Hidden 'risk' in polygenic scores: clinical use today could exacerbate health disparities

Alicia R. Martin[1,2,3], Masahiro Kanai[1,2,3,4,5], Yoichiro Kamatani[5,6], Yukinori Okada[5,7,8], Benjamin M. Neale[1,2,3], Mark J. Daly[1,2,3,9]

[1] Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA

[2] Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

[3] Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

[4] Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA

[5] Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan

[6] Kyoto-McGill International Collaborative School in Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto 606-8507, Japan

[7] Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita 565-0871, Japan

[8] Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita 565-0871, Japan

[9] Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland

Abstract

Polygenic risk scores have the potential to improve health outcomes for a variety of complex diseases and are poised for clinical translation, driven by the low cost of genotyping (<$50 per person), the ability to predict genetic risk of many diseases with a single test, and the dramatically increasing scale and power of genetic studies that aid prediction accuracy. However, the major *ethical* and *scientific* challenge surrounding clinical implementation is the observation that they are currently of far greater predictive value in European ancestry individuals than others. The better performance of such risk scores in European populations is an inescapable consequence of the heavily biased makeup of genome-wide association studies, with an estimated 79% of participants in all these existing studies being of European descent. Empirically, polygenic risk scores perform far better in European populations, with prediction accuracy reduced by approximately 2- to 5-fold in East Asian and African descent populations, respectively. This highlights that—unlike specific clinical biomarkers and prescription drugs, which may individually work better in some populations but do not ubiquitously perform far better in European populations–clinical uses of prediction today would systematically afford greater improvement to European descent populations. Early diversifying efforts, however, show promise in levelling this vast imbalance, even when non-European sample sizes are considerably smaller than the best-powered studies to date. Polygenic

46 risk scores provide a new opportunity to improve health outcomes for many diseases in
47 all populations, but to realize this full potential equitably, we must prioritize greater
48 inclusivity of diverse participants in genetic studies and open access to resulting
49 summary statistics to ensure that health disparities are not increased for those already
50 most underserved.
51
52 **Keywords**: health disparities, genetic risk prediction, polygenic risk scores, diversity,
53 population genetics, statistical genetics
54
55 Polygenic risk scores (PRS), which predict complex traits using genetic data, are of
56 burgeoning interest to the clinical community as researchers demonstrate their growing
57 power to improve clinical care, genetic studies of a wide range of phenotypes increase
58 in size and power, and genotyping costs plummet to less than US$50. Many earlier
59 criticisms of limited prediction power are now recognized to have been chiefly an issue
60 of insufficient sample size, which is no longer the case for many outcomes [1]. For
61 example, polygenic risk scores alone already predict breast and prostate cancer risk in
62 European descent patients more accurately than current clinical models [2-4]. Additionally,
63 integrated models of PRS together with other lifestyle and clinical factors have enabled
64 clinicians to more accurately quantify the risk of heart attack for patients; consequently,
65 they have more effectively targeted the reduction of LDL cholesterol and by extension
66 heart attack by prescribing statins to patients at the greatest overall risk of
67 cardiovascular disease [5-9]. Promisingly, return of genetic risk of complex disease to at-
68 risk patients does not induce significant self-reported negative behavior or psychological

function, and some potentially positive behavioral changes have been detected [10]. While we share enthusiasm about the potential of PRS to improve health outcomes through their eventual routine implementation as clinical biomarkers, we consider the consistent observation that they are currently of far greater predictive value in individuals of recent European descent than in others to be the major *ethical* and *scientific* challenge surrounding clinical translation and, at present, the most critical limitation to genetics in precision medicine. The scientific basis of this imbalance has been demonstrated in population genetics simulations, theoretically, and empirically across many traits and diseases [11-22].

All studies to date using well-powered genome-wide association studies (GWAS) to assess the predictive value of PRS in European and non-European descent populations have made a consistent observation: PRS predict individual risk far more accurately in Europeans than non-Europeans. In complex traits including height, body mass index (BMI), educational attainment, schizophrenia, and major depression, existing PRS computed with the largest available GWAS results predict outcomes far more accurately in new samples of European-descent than they do in non-Europeans, with the clearest study examples in East Asians and African Americans [15,16,18-24]. Rather than chance or biology, this is a predictable consequence of the fact that the genetic discovery efforts to date heavily underrepresent non-European populations globally. The correlation between true and genetically predicted phenotypes decays with genetic divergence from the makeup of the discovery GWAS, meaning that the accuracy of polygenic scores in different populations is highly dependent on the study population

92 representation in the largest existing 'training' GWAS. Here, we document study biases
93 that underrepresent non-European populations in current GWAS, and explain the
94 fundamental concepts contributing to reduced phenotypic variance explained with
95 increasing genetic divergence from populations included in GWAS.
96
97 **Predictable basis of disparities in polygenic risk score accuracy**
98 Poor generalizability of genetic studies across populations arises from the
99 overwhelming abundance of European descent studies and dearth of well-powered
100 studies in globally diverse populations [25-28]. According to the GWAS catalog, ~79% of all
101 GWAS participants are of European descent despite making up only 16% of the global
102 population (**Figure 1**). This is especially problematic as previous studies have shown
103 that Hispanic/Latino and African American studies have contributed an outsized number
104 of associations relative to studies of similar sizes in Europeans [27]. More concerningly,
105 the fraction of non-European individuals in GWAS has stagnated or declined since late
106 2014 (**Figure 1**), suggesting that we are not on a trajectory to correct this imbalance.
107 These numbers provide a composite metric of study availability, accessibility, and use—
108 i.e., cohorts that have been included in numerous GWAS are represented multiple
109 times, which may disproportionately include cohorts of European descent. Whereas the
110 average sample sizes of GWAS in Europeans continue to grow though, they have
111 stagnated and remain several-fold smaller in other populations (**Figure S1**).
112
113 The relative sample compositions of GWAS result in highly predictable disparities in
114 prediction accuracy; statistical and population genetics theory predicts that genetic risk

