of common frequency are closest to saturation in accuracy as their values were already close to one (**Figure 3A**).

DISCUSSION

GWAS requires large sample sizes to detect genetic associations to complex phenotypes, and more so as the field moves toward studying rare variants (Collins 2012; Amendola et al., 2018; Abul-Husn and Kenny 2019). Therefore, SNP array platforms will continue to inform GWAS even as whole-genome sequencing costs continue to drop. In this scenario, imputation tools and genome variation resources are vital to increasing the statistical power to discover associations in understudied populations. So far, GWAS have mainly focused on populations with European ancestry (Popejoy and Fullerton 2016; Mills and Rahal 2019) and, over the past years, interesting discoveries have been made (Visscher et al., 2017). However, not all GWAS results are portable between populations (Martin et al., 2017; Duncan et al., 2019; Sirugo, Williams, and Tishkoff 2019). To ensure that these advances reach all people equitably, we must expand these studies to other populations. Other recent projects around the world have sought to reverse this trend (Gurdasani et al., 2015, 2019; GenomeAsia100K Consortium 2019; Magalhães et al., 2018; Mulder et al., 2018) improving imputation accuracy, fine mapping of associations, and discovering novel associations to well-studied phenotypes. We sought to add to this trend by creating a Native American imputation reference panel merging publicly available Native American genomes (Mallick et al., 2016; Romero-Hidalgo et al., 2017; Bergström et al., 2020) with 50 novel genomes.

One major caveat of our panel is that it does not comprehensively reflect the indigenous genetic variation across the Americas. Most of the data come from individuals from Mexico. Furthermore, the 134 genomes added are only a small increment (5%) with respect to 1KGP. The contribution of this

