Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations

Alicia R. Martin,1,2,5,4 Christopher R. Gignoux,4 Raymond K. Walters,1,2,3 Genevieve L. Wojcik,4 Benjamin M. Neale,1,2,3 Simon Gravel,5,6 Mark J. Daly,1,2,3 Carlos D. Bustamante,4 and Eimear E. Kenny7,8,9,10,*

The vast majority of genome-wide association studies (GWASs) are performed in Europeans, and their transferability to other populations is dependent on many factors (e.g., linkage disequilibrium, allele frequencies, genetic architecture). As medical genomics studies become increasingly large and diverse, gaining insights into population history and consequently the transferability of disease risk measurement is critical. Here, we disentangle recent population history in the widely used 1000 Genomes Project reference panel, with an emphasis on populations underrepresented in medical studies. To examine the transferability of single-ancestry GWASs, we used published summary statistics to calculate polygenic risk scores for eight well-studied phenotypes. We identify directional inconsistencies in all scores; for example, height is predicted to decrease with genetic distance from Europeans, despite robust anthropological evidence that West Africans are as tall as Europeans on average. To gain deeper quantitative insights into GWAS transferability, we developed a complex trait coalescent-based simulation framework considering effects of polygenicity, causal allele frequency divergence, and heritability. As expected, correlations between true and inferred risk are typically highest in the population from which summary statistics were derived. We demonstrate that scores inferred from European GWASs are biased by genetic drift in other populations even when choosing the same causal variants and that biases in any direction are possible and unpredictable. This work cautions that summarizing findings from large-scale GWASs may have limited portability to other populations using standard approaches and highlights the need for generalized risk prediction methods and the inclusion of more diverse individuals in medical genomics.

Introduction

The majority of genome-wide association studies (GWASs) have been performed in populations of European descent.1–4 An open question in medical genomics is the degree to which these results transfer to new populations. GWASs have yielded tens of thousands of common genetic variants significantly associated with human medical and evolutionary phenotypes, most of which have replicated in other ethnic groups.5–7 However, GWASs are optimally powered to discover common variant associations, and the European bias in GWAS results in associated SNPs with higher minor allele frequencies on average compared to other populations. The predictive power of GWAS findings and genetic diagnostic accuracy in non-Europeans are therefore limited by population differences in allele frequencies and linkage disequilibrium structure. For example, a previous study showed that the accuracy of breeding values and genomic prediction decays approximately linearly with increasing divergence between the discovery and target population.8 Additionally, multiple individuals with African ancestry have received false positive misdiagnoses of hypertrophic cardiomyopathy that would have been prevented with the inclusion of even small numbers of African Americans in these studies.9 Further, a previous study finding that 96% of GWAS participants are of European descent7 has recently been updated; although the non-European proportion of GWAS participants has increased to nearly 20%, this is primarily driven by Asian individuals, and the proportion of individuals with African and Hispanic/Latino ancestry in GWASs has remained essentially unchanged.4

As GWAS sample sizes grow to hundreds of thousands of samples, they also become better powered to detect rare variant associations.10–12 Large-scale sequencing studies have demonstrated that rare variants show stronger geographic clustering than common variants.13–15 Rare, disease-associated variants are therefore expected to track with recent population demography and/or be population restricted.14,16–18 As the next era of GWASs expands to evaluate the disease-associated role of rare variants, it is not only scientifically imperative to include multi-ethnic populations, it is also likely that such studies will encounter increasing genetic heterogeneity in very large study populations. A comprehensive understanding of the genetic diversity and demographic history of multi-ethnic populations is critical for appropriate applications of GWASs and ultimately for ensuring that genetics does not contribute to or enhance health disparities.4

1Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA; 2Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; 3Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; 4Department of Genetics, Stanford University, Stanford, CA 94305, USA; 5Department of Human Genetics, McGill University, Montreal, QC H3A 0G1, Canada; 6McGill University and Genome Quebec Innovation Centre, Montreal, QC H3A 0G1, Canada; 7Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; 8The Charles Bronfman Institute of Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; 9Center of Statistical Genetics, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; 10Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
*Correspondence: eimear.kenny@mssm.edu
http://dx.doi.org/10.1016/j.ajhg.2017.03.004.
© 2017 American Society of Human Genetics.