115 prediction accuracy will decay with increasing genetic divergence between the original
116 GWAS sample and target of prediction, a function of population history [13,14]. This pattern
117 can be attributed to several statistical observations which we detail below: 1) GWAS
118 favor the discovery of genetic variants that are common in the study population; 2)
119 linkage disequilibrium (LD) differentiates marginal effect size estimates for highly
120 polygenic traits across populations, even when causal variants are the same; and 3)
121 environment and demography differ across populations, which may interact to drive
122 differential forces of natural selection that in turn drive differences in causal genetic
123 architecture*. Of note, the first two of these degrade prediction performance across
124 populations substantially even when there exist no biological, environmental, or
125 diagnostic differences.
126
127 *Common discoveries and low-hanging fruit*
128 First, the power to discover an association in a genetic study depends on the effect size
129 and frequency of the variant [29]. This power dependence means that the most significant
130 associations tend to be more common in the populations in which they are discovered
131 than in other populations [13,30]. For example, GWAS catalog variants are on average
132 more common in European populations compared to East Asian and African
133 populations (**Figure 2B**), an observation not representative of genomic variants at large.
134 Understudied populations offer low-hanging fruit for genetic discovery because variants
135 that are common in these groups but rare or absent in European populations could not

* We define the causal genetic architecture as the true effects (typically written as β in the GWAS literature) of variants that impact a phenotype that would be identified in a population of infinite sample size. Unlike effect size estimates (usually written as $\hat{\beta}$), true effects are typically modeled as invariant with respect to LD and allele frequency differences across populations.

be discovered even with very large European sample sizes. Some examples include *SLC16A11* and *HNF1A* associations with type II diabetes in Latino populations, *APOL1* associations with end-stage kidney disease, and associations with prostate cancer in African descent populations [31-34]. If we assume that causal genetic variants have an equal effect across all populations—an assumption with some empirical support that offers the best case scenario for transferability [35-40]—Eurocentric GWAS biases mean that variants that are common in European populations are preferentially discovered and associated with risk, and thus account for a larger fraction of the variance in polygenic risk [13]. Furthermore, imputation reference panels share the same study biases as in GWAS [41], creating challenges for imputing sites that are rare in European populations but common elsewhere when the catalog of non-European haplotypes is substantially smaller. These issues are insurmountable through statistical methods alone [13], but rather motivate substantial investments in more diverse populations to produce similar-sized GWAS of biomedical phenotypes as well as sequenced reference panels in other populations.

Linkage disequilibrium

Second, the correlation structure of the human genome, i.e. LD, varies across populations due to demographic history (**Figure 2A,C-E**). These LD differences in turn drive differences in effect size estimates (i.e. predictors) from GWAS across populations, even when causal effects are the same. (Mathematically, the marginal GWAS estimate $\hat{\beta}_j = \sum_{k=1}^{m} r_{j,k}\beta_k + \epsilon_j$, where $\hat{\beta}_j$ are effect size estimates at SNP *j*, $r_{j,k}$ is pairwise SNP LD between SNPs *j* and *k*, β_k is the causal SNP effect at nearby SNP *k*,

and ϵ is residual error from bias or noise. More simply, when causal effects are the same across populations, effect size *estimates* at SNPs tagging these causal variants from which we construct predictors will differ across populations in proportion to LD between tagging and causal SNP pairs.) While differences in effect size estimates due to LD differences may typically be small for most regions of the genome (**Figure 2C-E**), PRS sum across these effects, also aggregating these population differences. Statistical methods that account for LD differences across populations may help improve risk prediction accuracy within each population. While empirical studies suggest that causal effect sizes tend to be shared [35-40], it may not be feasible to fine-map most variants to a single locus to solve issues of low generalizability (i.e. using causal rather than correlated effect size estimates), even with very large GWAS. This is because complex traits are highly polygenic, meaning most of our prediction power comes from small effects that do not meet genome-wide significance and/or cannot be fine-mapped, even in many of the best-powered GWAS to date [42].

History, selection, the environment, and complex interactions

Lastly, other environmental, demographic, and cohort considerations may further worsen prediction accuracy differences across populations in less predictable ways. GWAS ancestry study biases and LD differences across populations are extremely challenging to address, but these issues actually make many favorable assumptions that all causal loci have the same impact and are under equivalent selective pressure in all populations. In contrast, other effects on polygenic adaptation or risk scores such as long-standing environmental differences across global populations that have resulted in differing responses of natural selection can impact populations differently based on their

unique histories. Additionally, residual uncorrected population stratification may impact risk prediction accuracy across populations, but the magnitude of its effect is currently unclear. These effects are particularly challenging to disentangle, as has clearly been demonstrated for height, where evidence of polygenic adaptation and/or its relative magnitude is under question [43,44]. Comparisons of geographically stratified phenotypes like height across populations with highly divergent genetic backgrounds and mean environmental differences, such as differences in resource abundance during development across continents, are especially prone to confounding from correlated environmental and genetic divergence [43,44]. This residual stratification can lead to over-predicted differences across geographical space [45].

Related to stratification, most polygenic scoring methods do not explicitly address recent admixture and none consider recently admixed individuals' unique local mosaic of ancestry—further methods development in this space is needed. Furthermore, comparing PRS across environmentally stratified cohorts, such as in some biobanks with healthy volunteer effects versus disease study datasets or hospital-based cohorts, requires careful consideration of technical differences, collider bias, as well as variability in baseline health status among studies. It is also important to consider differences in clinical definition of the phenotypes and heterogeneous constitution of sub-phenotypes among countries.