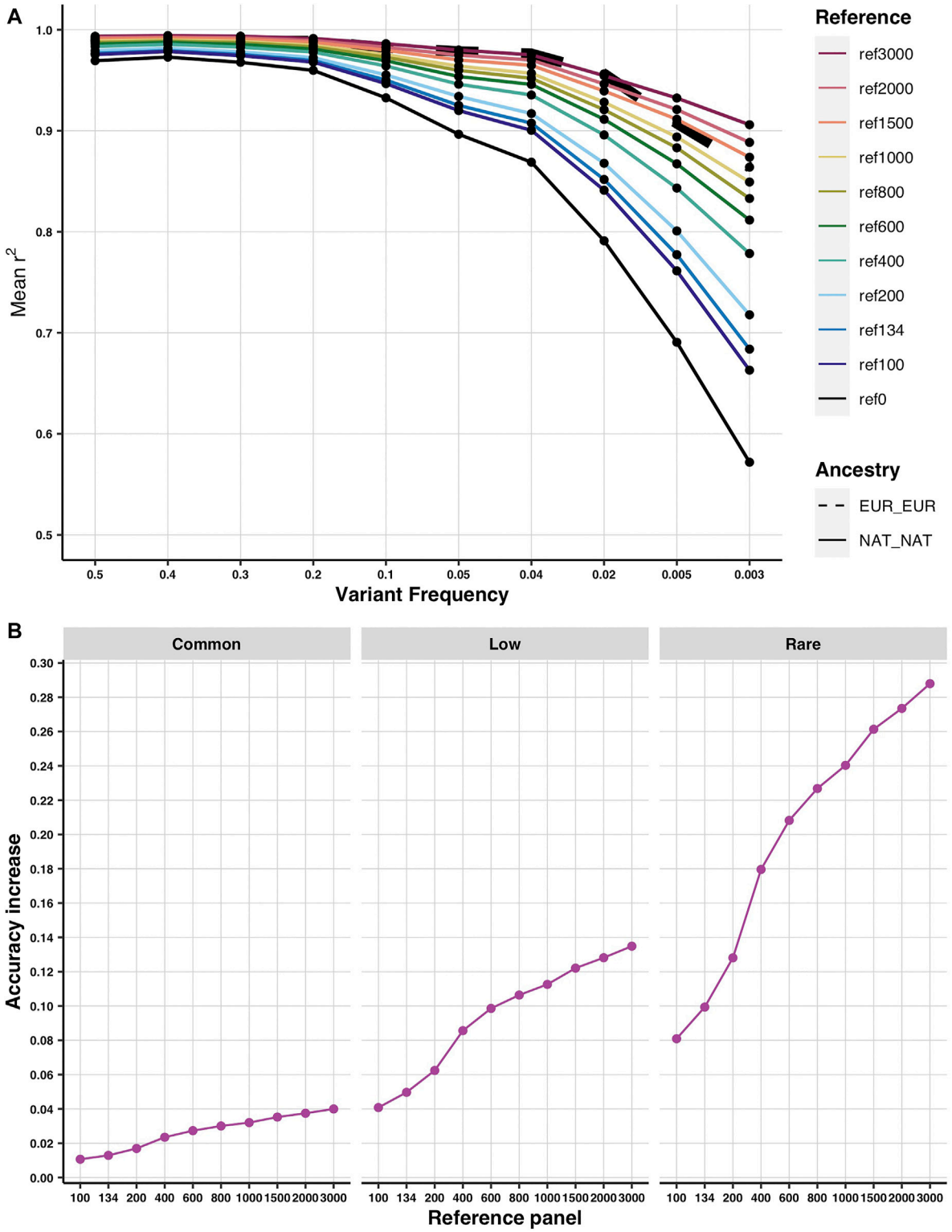

FIGURE 3 | Predicted imputation accuracy according to demographic simulations. **(A)** Imputation accuracy in the diploid Native American (solid colored lines) and diploid European (thick dashed line) ancestries using different simulated reference panels of incremental sizes. Ref 0 stands for the base reference (as it has 0 additional reference genomes). Given the available demographic model (Browning et al., 2018), a simulated Asian population was used as a proxy for Native American ancestry for the purpose of reproducing a three-way admixture process with similar ancestry proportions of African, European, and Native American sources to that observed in admixed Latino populations (see Methods for details). **(B)** Increase in imputation accuracy from the base reference in the Native American diploid ancestry at increasing sizes of the reference panel by allele frequency [common (0.5–0.05), low (0.05–0.01), and rare (0.01–0.003)].

panel is small in comparison to projects like the Uganda Genome Resource that sequenced 1,978 novel genomes (Gurdasani et al., 2019). Even with these limitations in mind, we were able to quantify the consequences of the lack of Native American genomes in commonly used imputation reference panels using empirical and simulated data analyses, while highlighting what this means for ongoing and future studies in the region.

Our panel increased the number of SNPs above the standard quality threshold for human genetic studies increasing statistical power in the four AMR populations of 1KGP. This mirrors what has been achieved by other studies in other populations (Ahmad et al., 2017; Magalhães et al., 2018; Gurdasani et al., 2019). The magnitude of this increase is positively correlated with the proportion of Native American ancestry. In other words, our panel has a stronger impact on individuals with higher Native American ancestry. However, even after using our panel, the majority of SNPs that were above the quality threshold are in chromosomal segments of the genome with European diploid ancestry, regardless of the proportion of European ancestry in the population, due to an over-representation of this ancestry in the reference panel. This means that, when doing a GWAS, the genetic signals predominantly found on the European ancestry will be easier to detect.

We were able to increase imputation accuracy in rare variants of Native American diploid ancestry in the MXL population. This was not the case for the other three populations. We expected that, since PEL is the population with the highest Native American ancestry proportion, it would also be the population most benefited by the use of our extended panel. However, there can be high levels of genetic differentiation among Native American groups, even if they are geographically close (Moreno-Estrada et al., 2014). In light of this fact, it is not a surprise that our panel, constructed with a majority of Native American individuals from Mexico, only improves accuracy in the MXL population. This suggests that to observe similar results in other populations, we should include more individuals of those populations in our panel. We also observed an increase in accuracy in some variants of European diploid ancestry. This could be attributed to the small fraction of European admixture present in the whole genomes of our extended panel, despite being enriched for Native American ancestry. Also, some of these European haplotypes could have better-captured variation found in European ancestry segments of MXL individuals. Finally, to achieve an overall increase in imputation accuracy across the whole spectrum of variant frequencies as achieved in other studies (Ahmad et al., 2017; Gurdasani et al., 2019), we would need a larger Native American reference panel, as quantified by our simulations.