The American Journal of Human Genetics 100, 1–15, April 6, 2017
The most recent release of the 1000 Genomes Project (phase 3) provides one of the largest global reference panels of whole-genome sequencing data, enabling a broad survey of human genetic variation. 19 The depth and breadth of diversity queried facilitates a deep understanding of the evolutionary forces (e.g., selection and drift) shaping existing genetic variation in present-day populations that contribute to adaptation and disease. 20–25 Studies of admixed populations have been particularly fruitful in identifying genetic adaptations and risk for diseases that are stratified across diverged ancestral origins. 26–31 Admixture patterns became especially complex during the peopling of the Americas, with extensive recent admixture spanning multiple continents. Processes shaping structure in these admixed populations include sex-biased migration and admixture, isolation-by-distance, differential drift in mainland versus island populations, and variable admixture timing. 14,32,33

Standard GWAS strategies approach population structure as a nuisance factor. A typical stepwise procedure first detects dimensions of global population structure in each individual, using principal-component analysis (PCA) or other methods, 34–37 and often excludes “outlier” individuals from the analysis and/or corrects for inflation arising from population structure in the statistical model for association. Such strategies reduce false positives in test statistics, but can also reduce power for association in heterogeneous populations and are less likely to work for rare variant association. 38,39 Recent methodological advances have leveraged patterns of global and local ancestry for improved association power, 27,40,41 fine-mapping 42 and genome assembly. 43 At the same time, population genetic studies have demonstrated the presence of fine-scale sub-continental structure in the African, Native American, and European components of populations from the Americas. 44–47 If trait-associated variants follow the same patterns of demography, then we expect that modeling sub-continental ancestry may enable their improved detection in admixed populations.

The dawn of the GWAS era saw limited success in identifying genome-wide significant loci associated with disease, and a major endeavor to better understand the genetic architecture of complex traits emerged. The peaks that met genome-wide significance typically did not explain a significant fraction of the phenotypic variance, and a major goal to estimate how many more signals remained yet to be discovered arose; this objective ushered in a wave of methodological development in heritability, linear mixed models, and polygenic risk prediction, as discussed and reviewed extensively elsewhere. 11,48–56 Numerous complex traits have been studied with cohort sizes in the hundreds of thousands, and yet in each case there are many more signals that improve prediction accuracy than meet genome-wide significance. 48,57–59 For example, including only genome-wide significant loci in the prediction of schizophrenia explains <3% of the phenotypic variance, whereas loci meeting the significance threshold that optimally balances signal versus noise (in this case, \( p \leq 0.1 \)) in the meta-analysis explains considerably more (>18%) of the phenotypic variance. 11 Because the prediction accuracy, which is usually measured via prediction \( R^2 \), Nagelkerke’s \( R^2 \), or receiver operator curve AUC, of polygenic risk scores is currently low for most traits, 56 genetic risk prediction is not clinically viable at present, but polygenic risk scores have nonetheless repeatedly proven valuable in research contexts across a multitude of complex traits 11,48,60–65 and will become increasingly useful as GWAS sample sizes grow. 59 Additionally, several methodological advancements to the standard approach have recently been undertaken. 58,66–68

In this study, we explore the impact of population diversity on the landscape of variation underlying human traits. We infer demographic history for the global populations in the 1000 Genomes Project, focusing particularly on admixed populations from the Americas, which are under-represented in medical genetic studies. 4 We disentangle local ancestry to infer the ancestral origins of these populations. We link this work to ongoing efforts to improve study design and disease variant discovery by quantifying biases in clinical databases and GWASs in diverse and admixed populations. These biases have a striking impact on genetic risk prediction; for example, a previous study calculated polygenic risk scores for schizophrenia in East Asians and Africans based on GWAS summary statistics derived from a European cohort and found that prediction accuracy was reduced by more than 50% in non-European populations. 52 To disentangle the role of demography on polygenic risk prediction derived from single-ancestry GWASs, we designed a coalescent-based simulation framework reflecting modern human population history and show that polygenic risk scores derived from European GWASs are biased when applied to diverged populations. Specifically, we identify reduced variance in risk prediction with increasing divergence from Europe reflecting decreased overall variance explained, and demonstrate that an enrichment of low-frequency risk and high-frequency protective alleles contribute to an overall protective shift in European inferred risk on average across traits. Our results highlight the need for the inclusion of more diverse populations in GWASs as well as genetic risk prediction methods improving transferability across populations.

Material and Methods

Ancestry Deconvolution

We used the phased haplotypes from the 1000 Genomes consortium. We phased reference haplotypes from 43 Native American samples from Mao et al. 69 inferred to have \( >0.99 \) Native ancestry in ADMIXTURE using SHAPEIT2 (v.2.r778), 70 then merged the haplotypes using scripts made publicly available. These combined phased haplotypes were used as input to the PopPhased version of RFMix v.1.5.471 with the following flags: -w 0.2, -e 1, -n 5, --use-reference-panels-in-EM, --forward-backward EM. The node size of 5 was selected to reduce bias in
دریافت فوری متن کامل مقاله

امکان دانلود نسخه تمام متن مقالات انگلیسی
امکان دانلود نسخه ترجمه شده مقالات
پذیرش سفارش ترجمه تخصصی
امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
امکان دانلود رایگان ۲ صفحه اول هر مقاله
امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
دانلود فوری مقاله پس از پرداخت آنلاین
پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات