Analysis of Polygenic Score Usage and Performance in Diverse Human Populations

Duncan, LE[a,1]
Shen, H[a]
Gelaye, B[b]
Ressler, KJ[c]
Feldman, MW[d]
Peterson, RE[e]
Domingue, BW[f]

[a] Stanford University, Department of Psychiatry and Behavioral Sciences, 401 Quarry Road, Stanford, CA 94305
[b] Harvard T.H. Chan School of Public Health, Department of Epidemiology, 667 Huntington Ave, Kresge 505, Boston, MA 02115
[c] McLean Hospital, Harvard Medical School, Mailman Research Center, 115 Mill St. Belmont, MA 02478
[d] Stanford University, Department of Biology, Herrin 478A, Stanford, CA 94305
[e] Virginia Institute for Psychiatric Genetics, Department of Psychiatry, Richmond, VA, P.O. Box 980003
[f] Stanford University, Department of Education, CERAS 510, Stanford CA, 94305

[1] Corresponding Author: Laramie Duncan
LaramieD@Stanford.edu
(650) 723-3258
401 Quarry Road
Room 3320
Stanford, CA 94305

Keywords: Polygenic score, human genetics, genome-wide association study (GWAS), diverse populations, complex genetic phenotypes, height, schizophrenia, BMI.

Abstract

Studies examining relationships between genotypic and phenotypic variation have historically been carried out on people of European ancestry. Efforts are underway to address this limitation, but until they succeed, the legacy of a Euro-centric bias will continue to hinder research, including the use of polygenic scores, which are individual-level metrics of genetic risk. Ongoing debate surrounds the generalizability of polygenic scores based on genome-wide association studies (GWAS) conducted in European ancestry samples, to non-European ancestry samples. We analyzed the first decade of polygenic scoring studies (2008-2017, inclusive), and found that 67% of studies included exclusively European ancestry participants and another 19% included only East Asian ancestry participants. Only 3.8% of studies were carried out on samples of African, Hispanic, or Indigenous peoples. We find that effect sizes for European ancestry-derived polygenic scores are only 36% as large in African ancestry samples, as in European ancestry samples (t=-10.056, df=22, p=5.5x10^{-10}). Analyzing global populations, we show that relationships between height polygenic scores and height are highly dependent on methodological choices in polygenic score construction, highlighting the need for caution in interpreting population level differences in distributions of polygenic scores, as currently calculated. These findings bolster the rationale for large-scale GWAS in diverse human populations and highlight the need for better handling of linkage disequilibrium and variant frequencies when applying scores to non-European samples.

Introduction

Awareness of the over-representation of participants of European ancestry in human genetics research has been broadly acknowledged[1–5], and increasing the representation of diverse populations has recently become a higher priority for the research community[5–10]. This has led funding agencies such as the National Institutes of Mental Health to make genetic studies of diverse populations a priority. Accordingly, representation of non-European ancestry participants in genome-wide association studies (GWAS) increased, from 4% in 2009[1] to 19% in 2016[3]. Most of the increase in non-European ancestry research is attributable to expansion of genetic studies of East Asian populations, as reported previously[3] and as observed in our data (see below). As such, most populations are still severely underrepresented. This lack of representation, if not mitigated, will limit our understanding of etiological factors predisposing to disease risk, and will hinder efforts to develop precision medicine. It is also important to understand the implications of the European-centric bias of earlier genetic studies, for work that builds upon existing research. For example, researchers want to know how the limited diversity in early medical genetic studies impacts the use of polygenic risk scores in non-European ancestry populations.

The use of polygenic risk scores[11,12] (PRS, also known as risk profile scoring, genetic scoring, and genetic risk scoring) has become widespread in biomedical and social science disciplines[13–15]. Businesses have commercialized this technology, including direct-to-consumer testing from 23&Me and other companies. Perhaps most importantly, there is hope that polygenic risk scores can improve health outcomes by accelerating diagnosis and matching patients to tailored treatments[16]. Polygenic scoring studies have demonstrated reliable, though modest, prediction using straightforward scoring methods[11,12] and genetic data alone, for many complex genetic phenotypes (e.g. blood pressure[13,17], height[18], diabetes[9,19], depression[7,20], and schizophrenia[14]). Polygenic risk scores are calculated by summing risk alleles, which are weighted by effect sizes derived from GWAS results[11,12,21]. Commonly used methods account for ancestry using principal components (calculated on pruned genetic data). In the parlance of polygenic scoring studies, the *training* GWAS is referred to as the "discovery" sample, and the *testing* dataset is referred to as the "target" sample. No overlap between discovery and target datasets is imperative, as is the removal of related individuals from analyses, as demonstrated by Wray and colleagues[21]. Methods of prediction that offer modest improvements on this basic framework are also available[22–25].

Polygenic scores can be constructed for any complex genetic phenotype for which appropriate GWAS (or other robust association) results are available. The challenges inherent in using polygenic scores – including modest predictive ability and considerations of statistical power in the interpretation of results – have been reviewed previously[21,26]. Recent research has focused on the generalizability of polygenic scores to non-European ancestry populations[27]. There is good reason to anticipate reduced predictive power in non-European samples because of differences in variant frequencies and linkage disequilibrium patterns between populations[12,28]. However, the expected magnitude of decreased performance is largely unknown. Few systematic studies of polygenic score performance across different ancestry groups are available, though see Hoffman and colleagues[13] for a thorough investigation of blood pressure metrics. To date, all available

information has been based on individual phenotypes or small numbers of observations[28,32]. Further, previous findings may need to be re-evaluated in light of newer findings about relationships between ancestry and GWAS results[29–31]. Thus, there is a need for systematic evaluation of polygenic score performance across multiple populations and phenotypes.

In addition to questions about the predictive performance of polygenic scores, a second major area of inquiry concerns differences in the distributions of polygenic scores for global populations, as currently calculated[27,30,31,33–39]. Multiple potential causes of observed distribution differences of polygenic scores have been reported, including drift[27], selection[33,36–39], artifactual differences due to uncorrected population stratification[30,31] and different environmental effects[40,41]. We investigate relationships between global principal components and polygenic scores, and assess relationships between polygenic scores and phenotypes for height, using 1000Genomes[42] data and country-level height information. These analyses will be helpful in calibrating researcher expectations about polygenic scores, as currently calculated, for diverse non-European ancestry populations.

Materials and Methods

This study has two major parts. First, we analyzed data extracted from previous polygenic scoring studies in order to describe trends in polygenic scoring research and to provide the most comprehensive analysis of polygenic scoring performance available to date. Second, we analyzed properties of polygenic scores as calculated for the 1000Genomes individuals, and compared polygenic scores to country-level information about height. Relevant procedures for these two parts are described below. This work received a notice of determination that this was not human subjects research from Stanford University.

Part 1: Extracting and analyzing data from previous polygenic scoring studies. We first identified studies with data suitable for extraction, via PubMed on January 23rd, 2018 using the following search terms: (Genome-Wide Association Stud* OR GWAS OR Genome Wide Association Stud*) and (polygenic risk score OR genetic risk score OR polygenic risk scor* OR genetic risk scor* OR risk profile scor* OR "genomic profile"). We sought to identify all polygenic scoring studies, of any complex genetic phenotype, from the first decade of polygenic scoring research. This yielded 1,226 studies, 733 of which were polygenic scoring studies (see **Figure 1**). From these 733 studies we extracted data about the ancestry of participants and methods of constructing polygenic scores.