Another consideration is the fact that, like other existing biomarkers, the predictive utility of PRS may change as a function of age, consistent with age-dependent heritability for

some traits [46]. For example, increasing age is associated with higher risk of coronary artery disease, and higher PRS accelerate this increased risk [47]. Consequently, the age of intervention e.g. with statins needs to be evaluated in aggregate with other clinical risk factors that change over time. Autism spectrum disorder and schizophrenia also have a genetic basis with differing developmental trajectories; their shared genetic influences decrease with age, whereas the genetic overlap between schizophrenia and social communication difficulties persists with age [48]. Further work on how prediction accuracy varies as a function of age across phenotypes is needed.

Differences in environmental exposure, gene × gene interactions, gene × environment interactions, historical population size dynamics, and other factors will further limit generalizability for genetic risk scores in an unpredictable, trait-specific fashion [49,50]. Complex traits do not behave in a genetically deterministic manner, with some environmental factors dwarfing individual genetic effects, creating outsized issues of comparability across globally diverse populations; among psychiatric disorders for example, whereas schizophrenia has a nearly identical genetic basis across East Asians and Europeans (r_g=0.98) [40], substantially different rates of alcohol use disorder across populations is partially explained by differences in availability and genetic differences impacting alcohol metabolism [51]. While non-linear genetic factors explain little variation in complex traits beyond a purely additive model [52], some unrecognized nonlinearities and gene × gene interactions can also induce genetic risk prediction challenges, as pairwise interactions are likely to vary more across populations than individual SNPs. Mathematically, we can simplistically think of this in terms of a two-

SNP model, in which the sum of two SNP effects is likely to explain more phenotypic variance than the product of the same SNPs. Some machine learning approaches may thus modestly improve genetic prediction accuracy beyond current approaches for some phenotypes [53], but these approaches are most likely to improve prediction accuracy for atypical traits with simpler architectures, known interactions, and poor prediction generalizability across populations, such as skin pigmentation [54].

Limited generalizability of genetic prediction across diverse populations

Previous work has assessed prediction accuracy across diverse populations in several traits and diseases for which GWAS summary statistics are available. These assessments are becoming increasingly feasible with the growth and public availability of global biobanks for quantitative traits as well as diversifying priorities from funding agencies [55,56]. As of yet, multi-ethnic work has been slow in most disease areas [57], limiting even the opportunity to assess prediction utility in non-European cohorts. Nonetheless, we have assembled prediction accuracy statistics from several studies using the largest European GWAS to predict several phenotypes in target European and non-European cohorts. For example, multiple schizophrenia studies consistently predicted risk on average 2.2-fold worse in East Asians relative to Europeans, (i.e. μ=0.46, σ=0.06), using summary statistics from a Eurocentric GWAS [15,18] (**Figure S2**), despite the fact that there is no significant genetic heterogeneity in schizophrenia between the two populations [40]. This finding is even more pronounced in African Americans, consistent with higher genetic divergence from Europeans than between Europeans and East Asians [30]. Across several phenotypes with a range of genetic

252 architectures in which empirical evaluations were available, including BMI, educational
253 attainment, height, and schizophrenia, prediction accuracy using European GWAS
254 summary statistics was on average 4.5-fold less accurate in African Americans than in
255 Europeans (i.e. μ=0.22, σ=0.09, **Figure S2**) [15,16,19-22]. By extension, prediction accuracy
256 is expected to be even lower in African Americans with higher than average African
257 ancestry or among populations with greater divergence from Europeans (e.g. some
258 southern African populations). These enormous disparities are not simply
259 methodological issues, as various approaches (e.g. pruning and threshold versus
260 LDPred) and accuracy metrics (R^2 for quantitative traits and various pseudo-R^2 metrics
261 for binary traits) illustrate this consistently poorer performance in populations distinct
262 from the discovery sample across a range of polygenic traits (**Table S4**). We assessed
263 how prediction accuracy decayed across globally diverse populations for 17
264 anthropometric and blood panel traits in the UK Biobank (UKBB) when using European-
265 derived summary statistics (**Methods**); consistent with previous studies, we find that
266 relative to European prediction accuracy, genetic prediction accuracy was far lower in
267 other populations (**Figure 3**, notably 1.6-fold lower in Hispanic/Latino Americans, 1.7-
268 fold lower in South Asians, 2.5-fold lower in East Asians, and 4.9-fold lower in Africans
269 on average).
270
271 **Prioritizing diversity shows early promise for polygenic prediction**
272 Early diversifying GWAS efforts have been especially productive for informing on these
273 questions surrounding risk prediction. Rather than varying the prediction target dataset,
274 some GWAS in diverse populations have increased the scale of non-European

summary statistics and also varied the study dataset in multi-ethnic PRS studies. For example, a BioBank Japan (BBJ) GWAS study (N=158,284) showed that compared to a 2× larger European GWAS (N=322,154), the variance in BMI explained in an independent Japanese cohort with Japanese GWAS summary statistics was on average 1.5-fold greater than with European GWAS summary statistics (R^2=0.154 vs 0.104 at $p < 0.05$, respectively) [23]. Relatedly, an East Asian schizophrenia study (N=22,778 cases and 35,362 controls) showed that compared to an effectively 3× larger European study, prediction accuracy in East Asians was on average 1.3-fold higher than with European summary statistics (liability R^2=0.029 vs 0.022, respectively) [40]. Thus, even when studies in non-European populations are only a fraction the size of the largest European study, they are likely to have disproportionate value for predicting polygenic traits in other individuals of similar ancestry.