These results are important with regard to not only GWAS but also their further applications. For instance, one of the applications of GWAS summary statistics is Polygenic Risk Scores (PRS). PRS calculates the genetic "risk" of an individual for a particular phenotype by summing the risk alleles present in that individual (Torkamani, Wineinger, and Topol 2018). PRS necessitates summary statistics calculated in a population as close as possible to the target individuals to be accurate. Previous studies have shown that this is not a trivial task (Tropf et al., 2017; Sirugo, Williams, and Tishkoff 2019; Mostafavi et al., 2020). Even among European populations, PRS estimates vary widely depending on the source of summary statistics due to population structure (Berg et al., 2019; Sohail et al., 2019). To have accurate PRS for the Latin American population, we need to have more studies in the region. Furthermore, our results show that we also need a better imputation panel for these populations to avoid a bias towards identifying genetic signals present on the European ancestry background.

The question of how much data are needed remained. To answer it, we employed demographic simulations. We replicated the same pattern of imputation accuracy of our data and of previous studies (Martin et al., 2017). Our strategy shows that we would need at least 3,000 Native American genomes to equal imputation accuracy of Native diploid ancestry to that of European diploid ancestry across all variant frequencies. This number holds for populations such as MXL with roughly similar ancestral proportions as the simulated admixed population. The minimum number of necessary new genomes will change depending on the proportion of native ancestry of the target population. Our study provides a framework for future projects to decide how many resources to allocate to the generation of whole-genome data. Furthermore, we have shown that rare variants are the most benefited by the addition of new data. This will prove particularly relevant as the field moves towards studying that end of the variant frequency spectrum (Cirulli et al., 2020; Minikel et al., 2020). Overall, our results show the importance of generating more diverse imputation panels to enable genetic discoveries in a broader spectrum of human diversity and to procure equity as scientific advancements in precision medicine should extend globally in benefit of all.

DATA AVAILABILITY STATEMENT

The newly generated data presented in the study are deposited in the European Genome-phenome Archive (EGA) repository, accession number EGAD00001008354 i.e. https://ega-archive.org/datasets/EGAD00001008354.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Research Ethics Committee (Approval CI-1479) of the National Institute of Public Health, Mexico. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AJ-K, AM-E, AYC, AC, SF-V, AJM, MS, AH, and DD-V designed the study. AM-E, LG-G, LF-R, LC-H, TT-L, HM-M, CA-S, and NM-R selected and provided DNA samples from the MX Biobank Project. AM-E, and AH sequenced the data. HK, NK, SS, MT,

CQ-C, and MP-M performed the whole-genome variant calling and curated the data. AJ-K, AYC, and SM-M analyzed the data. AJ-K and AM-E drafted the manuscript, with input from LG-G, CQ-C, LF-R, LC-H, TT-L, HM-M, CA-S, AH-C, MS, SM-M, NM-R, and GD-S. All authors read and approved the manuscript.

FUNDING

This work was supported by "The Mexican Biobank Project: Building Capacity for Big Data Science in Medical Genomics in Admixed Populations", a binational initiative between Mexico and the UK co-funded by CONACYT (Grant number FONCICYT/50/2016), and The Newton Fund through The Medical Research Council (Grant number MR/N028937/1) awarded to AME and AVSH. It was also supported by the International Center for Genetic Engineering and Biotechnology (ICGEB, Italy) grant number CRP/MEX20-01. MS was partially supported by the Chicago Fellows program of the University of Chicago. DODV is supported by the UC MEXUS CONACYT collaborative program (Grant number CN-19-29), and the UNAM PAPIIT funding program (Grant number IA200620).