We next identified studies that contained valid comparisons of the performance of polygenic scores in European ancestry participants and at least one other ancestry. Specifically, matched analyses (from two or more ancestry groups, from any given publication) had to use the same genotyping chip for all samples, the same weights for variants, the same algorithm for constructing polygenic scores, and the same methods of measuring phenotypes across all participants. Data from 29 studies met inclusion criteria. From these studies we then extracted effect size metrics for each ancestry group and then normalized score performance for all ancestry groups to performance within the European ancestry participants, by dividing all effect

sizes (within each study) by the effect size of the relevant European ancestry sample, within each study. We multiplied values by 100 so that performance for each non-European ancestry sample could be expressed as a percentage of European ancestry performance, which was standardized to 100%. For example, the first polygenic scoring study of schizophrenia[12] found that polygenic scores explained only 0.4% of phenotypic variance in an African ancestry sample, whereas 3.2% of phenotypic variance was explained in a matched European ancestry sample. Consequently, the value of 12.5% (100*(0.004/.032)) is represented as one yellow point in **Figure 2**, among the African ancestry observations displayed on the right side of the **Figure 2**. In **Figure 2**, each point represents one within-study comparison between a non-European ancestry sample and the matched (within-study) European ancestry sample. European ancestry performance is standardized to 100% in all comparisons and is represented by the blue circle on the left side of the figure. By normalizing within-study effect sizes to European ancestry effect sizes, we were able to combine observations across phenotypes, and therefore to obtain general estimates of polygenic score performance across ancestry groups and complex genetic phenotypes. We only analyzed results from studies employing the "classical" polygenic scoring approach, which includes "pruning and thresholding", and allele weights derived from an independent discovery GWAS[11,12,21,26]. For construction of the boxplots using R, we used the default settings except for specifying that the whisker ends should reflect minimum and maximum values ('range=0').

Part 2: Polygenic score properties for 1000Genomes individuals. For part 2, we used publicly available data from 1000Genomes[42]. Individual-level genotype data for 2,557 individuals was downloaded from ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/. Weights for constructing polygenic scores came from publicly-available sources of GWAS results for height, body mass index (BMI), and schizophrenia[14,18,46,48]. Data about average human height, for countries of origin for 1000Genomes populations, was downloaded from a pre-compiled table with male and female heights by country: https://en.wikipedia.org/wiki/List_of_average_human_height_worldwide.

Data preparation and analysis for 1000Genomes samples: The full 1000Genomes dataset was first filtered to include only bi-allelic single nucleotide polymorphisms (SNPs) with greater than 0.1% minor allele frequency. In order to calculate principal components across 1000Genomes genotypes, we used second generation PLINK[53] to obtain variants in approximate linkage equilibrium, and we also removed the MHC region of chromosome 6 (25-35Mb) and the large inversion region on chromosome 8 (7-13Mb). We then calculated 20 PCs across all individuals.

In order to obtain weights for constructing polygenic scores, summary statistics files (i.e. GWAS results) were pruned to include only variants in approximate linkage equilibrium, using second generation PLINK[53,54] and the following thresholds: --clump-kb 500, --clump-p1 1, --clump-p2 1, --clump-r2 0.2. 1000Genomes European ancestry data was used as the source of linkage disequilibrium information for pruning all summary statistic files given the primarily European ancestry of discovery GWAS datasets.

Second generation PLINK[53] was used to construct polygenic scores for each phenotype for the 1000Genomes participants, using 13 thresholds for including pruned discovery GWAS variants in scores, as follows: $p<5x10^{-8}$, $p<1x10^{-6}$, $p<1x10^{-4}$, $p<1x10^{-3}$, $p<1x10^{-2}$, $p<.05$, $p<.1$, $p<.2$, $p<.3$, $p<.4$, $p<.5$, $p<.75$, $p</=1$. The statistical package R[55] was used for plotting and statistical

analyses (t-tests, correlations). Height phenotype data was downloaded from a compiled table of average heights, for males and females, by countries. Heights for males and females were averaged. Certain populations were excluded from the analysis of correlations between polygenic risk scores for height and height phenotypes for three reasons: Four populations were excluded due to lack of height phenotype data: Puerto Rican in Puerto Rico (PUR), Bengali in Bangladesh (BEB), Punjabi in Lahore, Pakistan (PJL), Mende in Sierra Leone (MSL). Two populations were excluded due to the combination of highly mixed country ancestry (impacting validity of height phenotype) and admixture of the 1000Genomes population (impacting variability in the polygenic scores for height): African Ancestry in Southwest US (ASW), African Caribbean in Barbados (ACB). One population was excluded due to the absence of a single European country of origin (needed for height phenotype information used in this report): Utah residents with Northern and Western European ancestry (CEU). Details are given in **Supplementary_Table_3_1000Genomes_countries of origin.xlsx**

Results

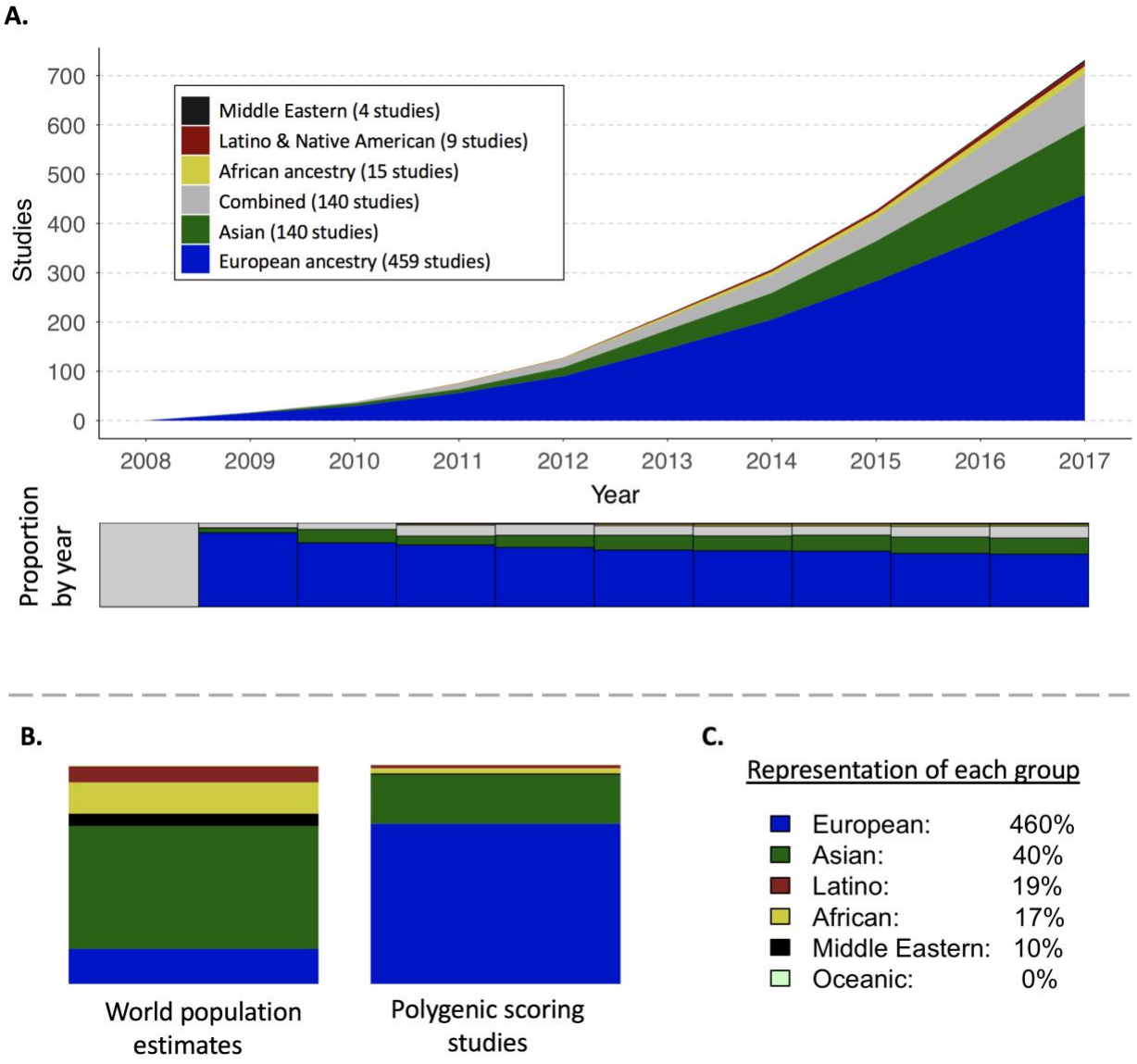
Usage and performance of polygenic scores in diverse human populations