Given this background, we performed a systematic evaluation of polygenic prediction accuracy across 17 quantitative anthropometric and blood panel traits and five disease endpoints in British and Japanese individuals [23,58,59] by performing GWAS with the exact same sample sizes in each population. We symmetrically demonstrate that prediction accuracy is consistently higher with GWAS summary statistics from ancestry-matched summary statistics (**Figure 4, Figure S8–12**). Keeping in mind issues of comparability described above, we note that BBJ is a hospital-based disease-ascertained cohort, whereas UKBB is a healthier than average population-based cohort; thus, differences in observed heritability among these cohorts (rather than among populations) due to differences in phenotype precision likely explain lower prediction accuracy from the BBJ

GWAS summary statistics for anthropometric and blood panel traits, but higher prediction accuracy for five ascertained diseases (**Table S5**). Indeed, other East Asian studies have estimated higher heritability for some quantitative traits than BBJ using the same methods, such as for height (h^2 = 0.48±0.04 in Chinese women) [60]. Some statistical fluctuations in the relative differences in prediction accuracy across populations are likely driven by differences in heritability measured in each population and/or trans-ethnic genetic correlation[†] (**Figure S4-7**, **Table S5–8**). These correlation estimates indicate that effect sizes were mostly highly correlated across ancestries, with a few traits that were somewhat lower than excepted (e.g. height and BMI, with ρ_{ge}=0.69 and 0.75, respectively); effect sizes can be different across populations due to natural selection differences, gene-gene interactions, gene-environment interactions, statistical noise, and/or other phenomena [61]. Prediction accuracy was far lower in individuals of African descent in the UK Biobank (**Figure S10 and S13**) using GWAS summary statistics from either European or Japanese ancestry individuals, consistent with reduced prediction accuracy with increasing genetic divergence (**Figures 3 and 4**). These population studies demonstrate the power and utility of increasingly diverse GWAS for prediction, especially in populations of non-European descent.

While many other traits and diseases have been studied in multi-ethnic settings, few have reported comparable metrics of prediction accuracy across populations. Cardiovascular research, for example, has led the charge towards clinical translation of PRS [1]. This enthusiasm is driven by observations that a polygenic burden of LDL-

[†] Trans-ethnic genetic correlation compares the estimated correlation of common variant effect sizes at SNPs common in two populations, here using Popcorn [61].

increasing SNPs can confer monogenic-equivalent risk of cardiovascular disease, with polygenic scores improving clinical models for risk assessment and statin prescription that can reduce coronary heart disease and improve healthcare delivery efficiency [5,6,8]. However, many of these studies have been conducted exclusively in European descent populations, with few studies rigorously evaluating population-level applicability to non-Europeans. Those existing findings indeed demonstrate a large reduction in prediction utility in non-European populations [11], though often with comparisons of odds ratios among arbitrary breakpoints in the risk distribution that make comparisons across studies challenging. To better clarify how polygenic prediction will be deployed in a clinical setting with diverse populations, more systematic and thorough evaluations of the utility of PRS within and across populations for many complex traits are still needed. These evaluations would benefit from rigorous polygenic prediction accuracy evaluations, especially for diverse non-European patients [62-64].

Translational genetic prediction may uniquely exacerbate disparities

Our impetus for raising these statistical issues limiting the generalizability of PRS across population stems from our concern that, while they are legitimately clinically promising for improving health outcomes for many biomedical phenotypes, they may have a larger potential to raise health disparities than other clinical factors for several reasons. The opportunities they provide for improving health outcomes means they inevitably will and should be pursued in the near term, but we urge that a concerted prioritization to make GWAS summary statistics easily accessible for diverse populations and a variety of traits and diseases is imperative, even when they are a fraction the size of the largest

existing European datasets. *Individual* clinical tests, biomarkers, and prescription drug efficacy may vary across populations in their utility, but are fundamentally informed by the same underlying biology [65,66]. Currently, guidelines state that as few as 120 individuals define reference intervals for clinical factors (though often smaller numbers from only one subpopulation are used) and there is no clear definition of who is "normal" [65]. Consequently, reference intervals for biomarkers can sometimes deviate considerably by reported ethnicity [67-69]. Defining ethnicity-specific reference intervals is clearly an important problem that can provide fundamental interpretability gains with implications for some major health benefits (e.g. need for dialysis and development of Type 2 diabetes based on ethnicity-specific serum creatinine and hemoglobin A1C reference intervals, respectively) [68]. Simply put, some biomarkers or clinical tests scale directly with health outcomes independent of ancestry, and many others may have distributional differences by ancestry but are equally valid after centering with respect to a readily collected population reference.

In contrast, PRS are uniformly less useful in understudied populations due to differences in genomic variation and population history [13,14]. No analogous solution of defining ethnicity-specific reference intervals would ameliorate health disparities implications for PRS or fundamentally aid interpretability in non-European populations. Rather, as we and others demonstrate, PRS are unique in that even with multi-ethnic population references, these scores are fundamentally less informative in populations more diverged from GWAS study cohorts.

366 The clinical use and deployment of genetic risk scores needs to be informed by the
367 issues surrounding tests that currently would unequivocally provide much greater
368 benefit to the subset of the world's population which is already on the positive end of
369 healthcare disparities[‡]. Conversely, African descent populations, which already endure
370 many of the largest health disparities globally, are often predicted marginally better, if at
371 all, compared to random (**Figure 4F**). They are therefore least likely to benefit from
372 improvements in precision healthcare delivery from genetic risk scores with existing
373 data due to human population history and study biases. This is a major concern globally
374 and especially in the U.S., which already leads other middle- and high-income countries
375 in both real and perceived healthcare disparities [70]. Thus, we would strongly urge that
376 any discourse on clinical use of polygenic scores include a careful, quantitative
377 assessment of the economic and health disparities impacts on underrepresented
378 populations that might be unintentionally introduced by the use of PRS and raise
379 awareness about how to eliminate these disparities.
380
381 **How do we even the ledger?**
382 What can be done? The single most important step towards parity in PRS accuracy in
383 diverse populations is by vastly increasing the diversity of participants included and
384 analyzed in genetic studies. Regulatory protections against genetic discrimination are
385 necessary to accompany calls for more diverse studies; while some already exist in the
386 U.S., including for health insurance and employment opportunities via the Genetic
387 Information Nondiscrimination Act (GINA), stronger protections in these and other areas

[‡] To maximally benefit all populations, the largest existing GWAS results should be used. Downsampling European GWAS for the sake of parity results in worse predictors for all individuals.

globally will be particularly important for minorities and/or marginalized groups. An equal investment in GWAS across all major ancestries and global populations is the most obvious solution to generate a substrate for equally informative risk scores but is not likely to occur any time soon absent a dramatic priority shift given the current imbalance and stalled diversifying progress over the last five years (**Figure 1, Figure S1**). While it may be challenging or in some cases infeasible to acquire sample sizes large enough for PRS to be equally informative in all populations, some much-needed efforts towards increasing diversity in genomics that support open sharing of GWAS summary data from multiple ancestries are underway. Examples include the *All of Us* Research Program, the Population Architecture using Genomics and Epidemiology (PAGE) Consortium, as well as some disease-focused consortia, such as the T2D-genes and Stanley Global initiatives on the genetics of type II diabetes and psychiatric disorders, respectively[§]. Supporting data resources such as imputation panels, multi-ethnic genotyping arrays, gene expression datasets from genetically diverse individuals, and other tools are necessary to similarly empower these diverse studies to most effectively improve predictive accuracy for all populations. The lack of supporting resources for diverse ancestries creates financial challenges for association studies with limited resources, e.g. raising questions about whether to genotype samples on GWAS arrays that may favor European allele frequencies versus sequence samples, and how dense of an array to choose or how deeply to sequence [71,72].

[§] For these diverse studies in the U.S. and globally, genetically determined ancestry and populations are important to delineate from self-identified race/ethnicity, as only controlling for the former can account for stratification of allele frequencies within a population.

Additional leading global efforts also provide easy unified access linking genetic, clinical record, and national registry data in more homogeneous continental ancestries, such as the UK Biobank, BioBank Japan, China Kadoorie Biobank, and Nordic efforts (e.g. in Danish, Estonian, Finnish, and other integrated biobanks). Notably, some of these biobanks such as UK Biobank have participants with considerable global genetic diversity that enables multi-ethnic comparisons; although minorities from this cohort provide the largest deeply phenotyped GWAS cohorts for several ancestries, these individuals are often excluded in current statistical analyses in favor of single ancestries, large sample sizes, and the simplicity afforded by genetic homogeneity. These considerations notwithstanding, there are critical needs and challenges for expanding the scale of genetic studies of heritable traits in diverse populations; this is especially apparent in Africa where humans originated and retain the most genetic diversity, as Africans are understudied but disproportionately informative for genetic analyses and evolutionary history [27,73]. The most notable investment here comes from the Human Heredity and Health in Africa (H3Africa) Initiative, increasing genomics research capacity in Africa through more than $216 million in funding from the NIH (USA) and Wellcome Trust (UK) for genetics research led by African investigators [56,74]. The increasing interest and scale of genetic studies in low- and middle-income countries (LMICs) raises ethical and logistical considerations about data generation, access, sharing, security, and analysis, as well as clinical implementation to ensure these advances do not only benefit high-income countries. Frameworks such as the H3ABioNet, a pan-African bioinformatics network designed to build capacity to enable

H3Africa researchers to analyze their data in Africa, provide cost-effective examples for training local scientists in LMICs [75].

The prerequisite data for dramatically increasing diversity also hypothetically exist in several large-scale publicly funded datasets such as the Million Veterans Project and Trans-Omics for Precision Medicine (TOPMed), but with problematic data access issues in which even summary data from GWAS within and across populations are not publicly shared. Existing GWAS consortia also need to carefully consider the granularity of summary statistics they release, as finer scale continental ancestries and phenotypes in large, multi-ethnic projects enable ancestry-matched analyses not possible with a single set of summary statistics. While there is an understandable patient privacy balance to strike when sharing individual-level data, GWAS summary statistics from all publicly funded and as many privately funded projects as possible should be made easily and publicly accessible to improve global health outcomes. Efforts to unify phenotype definitions, normalization approaches, and GWAS methods among studies will also improve comparability.

To enable progress towards parity, it will be critical that open data sharing standards be adopted for all ancestries and for genetic studies of all sample sizes, not just the largest European results. Locally appropriate and secure genetic data sharing techniques as well as equitable technology availability will need to be adopted widely in Asia and Africa as they are in Europe and North America, to ensure that maximum value is achieved from existing and ongoing efforts that are being developed to help counter the

current imbalance. Simultaneously, ethical considerations require that research capacity is increased in LMICs with simultaneous growth of diverse population studies to balance the benefits of these studies to scientists and patients globally versus locally to ensure that everyone benefits. Methodological improvements that better define risk scores by accounting for population allele frequency, LD, and/or admixture differences appropriately are underway and may help considerably but will not by themselves bring equality. All of these efforts are important and should be prioritized not just for risk prediction but more generally to maximize the use and applicability of genetics to inform on the biology of disease. Given the acute recent attention on clinical use of PRS, we believe it is paramount to recognize their potential to improve health outcomes for all individuals and many complex diseases. Simultaneously, we as a field must address the disparity in utility in an ethically thoughtful and scientifically rigorous fashion, lest we inadvertently enable genetic technologies to contribute to, rather than reduce, existing health disparities.

Acknowledgments

We thank Amit Khera for helpful discussions. We also thank Michiaki Kubo, Yoshinori Murakami, Masato Akiyama, and Kazuyoshi Ishigaki for their support in the BioBank Japan Project analysis. We are grateful to Steven Gazal for his help in calculating LD scores. This work was supported by funding from the National Institutes of Health (K99MH117229 to A.R.M.). UK Biobank analyses were conducted via application 31063. The BioBank Japan Project was supported by the Tailor-Made Medical Treatment Program of the Ministry of Education, Culture, Sports, Science, and

Technology (MEXT) and the Japan Agency for Medical Research and Development (AMED). M.K. was supported by a Nakajima Foundation Fellowship and the Masason Foundation.

Figures

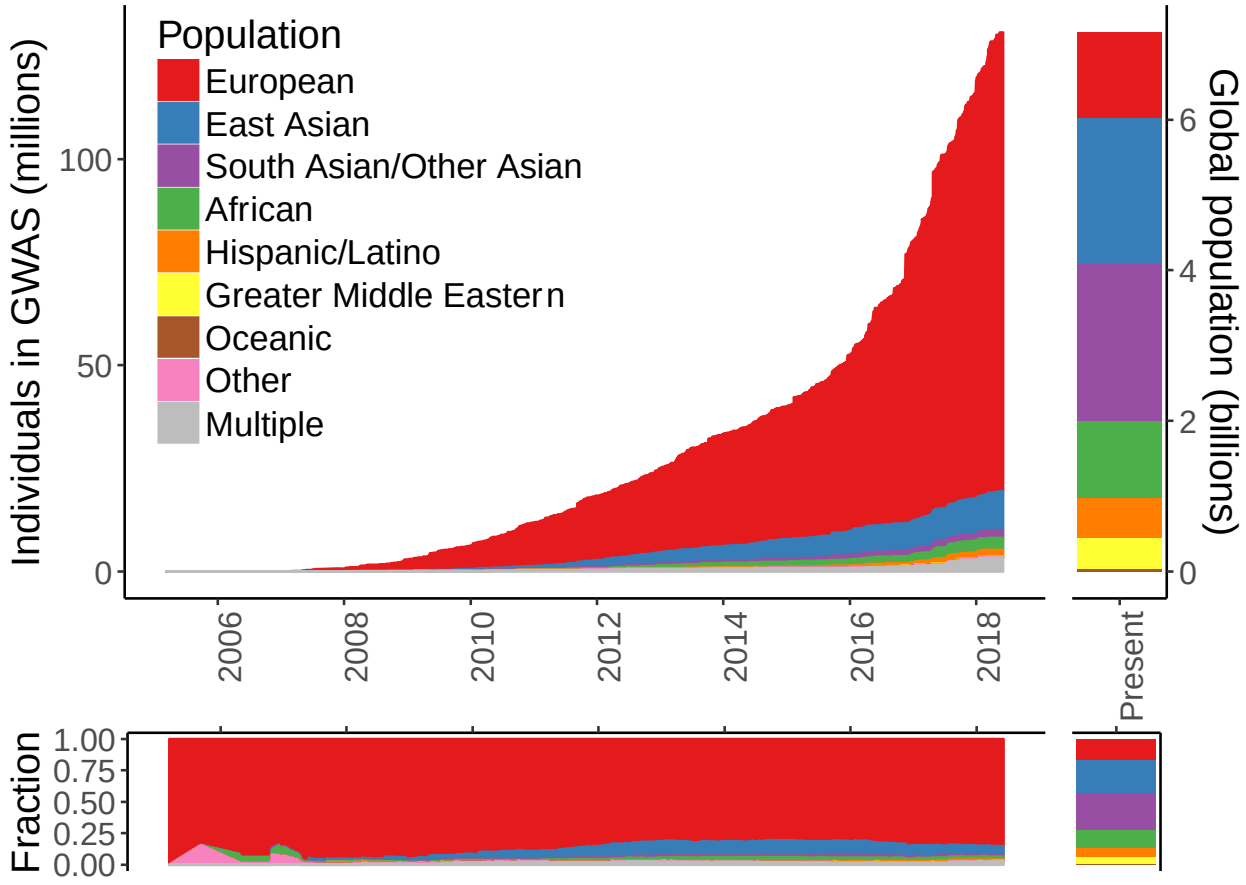

Figure 1 – Ancestry of GWAS participants over time compared to the global population. Cumulative data as reported by the GWAS catalog [27]. A notable caveat is that because some cohorts are included in numerous studies, some individuals are represented multiple times. This bias in multiple counting is especially likely for publicly available cohorts, which are more likely to be of European or East Asian descent. Individuals whose ancestry is "not reported" are not shown.

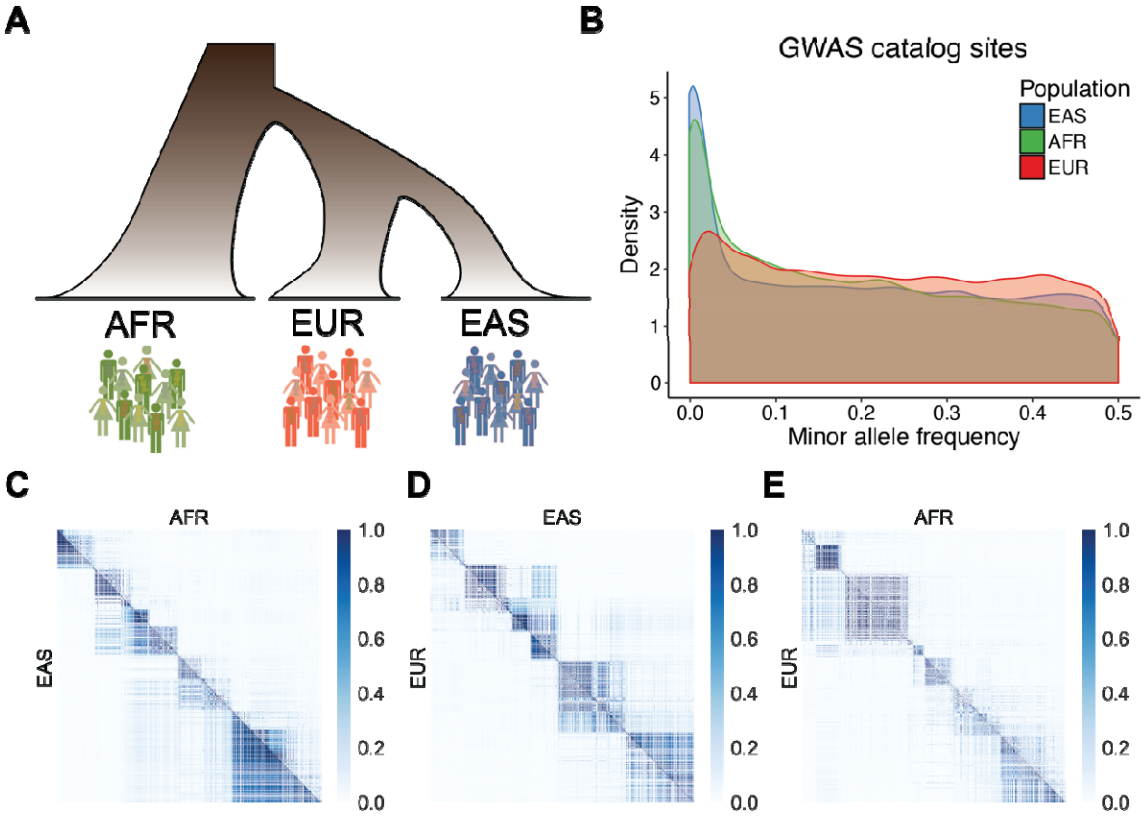

Figure 2 – Demographic relationships, allele frequency differences, and local LD patterns between population pairs. Data analyzed from 1000 Genomes, in which population labels are: AFR = continental African, EUR = European, and EAS = East Asian. A) Cartoon relationships among AFR, EUR, and EAS populations. B) Allele frequency distributions in AFR, EUR, and EAS populations of variants from the GWAS catalog. C-E) Color axis shows LD scale (r^2). LD comparisons between pairs of populations show the same region of the genome for each comparison (representative region is chr1, 51572kb-52857kb) among pairs of SNPs polymorphic in both populations, illustrating that different SNPs are polymorphic across some population pairs, and that these SNPs have variable LD patterns across populations.

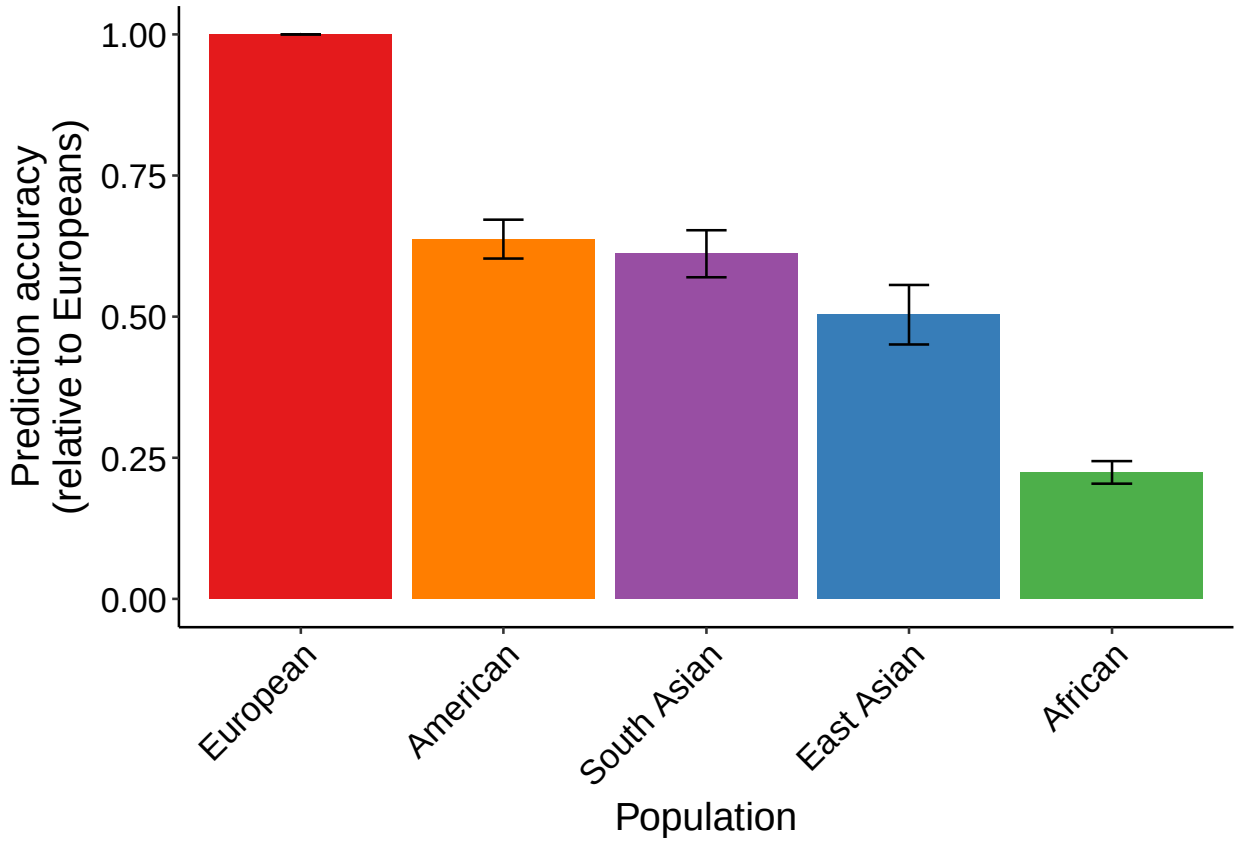

Figure 3 – Prediction accuracy relative to European ancestry individuals across 17 quantitative traits and 5 continental populations in UKBB. All phenotypes shown here are quantitative anthropometric and blood panel traits, as described in **Table S1**, including sample sizes. Bars show mean values, and error bars show standard errors of the means.