ACKNOWLEDGMENTS

We thank the participants of the *Encuesta Nacional de Salud, 2000* (2000 National Health Survey, ENSA 2000), conducted in Mexico nationwide by the *Secretaría de Salud* (Health Secretariat) and the *Instituto Nacional de Salud Pública* (National Institute of Public Health, INSP). We are grateful to Mitzi Flores and Adriana Garmendia for project management support and to Carlos Conde, Victor Guerrero Lemus, Armando Mendez Herrera, Cruz Portugal García, Ma. Luisa Ordóñez-Sánchez, Rosario Rodriguez-Guillen, and Manuel Velazquez Mesa for biobank maintenance and sample preparation. We also thank Mary Ortega, Cecilia Gutiérrez, and Sara García for technical assistance, Jacob Cervantes for IT support, and Aaron Ragsdale for comments on earlier versions of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.719791/full#supplementary-material

REFERENCES

Abul-Husn, N. S., and Kenny, E. E. (2019). Personalized Medicine and the Power of Electronic Health Records. *Cell* 177 (1), 58–69. doi:10.1016/j.cell.2019.02.039

Adrion, J. R., Cole, C. B., Dukler, N., Galloway, J. G., Gladstein, A. L., Gower, G., et al. (2020). A Community-Maintained Standard Library of Population Genetic Models. *eLife* 9, e54967. doi:10.7554/eLife.54967

Agrawal, N., and Brown, M. A. (2014). Genetic Associations and Functional Characterization of M1 Aminopeptidases and Immune-Mediated Diseases. *Genes Immun.* 15 (8), 521–527. doi:10.1038/gene.2014.46

Aguilar-Ordoñez, I., Pérez-Villatoro, F., García-Ortiz, H., Barajas-Olmos, F., Ballesteros-Villascán, J., González-Buenfil, R., et al. (2021). Whole Genome Variation in 27 Mexican Indigenous Populations, Demographic and Biomedical Insights. *PLoS One* 16 (4), e0249773. doi:10.1371/journal.pone.0249773

Ahmad, M., Sinha, A., Ghosh, S., Kumar, V., Davila, S., Yajnik, C. S., et al. (2017). Inclusion of Population-Specific Reference Panel from India to the 1000 Genomes Phase 3 Panel Improves Imputation Accuracy. *Sci. Rep.* 7 (1), 6733. doi:10.1038/s41598-017-06905-6

Alexander, D. H., Novembre, J., and Lange., K. (2009). Fast Model-Based Estimation of Ancestry in Unrelated Individuals. *Genome Res.* 19 (9), 1655–1664. doi:10.1101/gr.094052.109

Amendola, L. M., Berg, J. S., Horowitz, C. R., Angelo, F., Bensen, J. T., Biesecker, B. B., et al. (2018). The Clinical Sequencing Evidence-Generating Research Consortium: Integrating Genomic Sequencing in Diverse and Medically Underserved Populations. *Am. J. Hum. Genet.* 103 (3), 319–327. doi:10.1016/j.ajhg.2018.08.007

Berg, J. J., Harpak, A., Sinnott-Armstrong, N., Joergensen, A. M., Mostafavi, H., Field, Y., et al. (2019). Reduced Signal for Polygenic Adaptation of Height in UK Biobank. *eLife* 8, e39725. doi:10.7554/eLife.39725

Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., et al. (2020). Insights into Human Genetic Variation and Population History from 929 Diverse Genomes. *Science* 367 (6484), eaay5012. doi:10.1126/science.aay5012

Biddanda, A., Rice, D. P., and Novembre, J. (2020). A Variant-Centric Perspective on Geographic Patterns of Human Allele Frequency Variation. *eLife* 9, e60107. doi:10.7554/eLife.60107