How well different ancestry groups have been represented in the first decade of polygenic scoring research (2008-2017, inclusive) is shown in **Figure 1A**, which presents cumulative distributions of studies for specific ancestry groups across time. The field has been dominated by European ancestry studies. Across the 733 studies examined (see **Methods** for inclusion criteria and **Supplementary Table 1** for a list of studies), 67% included exclusively European ancestry participants. There have also been 140 studies conducted in exclusively Asian populations (19%), most commonly in East Asian countries (e.g., China and Japan). Only 3.8% of the polygenic studies from the first decade of polygenic scoring research concerned populations of African, Latino/Hispanic, or Indigenous peoples combined[*]. These results are similar to those reported by Popejoy and Fullerton[3], who noted that non-European ancestry representation in GWASs was almost exclusively in Asian populations, and East Asian populations in particular.

By comparing representation of particular ancestry groups with world population estimates for those groups (**Figure 1B**), it is possible to quantify the over- or under-representation of each major ancestry group. European ancestry representation was approximately 460% of what it would be if representation was proportional to world ancestry. In contrast, African ancestry (17%) and Latino samples (19%) were under-represented relative to world populations. East and South Asian samples are combined in this figure, but it should be noted that representation of East Asian samples is much higher than South Asian samples, which have been included in very few polygenic scoring studies to date. Middle Eastern and Oceanic populations have the lowest representation in polygenic scoring studies relative to world populations for these groups (10% and 0%, respectively).

[*]Note that we retain populations names from the original reports (e.g. "Native American" and "Middle Eastern") for Figure 1 in order to maintain consistency in terminology. "Combined" means that more than one ancestry group was included in the study (e.g. European ancestry and Asian ancestry participants).

Figure 1. The first decade of polygenic scoring studies (2008-2017) focused primarily on European ancestry samples (N=733 studies). A. Cumulative numbers of studies by year are denoted by color. The stacked bar graph below the cumulative distribution plot shows proportional ancestry by year. **B.** Pie charts depict world ancestry representation (left) and polygenic scoring study representation (right). **C.** The percentage representation for each ancestry group is given, such that 100% would indicate equal representation in the world and in polygenic scoring studies. For example, European ancestry samples are over-represented (460%) whereas African ancestry samples are under-represented (17%).

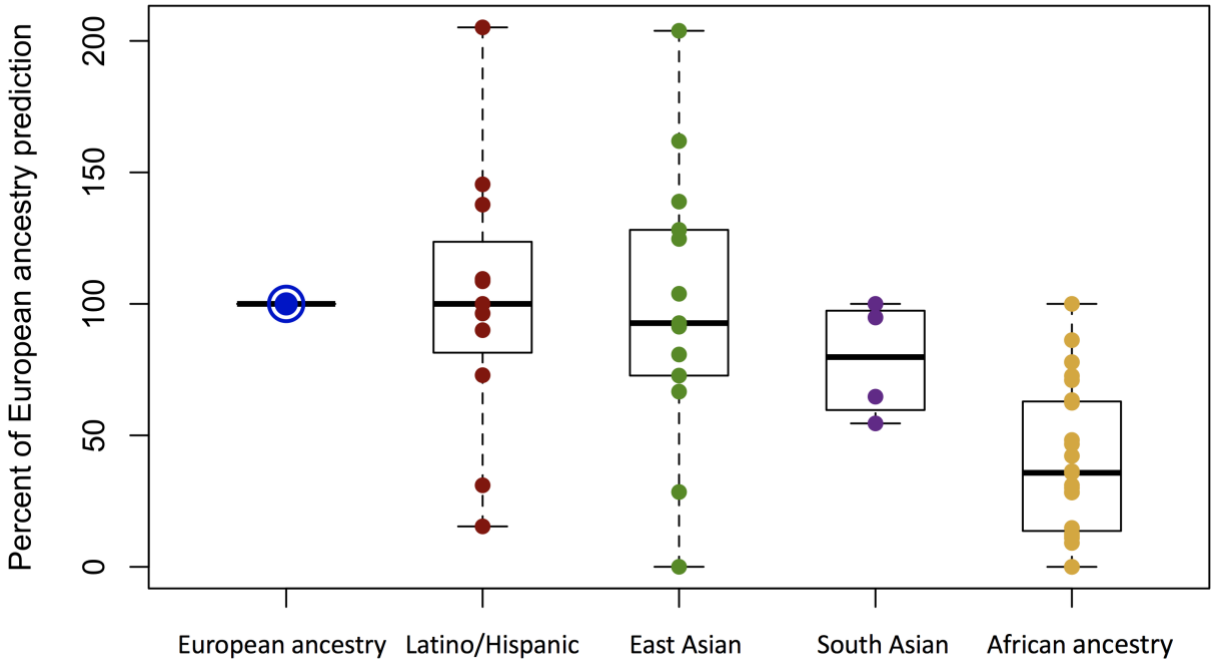


Having analyzed the *use* of polygenic scores in different ancestry groups (above), we next assessed the *performance* of polygenic scores in multiple ancestry groups. Since most large-scale GWAS have been conducted in primarily (or exclusively) European ancestry individuals, our *a priori* hypothesis was that polygenic scores would perform best among European ancestry individuals, and less well for other populations. **Figure 2** provides an overview of polygenic

score performance across ancestry groups. Results from all complex genetic phenotypes are analyzed together in order to increase the amount of data available for analysis.

Polygenic score performance was worst among African ancestry samples. The median effect size of polygenic scores in African ancestry samples was only 36% that of matched European ancestry samples (t=-10.056, df=22, p=5.5x10^{-10}). Relative to matched European ancestry samples, performance was also lower in South (80%) and East Asian (93%) samples, but not significantly so. In sum, an expectation of poorer polygenic score performance in non-European ancestry populations seems reasonable given these data. Attenuation of predictive performances is likely to be most extreme in samples of African ancestry, consistent with, on average, greater genetic distance between European and African ancestry populations, than between European and other ancestry populations[28,45].

Figure 2 Performance of polygenic scores in Latino, East Asian, South Asian, and African ancestry samples, relative to performance in matched European ancestry samples (29 total studies). In order to make data comparable across studies, performance in each study was standardized to European ancestry performance, hence the single European ancestry y-axis value of 100%. Each point represents one pair of polygenic scoring analyses between a European ancestry sample and a matched sample from another ancestry (see text for details). Boxplots denote minimum, first quartile, median, third quartile, and maximum values.