Browning, S. R., Browning, B. L., Daviglus, M. L., Durazo-Arvizu, R. A., Schneiderman, N., Kaplan, R. C., et al. (2018). Ancestry-Specific Recent Effective Population Size in the Americas. *PLoS Genet.* 14 (5), e1007385. doi:10.1371/journal.pgen.1007385

Chacón-Duque, J-C., Adhikari, K., Fuentes-Guajardo, M., Mendoza-Revilla, J., Acuña-Alonzo, V., Rodrigo, B., et al. (2018). Latin Americans Show Wide-Spread Converso Ancestry and Imprint of Local Native Ancestry on Physical Appearance. *Nat. Commun.* 9 (1), 5388. doi:10.1038/s41467-018-07748-z

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

Chatterjee, N., Shi, J., and García-Closas, M. (2016). Developing and Evaluating Polygenic Risk Prediction Models for Stratified Disease Prevention. *Nat. Rev. Genet.* 17 (7), 392–406. doi:10.1038/nrg.2016.27

Cirulli, E. T., White, S., Read, R. W., Elhanan, G., Metcalf, W. J., Tanudjaja, F., et al. (2020). Genome-Wide Rare Variant Analysis for Thousands of Phenotypes in over 70,000 Exomes from Two Cohorts. *Nat. Commun.* 11 (1), 542. doi:10.1038/s41467-020-14288-y

Collins, R. (2012). What Makes UK Biobank Special? *Lancet* 379 (9822), 1173–1174. doi:10.1016/s0140-6736(12)60404-8

Danecek, P., Adam, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The Variant Call Format and VCFtools. *Bioinformatics* 27 (15), 2156–2158. doi:10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve Years of SAMtools and BCFtools. *GigaScience* 10 (2). giab008. doi:10.1093/gigascience/giab008

Delaneau, O., and Marchini, J.1000 Genomes Project Consortium; 1000 Genomes Project Consortium (2014). Integrating Sequence and Array Data to Create an Improved 1000 Genomes Project Haplotype Reference Panel. *Nat. Commun.* 5, 3934. doi:10.1038/ncomms4934

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., et al. (2019). Analysis of Polygenic Risk Score Usage and Performance in Diverse Human Populations. *Nat. Commun.* 10 (1), 3328. doi:10.1038/s41467-019-11112-0

Faust, G. G., and Hall., I. M. (2014). SAMBLASTER: Fast Duplicate Marking and Structural Variant Read Extraction. *Bioinformatics* 30 (17), 2503–2505. doi:10.1093/bioinformatics/btu314

Flannick, J., Thorleifsson, G., Beer, N. L., Jacobs, S. B., Grarup, N., Burtt, N. P., et al. (2014). Loss-of-Function Mutations in SLC30A8 Protect against Type 2 Diabetes. *Nat. Genet.* 46 (4), 357–363. doi:10.1038/ng.2915

GenomeAsia 100K Consortium (2019). The GenomeAsia 100K Project Enables Genetic Discoveries across Asia. *Nature* 576 (7785), 106–111. doi:10.1038/s41586-019-1793-z

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., et al. (2015). The African Genome Variation Project Shapes Medical Genetics in Africa. *Nature* 517 (7534), 327–332. doi:10.1038/nature13997

Gurdasani, D., Carstensen, T., Fatumo, S., Chen, G., Franklin, C. S., Prado-Martinez, J., et al. (2019). Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa. *Cell* 179 (4), 984e36–1002. doi:10.1016/j.cell.2019.10.004

HarrisDaniel, N., Wei, S., Amol, C., ShettyKelly, S., et al. (2018). "Evolutionary Genomic Dynamics of Peruvians Before, During, and after the Inca Empire.," in Proceedings of the National Academy of Sciences of the United States of America 115 (28), E6526–E6535.

Howie, B., Marchini, J., and Stephens, M. (2011). Genotype Imputation with Thousands of Genomes. *G3 Genes|Genomes|Genetics* 1 (6), 457–470. doi:10.1534/g3.111.001198

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G. R. (2012). Fast and Accurate Genotype Imputation in Genome-Wide Association Studies through Pre-phasing. *Nat. Genet.* 44 (8), 955–959. doi:10.1038/ng.2354

Kehdy, F. S. G., Gouveia, M. H., Machado, M., Magalhães, W. C. S., Horimoto, A. R., Horta, B. L., et al. (2015). Origin and Dynamics of Admixture in Brazilians and its Effect on the Pattern of Deleterious Mutations. *Proc. Natl. Acad. Sci. United States Am.* 112 (28), 8696–8701. doi:10.1073/pnas.1504447112

Kelleher, J., Etheridge, A. M., and McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Comput. Biol.* 12 (5), e1004842. doi:10.1371/journal.pcbi.1004842

Li, H., and Durbin, R. (2009). Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 25 (14), 1754–1760. doi:10.1093/bioinformatics/btp324

Li, H. (2011). A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data. *Bioinformatics* 27 (21), 2987–2993. doi:10.1093/bioinformatics/btr509

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 (D1), D896–D901. doi:10.1093/nar/gkw1133

Magalhães, W. C. S., Araujo, N. M., Leal, T. P., Araujo, G. S., Viriato, P. J. S., Kehdy, F. S., et al. (2018). EPIGEN-Brazil Initiative Resources: A Latin American Imputation Panel and the Scientific Workflow. *Genome Res.* 28 (7), 1090–1095. doi:10.1101/gr.225458.117

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., et al. (2016). The Simons Genome Diversity Project: 300 Genomes from 142 Diverse Populations. *Nature* 538 (7624), 201–206. doi:10.1038/nature18964

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 (2), 278–288. doi:10.1016/j.ajhg.2013.06.020

Marchini, J., and Howie, B. (2010). Genotype Imputation for Genome-Wide Association Studies. *Nat. Rev. Genet.* 11 (7), 499–511. doi:10.1038/nrg2796

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 (7), 906–913. doi:10.1038/ng2088

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.* 107 (4), 788–789. doi:10.1016/j.ajhg.2017.03.004

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical Use of Current Polygenic Risk Scores May Exacerbate Health Disparities. *Nat. Genet.* 51 (4), 584–591. doi:10.1038/s41588-019-0379-x

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: A MapReduce Framework for Analyzing Next-Generation DNA Sequencing Data. *Genome Res.* 20 (9), 1297–1303. doi:10.1101/gr.107524.110

Micheletti, S. J., Bryc, K., Ancona Esselmann, S. G., Freyman, W. A., Moreno, M. E., Poznik, G. D., et al. (2020). Genetic Consequences of the Transatlantic Slave Trade in the Americas. *Am. J. Hum. Genet.* 107 (2), 265–277. doi:10.1016/j.ajhg.2020.06.012

Mills, M. C., and Rahal., C. (2019). A Scientometric Review of Genome-Wide Association Studies. *Commun. Biol.* 2, 9. doi:10.1038/s42003-018-0261-x

Minikel, E. V., Karczewski, K. J., Martin, H. C., Cummings, B. B., Whiffin, N., Rhodes, D., et al. (2020). Evaluating Drug Targets through Human Loss-Of-Function Genetic Variation. *Nature* 581 (7809), 459–464. doi:10.1038/s41586-020-2267-z

Moreno-Estrada, A., Gignoux, C. R., Fernández-López, J. C., Zakharia, F., Martin, S., Contreras, A. V., et al. (2014). The Genetics of Mexico Recapitulates Native American Substructure and Affects Biomedical Traits. *Science* 344 (6189), 1280–1285. doi:10.1126/science.1251688

Mostafavi, H., Harpak, A., Agarwal, I., Conley, D., Pritchard, J. K., and Przeworski, M. (2020). Variable Prediction Accuracy of Polygenic Scores within an Ancestry Group. *eLife* 9, e48376. doi:10.7554/eLife.48376

Mulder, N., Adebamowo, S. N., de Vries, J., Matimba, A., Olowoyo, P., Ramsay, M., et al. (2018). H3Africa: Current Perspectives. *Pharmacogenomics Pers. Med.* 11, 59–66. doi:10.2147/pgpm.s141546

Nadkarni, G. N., Gignoux, C. R., Sorokin, E. P., Rahman, R., Barnes, K. C., and Wassel, C. L. (2018). Worldwide Frequencies of APOL1 Renal Risk Variants. *New Engl. J. Med.* 379 (26), 2571–2572. doi:10.1056/nejmc1800748

Nelson, M. R., Tipney, H., Painter, J. L., Shen, J., Nicoletti, P., Shen, Y., et al. (2015). The Support of Human Genetic Evidence for Approved Drug Indications. *Nat. Genet.* 47 (8), 856–860. doi:10.1038/ng.3314

Popejoy, A. B., and Fullerton, S. M. (2016). Genomics Is Failing on Diversity. *Nature* 538 (7624), 161–164. doi:10.1038/538161a

Romero-Hidalgo, S., Ochoa-Leyva, A., Garcíarrubio, A., Acuña-Alonzo, V., Antúnez-Argüelles, E., Balcazar-Quintero, M., et al. (2017). Demographic History and Biologically Relevant Genetic Variation of Native Mexicans Inferred from Whole-Genome Sequencing. *Nat. Commun.* 8 (1), 1005. doi:10.1038/s41467-017-01194-z

SIGMA Type 2 Diabetes ConsortiumWilliams, A. L., Jacobs, S. B. R., Moreno-Macías, H., Huerta-Chagoya, A., Churchhouse, C., Márquez-Luna, C., et al. (2014). Sequence Variants in SLC16A11 Are a Common Risk Factor for Type 2 Diabetes in Mexico. *Nature* 506 (7486), 97–101. doi:10.1038/nature12828

Sirugo, G., Williams, S. M., and Tishkoff, S. A. (2019). The Missing Diversity in Human Genetic Studies. *Cell* 177 (4), 1080. doi:10.1016/j.cell.2019.04.032

Soares-Souza, G., Borda, V., Kehdy, F., and Tarazona-Santos, E. (2018). Admixture, Genetics and Complex Diseases in Latin Americans and US Hispanics. *Curr. Genet. Med. Rep.* 6 (4), 208–223. doi:10.1007/s40142-018-0151-z

Sohail, M., Maier, R. M., Ganna, A., Bloemendal, A., Martin, A. R., Turchin, M. C., et al. (2019). Polygenic Adaptation on Height Is Overestimated Due to Uncorrected Stratification in Genome-Wide Association Studies. *eLife* 8, e39702. doi:10.7554/eLife.39702

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., and Prins, P. (2015). Sambamba: Fast Processing of NGS Alignment Formats. *Bioinformatics* 31 (12), 2032–2034. doi:10.1093/bioinformatics/btv098

The 1000 Genomes Project ConsortiumAuton, 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 (7571), 68–74. doi:10.1038/nature15393

Torkamani, A., Wineinger, N. E., and Topol, E. J. (2018). The Personal and Clinical Utility of Polygenic Risk Scores. *Nat. Rev. Genet.* 19 (9), 581–590. doi:10.1038/s41576-018-0018-x

Tropf, F. C., Lee, S. H., Verweij, R. M., Stulp, G., van der Most, P. J., de Vlaming, R., et al. (2017). Hidden Heritability Due to Heterogeneity across Seven Populations. *Nat. Hum. Behav.* 1 (10), 757–765. doi:10.1038/s41562-017-0195-1

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 (1), 5–22. doi:10.1016/j.ajhg.2017.06.005

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., et al. (2019). Genetic Analyses of Diverse Populations Improves Discovery for Complex Traits. *Nature* 570 (7762), 514–518. doi:10.1038/s41586-019-1310-4

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

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: frontiersin.org/about/contact

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