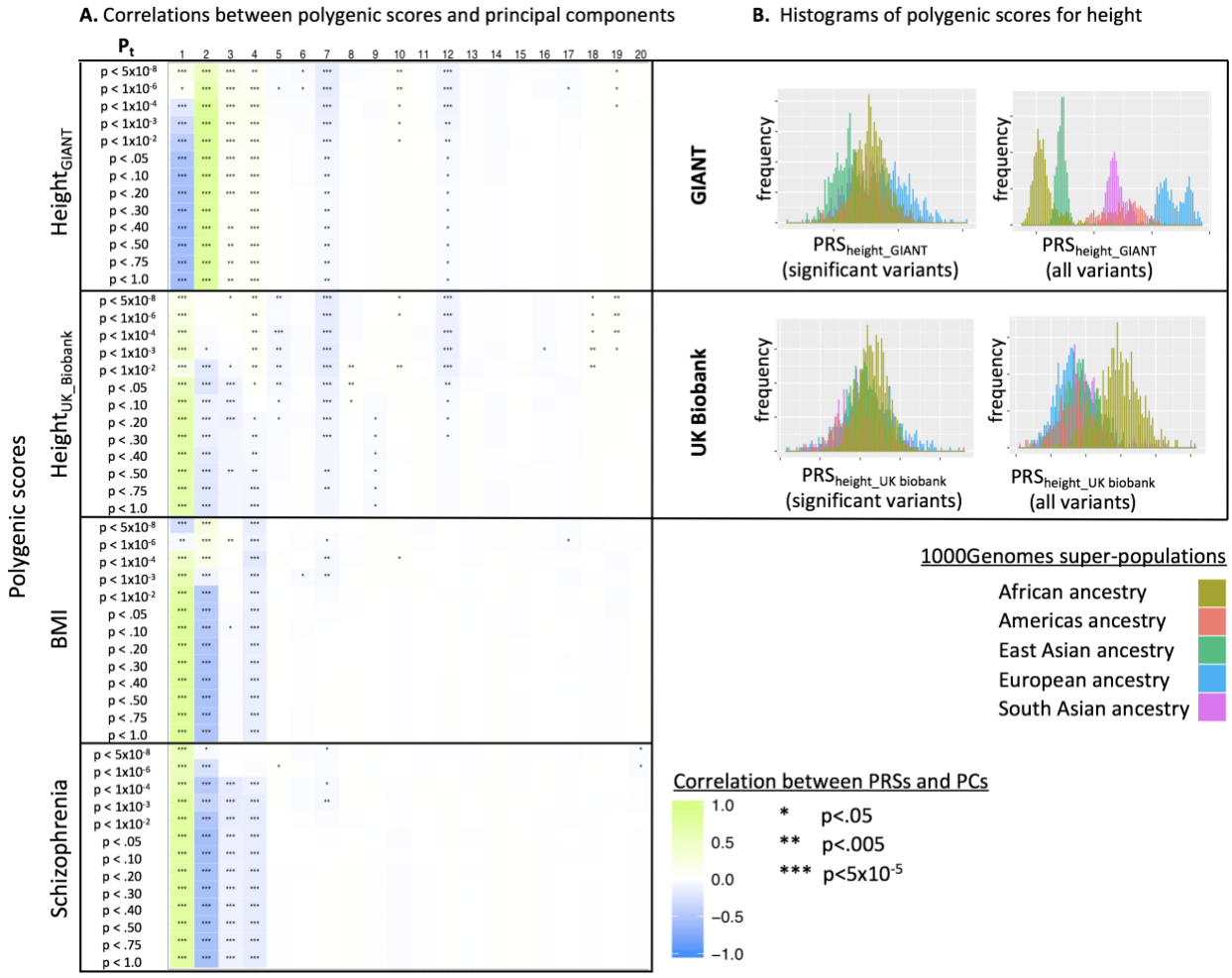


Correlations between global principal components (PCs) and polygenic scores, as currently calculated

We now consider questions about possible differences in polygenic scores among ancestral populations. Polygenic scores, as currently calculated, vary with ancestry. Indeed, polygenic scoring practices from as early as 2009 accounted for this[12]. The method used by Purcell and

colleagues in 2009 (and frequently since) includes two steps for mixed ancestry samples. First, samples are separated into more ancestrally homogeneous subgroups (using visual inspection of plots of principal components calculated on all genetic data from all samples). Second, principal components are calculated again within each of these more ancestrally homogeneous subgroups, and are used as covariates in polygenic scoring analyses, which are conducted separately within each subgroup. **Figure 3** demonstrates why these ancestry analysis procedures have been used, given that many global PCs are strongly correlated with polygenic risk scores, as currently calculated.

Figure 3 Properties of polygenic scores for 1000Genomes participants, across phenotypes (height$_{GIANT}$, height$_{UK_Biobank}$, body mass index, and schizophrenia) and across a range of p-values thresholds used to construct polygenic scores. A. Polygenic scores are often correlated with global principal components. For each phenotype, polygenic scores were constructed using 13 different p-value thresholds as applied to the discovery GWAS (denoted in the left-most column). Twenty global principal components (PCs) were calculated on 1000Genomes participants and are denoted across the top. Within the plot, correlations between each PC and each polygenic score are color-coded to reflect magnitude and direction of correlations from sky blue=-1 to lime=1. Stars indicate statistical significance of each correlation as follows: *p<0.05, **p<0.005, ***Bonferroni significant ($p<5x10^{-5}$). **B.** Histograms of height polygenic scores for 1000Genomes participants are color-coded according to super-population. Two height GWAS were used to construct scores: GIANT (top) and UK biobank (bottom). Scores were constructed using two p-value thresholds: genome-wide significant variants (left) and using all variants (right). As can be seen for both GIANT and UK Biobank-based scores, the choice of p-value threshold applied to the discovery GWAS has a dramatic impact on score distributions among different populations. In general, inclusion of more variants in polygenic scores leads to greater differentiation among distributions of polygenic scores for global populations.



P_t=p-value threshold applied to discovery GWAS in order to construct polygenic scores, PRS=polygenic risk score, GIANT=Genetic Investigation of ANthropomorphic Traits, UK=United Kingdom, BMI=body mass index, PCs=principal components.

Figure 3 shows that polygenic risk scores for the complex genetic phenotypes of height, body mass index (BMI), and schizophrenia are all significantly correlated with various global PCs (**3A**) and also that distributions of scores, as currently calculated, vary across global populations (**3B**). In Figure **3A**, it is primarily the first few PCs that are significantly correlated with polygenic scores, but non-consecutive and later PCs (e.g. 7 and 12 with height polygenic scores) may also be correlated. In order to construct Figure 3, we calculated polygenic scores and global PCs on all 1000Genomes individuals. We applied standard procedures including pruning and weighting alleles based on discovery GWAS results (see **Methods** for additional details). The results show that multiple, sometimes non-consecutive, PCs are strongly correlated with polygenic risk scores, as currently calculated. Many significant correlations are higher than r=.5, and 17.3% of correlations in **Figure 3** are significant after Bonferroni correction for 1,040 tests (i.e. 0.05/1,040= $p<5x10^{-5}$; 20 PCs x 4 phenotypes x 13 p-value thresholds for each discovery GWAS = 1,040 tests). See **Supplementary Table 2** for correlations and corresponding p-values for **Figure 3A**.

In addition to showing significant correlations between global PCs and polygenic scores, as currently calculated, **Figure 3** also demonstrates that the choice of p-value threshold applied to the discovery GWAS (in the construction of polygenic scores) has a dramatic effect on score distributions across populations. The differences are so pronounced that the direction of the correlation between individuals' values for a given PC and their polygenic scores may reverse across the range of p-value thresholds used to construct polygenic scores. For example, 1000Genomes participants' scores for the first global principal component and their GIANT[18]-based polygenic scores for height are modestly positively correlated when only genome wide significant variants are used to construct scores (r=.14, p=$3.4x10^{-12}$, faint green), whereas they are strongly negatively correlated when using all variants to construct polygenic scores (r=-.59, p=$1.9x10^{-223}$, blue). In other words, there are many highly significant correlations, which vary not only in magnitude, but also in direction across the range of p-values used in the construction of polygenic scores. The effects of the methodological choice of p-value threshold on polygenic scores is further demonstrated on the right-hand side of **Figure 3B**, which shows the distributions of GIANT-based height polygenic scores for 1000Genomes participants, using two choices of p-value thresholds (left: genome-wide significant variants, right: all variants). Plots of distributions of UK Biobank-based[46] polygenic scores for height are also shown on the right (bottom two plots).

Figure 3 demonstrates key points relevant to the use of polygenic scores in diverse human populations. First, polygenic scores are often correlated with global PCs, and the correlated PCs are not necessarily consecutive (e.g. global PCs 1-4, 7, and 12 are correlated with height polygenic scores). Second, methodological choices of p-value threshold and discovery GWAS can have dramatic effects on polygenic scores, such that the magnitude and even the direction of observed relationships (e.g. between polygenic scores and PCs) may change across the range of commonly used parameters (e.g. across the range of p-value thresholds used to construct scores). These findings highlight the importance of treating ancestry properly in all analyses involving polygenic risk scores, and suggest that a conservative approach that analyzes polygenic scores separately in each ancestry group may be warranted, at least until a better understanding of polygenic score differences among populations (and across different phenotypes) is achieved.

As noted by Chen and colleagues, explicit modeling of ancestry may afford even greater predictive power with polygenic scores[47].

Assessing putative correlations between global phenotypes and polygenic scores

Finally, we turn to the most difficult question: what causes differences in polygenic scores, as currently calculated, among different populations (e.g. see Figure 3B)? Differences could be real or artifactual (i.e. due to bias in data and/or methods), and five categories of explanations are listed below.

1) True differences due to drift
2) True differences due to selection
3) True differences in genetic effects due to environmental differences (gene-environment interactions)
4) Bias due to uncorrected population stratification in discovery and/or training samples
5) Bias due to discovery/training population data and/or polygenic scoring methods. Specifically, linkage disequilibrium (LD) structure and variant frequency are captured imperfectly with current methods (including genotyping and imputation), and they vary across populations, and currently available data resources are unequally representative of diverse global populations.

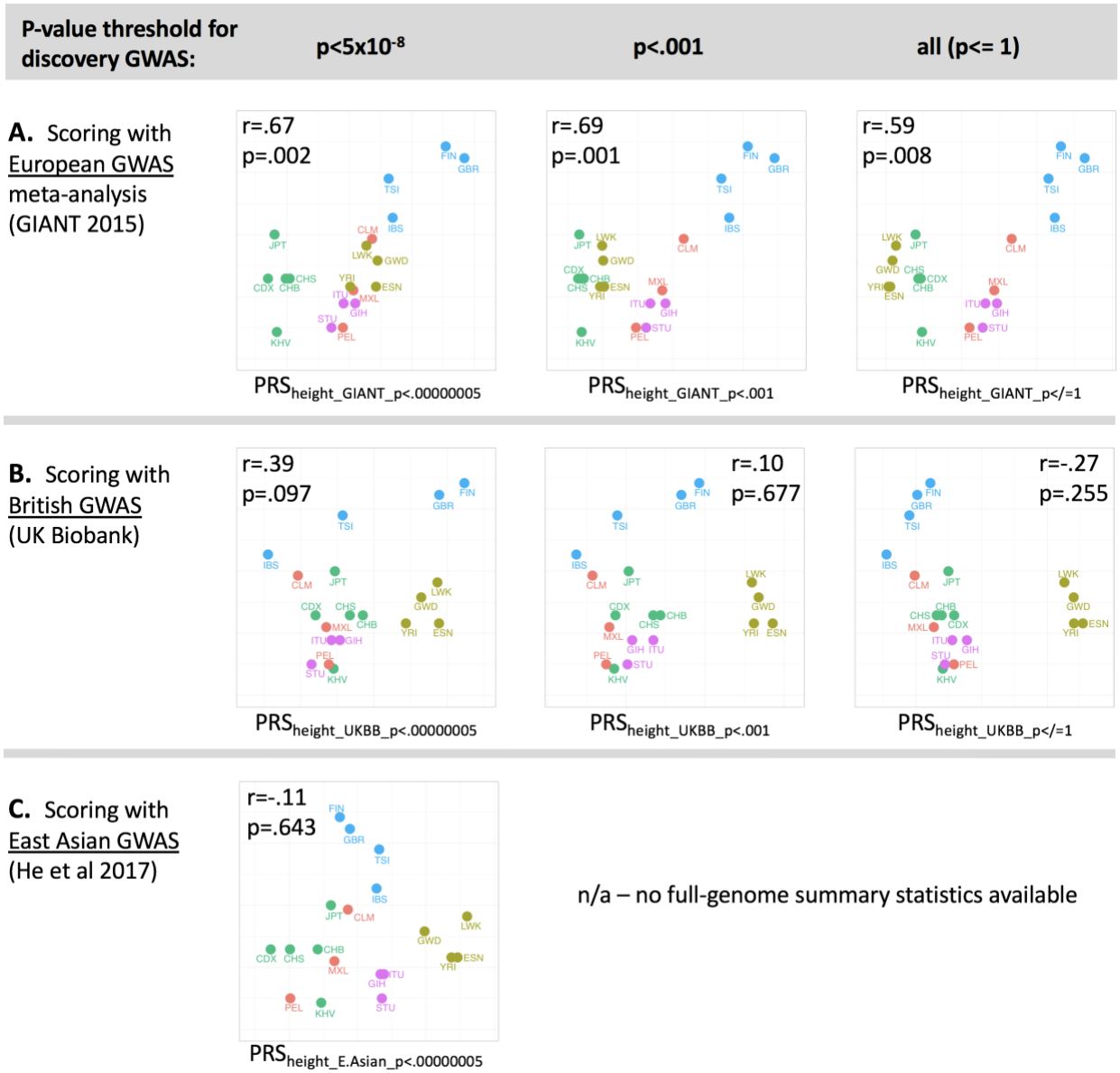
Drift has been implicated as an explanation for population differences in polygenic scores among populations[27], but others have reported that drift is insufficient to explain such differences[33]. Further, initial estimates of the strength of polygenic selection on height in European ancestry populations[33,37] have recently been greatly reduced[30,31], based on findings of uncorrected population stratification in summary statistics from the GIANT Consortium[30,31]. There is also disagreement about whether or not differences in average polygenic scores among populations might contribute to differences in phenotypic values among the same populations (which could also be due to environmental variation). Some have noted apparent positive correlations between average polygenic scores and phenotypes for BMI[34], lupus[35], and height as calculated using GIANT Consortium scores[33,36,37] and one group argued that there is no such correlation for height based on GIANT Consortium scores[27]. As described below, we include more data than used previously to address questions about potential correlations between global height polygenic scores and height phenotypes.

Using 1000Genomes data (as described in the **Methods**), we provide evidence consistent with artifacts contributing to differences in polygenic scores among global populations. In **Figure 4** we plot average polygenic scores for height of 1000Genomes populations on the x-axis, using three sources of weights for constructing scores (PRS=polygenic risk score):

- **4A** (top row) GIANT Consortium[18] based scores: PRS_{height_GIANT}
- **4B** (middle row) UKBiobank[46] based scores from the NealeLab: $PRS_{height_UKBiobank}$
- **4C** (bottom row) East-Asian GWAS based scores[48] from He et al: $PRS_{height_EastAsian}$

On the y-axis, we plot average height for countries of origin for 1000Genomes populations, when available (see **Methods** for details and exclusions).

Figure 4 Scatterplots of height polygenic scores (x-axis) and phenotypic height (y-axis) show that correlations are not consistent across discovery GWAS. The y-values for phenotypic height are the same for each plot, and reflect average height of individuals in the country of origin or each population included. Three different GWAS of height were used to construct polygenic scores (i.e. three rows), and three p-value thresholds (i.e. three columns) were used for polygenic score construction applied to the relevant discovery GWAS. **A.** GIANT-based polygenic scores for height. **B.** UK Biobank-based scores for height. **C.** East-Asian-based polygenic scores for height. The last two plots are missing because only genome-wide significant variants were available for the East Asian GWAS of height[48].



GWAS=genome-wide association study, GIANT=Genetic Investigation of ANthropomorphic Traits, PRS=polygenic risk score, UK=United Kingdom, PCs=principal components. Population abbreviations are standard abbreviations from the 1000Genomes project[42,45] and are included in Supplementary Table 3.

As shown in **Figure 4A**, height phenotypes for global populations (y-axes) are positively correlated with GIANT-based[18] polygenic scores for height (x-axes), but not with UK-Biobank-

based polygenic scores (**4B**) or East-Asian GWAS based polygenic scores (**4C**). Polygenic scores constructed using only genome-wide significant variants from GIANT (top left) were positively correlated with height phenotypes (r=.67, p=.002), as were scores constructed using larger numbers of GIANT-based variants (e.g. all variants, top right, r=.59, p=.008). Results in **4B** and **4C** demonstrate that correlations (or lack of correlations) between height and polygenic scores for height are dependent on discovery GWAS. Recent findings suggest that correction for population stratification may not have been adequate in GIANT[30,31], and therefore the positive correlations observed in **4A** may be partially due to uncorrected population stratification. The dependence of correlation estimates on discovery GWAS is further illustrated in **4C**, in which the point estimate for correlation between height and East Asian GWAS based polygenic scores for height is negative (r=-.11, p=.643). Power in discovery GWAS is also relevant, and greater confidence should be attributed to the results in **4A** and **4B** because both European ancestry discovery GWAS were adequately powered to detect hundreds of height loci, whereas the East Asian height GWAS was only adequately powered to detect 17 loci.

These results suggest that both the ancestry of the participants in the discovery GWAS (e.g. European[18,46] vs. East Asian[48]) and uncorrected population stratification[30,31] contribute to observed positive correlations between GIANT-based polygenic scores for height and height across global populations. UK Biobank-based polygenic scores provide no evidence of a positive correlation between heights and mean height polygenic scores for populations. More research is needed to better understand the exact causes of differences in score distributions across populations and their putative relationships to phenotypes. Future research must also account for environmental effects on phenotypes, as well as variability in measurement validity and reliability across populations. Even for the relatively simple example of height (which is easily measured and for which major environmental influences are relatively well-understood) our analyses suggest that a great deal of caution should be used in drawing conclusions about polygenic score differences underling global phenotypic differences, until data resources are significantly improved (i.e. well-powered GWAS in diverse populations), and until a deeper understanding of relevant population genetics principles has emerged. As discussed further below, even more caution will be required for other phenotypes such as psychiatric disorders.

Discussion

As conversations about personalized medicine make their way to the general public, it is important to recognize the need to include underrepresented populations in genetic studies. Among other concerns, the inclusion of participants representing diverse ancestries in research is imperative to ensure equitable benefit from scientific discoveries for diverse populations, and to prevent further increase in health disparities. Relevant to these longer-term objectives, our findings provide foundational information about polygenic risk score usage among diverse populations, summarized in four key points. First, polygenic scoring studies have primarily been conducted in European and East Asian ancestry populations. Second, the performance of polygenic scores in non-European populations is generally poorer than performance in European ancestry samples, and much worse in African ancestry samples. Third, polygenic scores for complex genetic phenotypes are often correlated with global principal components. Fourth, appropriate data resources are lacking to address most questions about putative differences in

polygenic scores across global (non-European ancestry) populations. The straightforward, albeit expensive and time-consuming, solution to improving polygenic score performance across diverse populations is to create well-powered GWAS data resources for many different global populations.

Our finding that the predictive power of polygenic risk scores in poorer in non-European populations, particularly among African ancestry individuals, is almost certainly due to the considerably greater genetic diversity within many African populations, the phenotypic consequences of which is not well-captured by European-ancestry GWAS. Indeed, the relative impoverishment of genetic diversity among European ancestry and other non-African populations (due to past population bottlenecks) is a natural limitation of European ancestry GWAS. In addition to the important goal of collection of more samples from more populations, there can also be improvements in the manner in which variant frequencies and linkage disequilibrium are handled in polygenic scoring studies. The results presented here provide benchmarks for the relative performance of polygenic scores in diverse populations, as compared to performance in populations of European ancestry. This is important, because it not only informs power calculations for future research, but also highlights relative differences in predictive utility across diverse populations, which must be considered in public health calculations regarding when and how to use polygenic risk scores. Available data, confined mostly to European ancestry samples, provides a best-case-scenario upon which policy decisions may be based, rather than an accurate portrayal of how polygenic scores will work in ancestrally diverse patient samples.

Concerning expectations about polygenic score performance in non-European ancestry samples we note that polygenic scores for many complex genetic phenotypes are strongly correlated with global PCs, which highlights the critical importance of appropriate statistical methods for the analysis of genetic data from non-European ancestry populations. Testing future polygenic scoring results for robustness to the inclusion of variable combinations of PCs will reduce the chances for spurious results (that are actually attributable to ancestry). In sum, the preponderance of genetic studies based on European ancestry samples has led to a situation in which polygenic scores are approximately one-third as informative for African ancestry individuals, as they are for European ancestry individuals. This is presumably true for commercially available tests as well, and consumers should be aware of the differential performance of tests across individuals.

Regarding scientific and public perception of polygenic scores, it is important to address apparent differences in polygenic score distributions across populations. Our findings suggest that it is currently not possible to know precisely the distribution of polygenic scores for diverse non-European populations, for any complex genetic phenotype, because data resources for most populations are currently inadequate. Further, as we have shown, the ordering of population distributions of polygenic scores varies under accepted methods of constructing these scores (i.e. using different p-value thresholds for variant inclusion in scores and using alternative discovery GWAS). Explanations for these differences are currently incomplete. Until vastly superior data resources are available – including large scale GWAS in multiple globally representative populations – scientists are unlikely to reach consensus regarding the existence, nature, and exact causes of polygenic score differences among populations.

In our analysis of possible relationships between average phenotypes for global populations and average polygenic scores for those phenotypes in global populations, we chose to examine height because it is easily measured and because factors affecting height (e.g. nutrition) are also relatively easily quantified. In contrast, research on other variables such as weight, smoking status, psychological symptoms, and cognitive performance requires more careful control for environmental confounders (including variables like social status), which are often correlated with ancestry and therefore may also be correlated with global principal components and polygenic scores (as currently calculated). This means that confounding of environmental and genetic effects is likely. For example, social experiences such as being subjected to racism are prime candidates for confounding in genetic studies.

In closing, we emphasize the need to engage experts from other disciplines, such as social psychology and bioethics, as geneticists attempt to characterize genetic effects on complex genetic phenotypes impacted by social factors (especially psychological and cognitive variables). This is necessary because societal influences including socioeconomic status and discrimination can powerfully influence these phenotypes[52], and these causal social factors often co-vary with ancestry. In genetic research, there is potential for relative blindness to non-genetic influences on phenotypes. Consequently, experts must be consulted in order to properly account for non-genetic influences on many complex genetic phenotypes and gene-environment interplay (i.e. correlations and interactions). Nevertheless, with cautious and broadly-informed research, potential medical benefits of correctly interpreting polygenic variation within and among populations can be realized.

Acknowledgements

Pilot grant to LED and BD from the Stanford Center for Clinical and Translation Research and Education (UL1 TR001085, PI Greenberg) helped fund this work. The Stanford Center for Computational, Evolutionary, and Human Genetics, CEHG, supported this work.

Author Contributions

LD and BD designed the project, LD and HS conducted the analyses, BG and KR contributed data, RP and MF provided critical feedback and framing, all authors contributed to writing of the manuscript.

Competing Interests

The authors declare no competing interests.

Materials and Correspondence

Correspondence and materials requests should be addressed to LD, Laramied@Stanford.edu

References

1. Bustamante CD, Burchard EG, De la Vega FM. Genomics for the world. Nature. 2011;475(7355):163–165. doi:10.1038/475163a

2. Petrovski S, Goldstein DB. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. Genome Biology. 2016;17:157. doi:10.1186/s13059-016-1016-y

3. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016;538(7624):161–164. doi:10.1038/538161a

4. Duncan LE, Pollastri AR, Smoller JW. Mind the gap: Why many geneticists and psychological scientists have discrepant views about gene-environment interaction (G×E) research. The American psychologist. 2014;69(3):249–268. doi:10.1037/a0036320

5. Dalvie S, Koen N, Duncan L, Abbo C, Akena D, Atwoli L, Chiliza B, Donald KA, Kinyanda E, Lochner C, et al. Large Scale Genetic Research on Neuropsychiatric Disorders in African Populations is Needed. EBioMedicine. 2015;2(10):1259–1261. doi:10.1016/j.ebiom.2015.10.002

6. Wojcik G, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, Highland HM, Patel YM, Sorokin EP, Avery CL, et al. Genetic Diversity Turns a New PAGE in Our Understanding of Complex Traits. bioRxiv. 2017 Sep 15:188094. doi:10.1101/188094

7. CONVERGE Consortium. Sparse whole-genome sequencing identifies two loci for major depressive disorder. Nature. 2015;523(7562):588–591. doi:10.1038/nature14659

8. Vargas JD, Manichaikul A, Wang X-Q, Rich SS, Rotter JI, Post WS, Polak JF, Budoff MJ, Bluemke DA. Common genetic variants and subclinical atherosclerosis: The Multi-Ethnic Study of Atherosclerosis (MESA). Atherosclerosis. 2016;245:230–236. doi:10.1016/j.atherosclerosis.2015.11.034

9. Qi Q, Stilp AM, Sofer T, Moon J-Y, Hidalgo B, Szpiro AA, Wang T, Ng MCY, Guo X, MEta-analysis of type 2 DIabetes in African Americans (MEDIA) Consortium, et al. Genetics of Type 2 Diabetes in U.S. Hispanic/Latino Individuals: Results From the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). Diabetes. 2017;66(5):1419–1425. doi:10.2337/db16-1150

10. Duncan LE, Ratanatharathorn A, Aiello AE, Almli LM, Amstadter AB, Ashley-Koch AE, Baker DG, Beckham JC, Bierut LJ, Bisson J, et al. Largest GWAS of PTSD (N=20 070) yields genetic overlap with schizophrenia and sex differences in heritability. Molecular Psychiatry. 2017 Apr 25. doi:10.1038/mp.2017.77

11. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Research. 2007;17(10):1520–1528. doi:10.1101/gr.6665407

12. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460(7256):748–752. doi:10.1038/nature08185

13. Hoffmann TJ, Ehret GB, Nandakumar P, Ranatunga D, Schaefer C, Kwok P-Y, Iribarren C, Chakravarti A, Risch N. Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. Nature Genetics. 2017;49(1):54–64. doi:10.1038/ng.3715

14. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511(7510):421–427. doi:10.1038/nature13595

15. Knowles JW, Ashley EA. Cardiovascular disease: The rise of the genetic risk score. PLOS Medicine. 2018;15(3):e1002546. doi:10.1371/journal.pmed.1002546

16. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nature Reviews Genetics. 2018 May 22:1. doi:10.1038/s41576-018-0018-x

17. Ehret GB, Ferreira T, Chasman DI, Jackson AU, Schmidt EM, Johnson T, Thorleifsson G, Luan J, Donnelly LA, Kanoni S, et al. The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals. Nature Genetics. 2016;48(10):1171–1184. doi:10.1038/ng.3667

18. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J, Kutalik Z, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nature Genetics. 2014;46(11):1173–1186. doi:10.1038/ng.3097

19. Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, Julier C, Morahan G, Nerup J, Nierras C, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nature Genetics. 2009;41(6):703–707. doi:10.1038/ng.381

20. Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, Adams MJ, Agerbo E, Air TM, Andlauer TMF, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nature Genetics. 2018;50(5):668–681. doi:10.1038/s41588-018-0090-3

21. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. Pitfalls of predicting complex traits from SNPs. Nature Reviews. Genetics. 2013;14(7):507–515. doi:10.1038/nrg3457

22. Grinde KE, Qi Q, Thornton TA, Liu S, Shadyab AH, Chan KHK, Reiner AP, Sofer T. Generalizing Genetic Risk Scores from Europeans to Hispanics/Latinos. bioRxiv. 2018 Jan 4:242404. doi:10.1101/242404

23. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, Genovese G, Loh P-R, Bhatia G, Do R, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. American Journal of Human Genetics. 2015;97(4):576–592. doi:10.1016/j.ajhg.2015.09.001

24. Okser S, Pahikkala T, Airola A, Salakoski T, Ripatti S, Aittokallio T. Regularized Machine Learning in the Genetic Prediction of Complex Traits. PLOS Genetics. 2014;10(11):e1004754. doi:10.1371/journal.pgen.1004754

25. Paré G, Mao S, Deng WQ. A machine-learning heuristic to improve gene score prediction of polygenic traits. Scientific Reports. 2017;7(1):12665. doi:10.1038/s41598-017-13056-1

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

27. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, Bustamante CD, Kenny EE. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. American Journal of Human Genetics. 2017;100(4):635–649. doi:10.1016/j.ajhg.2017.03.004

28. Scutari M, Mackay I, Balding D. Using Genetic Distance to Infer the Accuracy of Genomic Prediction. PLOS Genetics. 2016;12(9):e1006288. doi:10.1371/journal.pgen.1006288

29. Curtis D. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. bioRxiv. 2018 Mar 23:287136. doi:10.1101/287136

30. Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, Chiang CWK, Hirschhorn JN, Daly MJ, Patterson N, et al. Signals of polygenic adaptation on height have been overestimated due to uncorrected population structure in genome-wide association studies. bioRxiv. 2018 Jun 28:355057. doi:10.1101/355057

31. Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, Boyle EA, Zhang X, Racimo F, Pritchard JK, et al. Reduced signal for polygenic adaptation of height in UK Biobank. bioRxiv. 2018 Jun 27:354951. doi:10.1101/354951

32. Ware EB, Schmitz LL, Faul JD, Gard A, Mitchell C, Smith JA, Zhao W, Weir D, Kardia SL. Heterogeneity in polygenic scores for common human traits. 2017 Feb 5. doi:10.1101/106062

33. Guo J, Wu Y, Zhu Z, Zheng Z, Trzaskowski M, Zeng J, Robinson MR, Visscher PM, Yang J. Global genetic differentiation of complex traits shaped by natural selection in humans. Nature Communications. 2018;9(1):1865. doi:10.1038/s41467-018-04191-y

34. Mao L, Fang Y, Campbell M, Southerland WM. Population differentiation in allele frequencies of obesity-associated SNPs. BMC genomics. 2017;18(1):861. doi:10.1186/s12864-017-4262-9

35. Morris DL, Sheng Y, Zhang Y, Wang Y-F, Zhu Z, Tombleson P, Chen L, Cunninghame Graham DS, Bentham J, Roberts AL, et al. Genome-wide association meta-analysis in Chinese and European individuals identifies ten new loci associated with systemic lupus erythematosus. Nature Genetics. 2016;48(8):940–946. doi:10.1038/ng.3603

36. Robinson MR, Hemani G, Medina-Gomez C, Mezzavilla M, Esko T, Shakhbazov K, Powell JE, Vinkhuyzen A, Berndt SI, Gustafsson S, et al. Population genetic differentiation of height and body mass index across Europe. Nature Genetics. 2015;47(11):1357–1362. doi:10.1038/ng.3401

37. Racimo F, Berg JJ, Pickrell JK. Detecting Polygenic Adaptation in Admixture Graphs. Genetics. 2018 Jan 18:genetics.300489.2017. doi:10.1534/genetics.117.300489

38. Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, Yengo L, Rocheleau G, Froguel P, McCarthy MI, et al. Detection of human adaptation during the past 2000 years. Science. 2016;354(6313):760–764. doi:10.1126/science.aag0776

39. Berg JJ, Coop G. A population genetic signal of polygenic adaptation. PLoS genetics. 2014;10(8):e1004412. doi:10.1371/journal.pgen.1004412

40. Cavalli-Sforza L, Feldman MW. Models for cultural inheritance I. Group mean and within group variation. Theoretical Population Biology. 1973;4(1):42–55. doi:10.1016/0040-5809(73)90005-1

41. Creanza N, Kolodny O, Feldman MW. Cultural evolutionary theory: How culture evolves and why it matters. Proceedings of the National Academy of Sciences. 2017 Jul 20:201620732. doi:10.1073/pnas.1620732114

42. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061–1073. doi:10.1038/nature09534

43. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, Ripke S, Lee JC, Jostins L, Shah T, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nature Genetics. 2015;47(9):979–986. doi:10.1038/ng.3359

44. Brown BC, Ye CJ, Price AL, Zaitlen N. Transethnic Genetic-Correlation Estimates from Summary Statistics. The American Journal of Human Genetics. 2016;99(1):76–88. doi:10.1016/j.ajhg.2016.05.001

45. Consortium T 1000 GP. A global reference for human genetic variation. Nature. 2015;526(7571):68–74. doi:10.1038/nature15393

46. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, et al. Genome-wide genetic data on ~500,000 UK Biobank participants. bioRxiv. 2017 Jul 20:166298. doi:10.1101/166298

47. Chen C-Y, Han J, Hunter DJ, Kraft P, Price AL. Explicit modeling of ancestry improves polygenic risk scores and BLUP prediction. Genetic epidemiology. 2015;39(6):427–438. doi:10.1002/gepi.21906

48. He M, Xu M, Zhang B, Liang J, Chen P, Lee J-Y, Johnson TA, Li H, Yang X, Dai J, et al. Meta-analysis of genome-wide association studies of adult height in East Asians identifies 17 novel loci. Human Molecular Genetics. 2015;24(6):1791–1800. doi:10.1093/hmg/ddu583

49. O'Sullivan BP, Freedman SD. Cystic fibrosis. Lancet (London, England). 2009;373(9678):1891–1904. doi:10.1016/S0140-6736(09)60327-5

50. Rees DC, Williams TN, Gladwin MT. Sickle-cell disease. Lancet (London, England). 2010;376(9757):2018–2031. doi:10.1016/S0140-6736(10)61029-X

51. PriceMay. 22 M, 2018, Pm 2:00. 'It's a toxic place.' How the online world of white nationalists distorts population genetics. Science | AAAS. 2018 May 22 [accessed 2018 Aug 8]. http://www.sciencemag.org/news/2018/05/it-s-toxic-place-how-online-world-white-nationalists-distorts-population-genetics

52. Salvatore J, Shelton JN. Cognitive Costs of Exposure to Racial Prejudice. Psychological Science. 2007;18(9):810–815. doi:10.1111/j.1467-9280.2007.01984.x

53. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4(1):7. doi:10.1186/s13742-015-0047-8

54. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. American Journal of Human Genetics. 2007;81(3):559–575.

55. Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. 2005 [accessed 2013 Oct 4]. http://www.R-project.org

Figure Legends

For ease of review, figure legends are also supplied in the main text.

Figure 1. The first decade of polygenic scoring studies (2008-2017) focused primarily on European ancestry samples (N=733 studies). A. Cumulative numbers of studies by year are denoted by color. The stacked bar graph below the cumulative distribution plot shows proportional ancestry by year. **B.** Pie charts depict world ancestry representation (left) and polygenic scoring study representation (right). **C.** The percentage representation for each ancestry group is given, such that 100% would indicate equal representation in the world and in polygenic scoring

studies. For example, European ancestry samples are over-represented (460%) whereas African ancestry samples are under-represented (17%).

Figure 2 Performance of polygenic scores in Latino, East Asian, South Asian, and African ancestry samples, relative to performance in matched European ancestry samples (29 total studies). In order to make data comparable across studies, performance in each study was standardized to European ancestry performance, hence the single European ancestry y-axis value of 100%. Each point represents one pair of polygenic scoring analyses between a European ancestry sample and a matched sample from another ancestry (see text for details).

Figure 3 Properties of polygenic scores for 1000Genomes participants, across phenotypes (height$_{GIANT}$, height$_{UK_Biobank}$, body mass index, and schizophrenia) and across a range of p-values thresholds used to construct polygenic scores. A. Polygenic scores are oftentimes correlated with global principal components. For each phenotype, polygenic scores were constructed using 13 different p-value thresholds as applied to the discovery GWAS (denoted in the left-most column). Twenty global principal components (PCs) were calculated on 1000Genomes participants and are denoted across the top. Within the plot, correlations between each PC and each polygenic score are color-coded to reflect magnitude and direction of correlations from sky blue=-1 to lime=1. Stars indicate statistical significance of each correlation as follows: *p<0.05, **p<0.005, ***Bonferroni significant. **B.** Histograms of height polygenic scores for 1000Genomes participants are color-coded according to super-population. Two height GWAS were used to construct scores: GIANT (top) and UK biobank (bottom). Scores were constructed using two p-value thresholds: genome-wide significant variants (left) and using all variants (right). As can be seen for both GIANT and UK Biobank-based scores, the choice of p-value threshold applied to the discovery GWAS has a dramatic impact on score distributions among different populations. In general, inclusion of more variants in polygenic scores leads to greater differentiation in distributions of polygenic scores for global populations.

Figure 4 Scatterplots of height polygenic scores (x-axis) and phenotypic height (y-axis) show that correlations are not consistent across discovery GWAS. The y-values for phenotypic height are the same for each plot, and reflect average height of individuals in the country of origin or each population included. Three different GWAS of height were used to score populations (i.e. three rows), and three p-value thresholds (i.e. three columns) were used for polygenic score construction as applied to the relevant discovery GWAS. **A.** GIANT-based polygenic scores for height. **B.** UK Biobank-based scores for height. **C.** East-Asian-based polygenic scores for height. The last two plots are missing because only genome-wide significant variants were available for the East Asian GWAS of height[48].