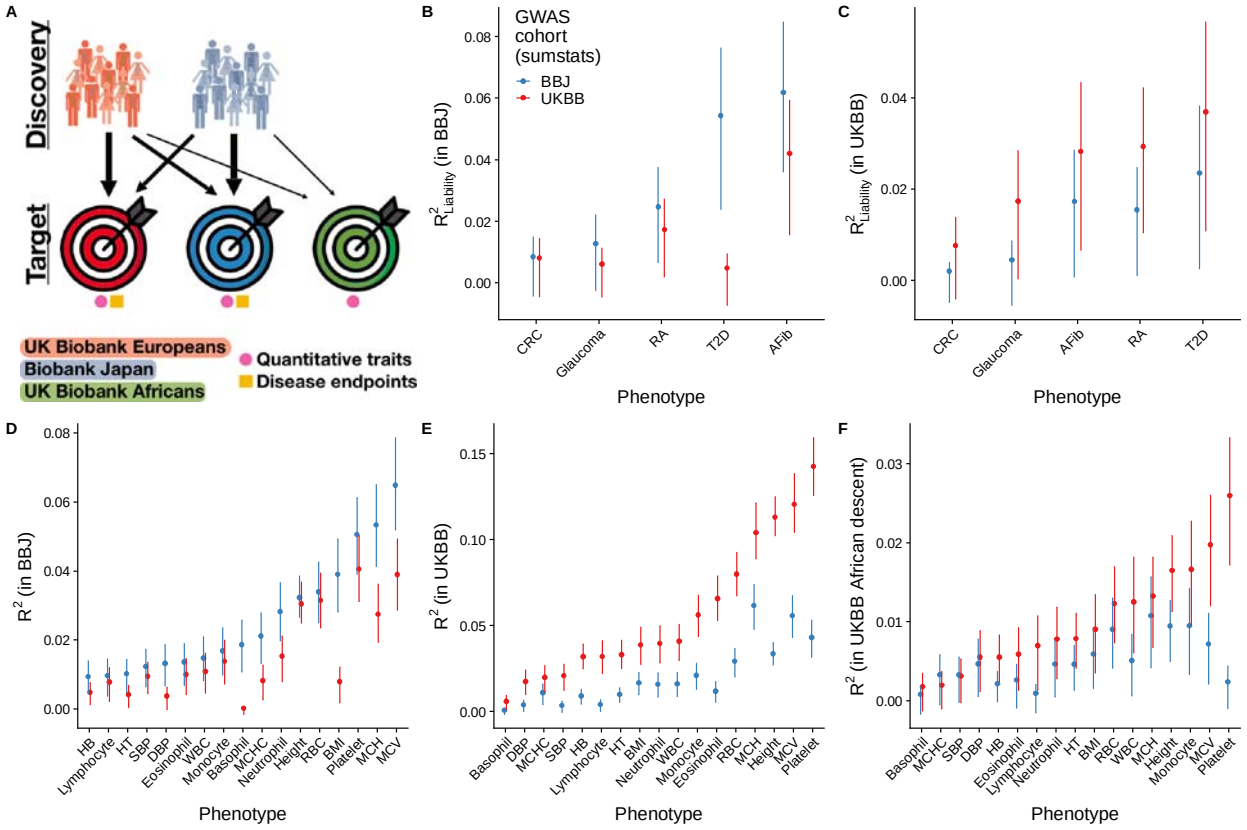

Figure 4 – Polygenic risk prediction accuracy in Japanese, British, and African descent individuals using independent GWAS of equal sample sizes in the BioBank Japan and UK Biobank. All target prediction cohorts are withheld from the GWAS and thus independent. Sample sizes in each GWAS are identical by design between BBJ and UKBB (**Table S1 and S3**). To optimize signal to noise, each point shows the maximum R^2 (i.e. best predictor) across five p-value thresholds. R^2 values for all p-value thresholds are shown in **Figures S8-S12**. Prediction accuracy tends to be higher in the UK Biobank for quantitative traits than in BioBank Japan and vice versa for disease endpoints, likely because of concomitant phenotype precision and

consequently observed heritability for these classes of traits (**Table S5-S7**)[**].

Abbreviations are as follows: AFib = atrial fibrillation, BMI = body mass index, CRC = colorectal cancer, DBP = diastolic blood pressure, Hb = hemoglobin, Ht = Hematocrit, MCH = mean corpuscular hemoglobin, MCHC = mean corpuscular hemoglobin concentration, MCV = mean corpuscular volume, RA = rheumatoid arthritis, RBC = red blood cell count, SBP = systolic blood pressure, T2D = type 2 diabetes, WBC = white blood cell count. A) Explanatory diagram showing the different discovery and target cohorts/populations, and quantitative traits versus disease endpoints. B) Genetic prediction accuracy for five diseases in Japanese individuals using summary statistics from GWAS of independent BioBank Japan versus UK Biobank samples. C) Genetic prediction accuracy for the same five diseases in independent British individuals using summary statistics from GWAS of independent BioBank Japan versus the UK Biobank samples. D) Genetic prediction accuracy for 17 anthropometric and blood panel traits in Japanese individuals using summary statistics from GWAS of independent BioBank Japan versus UK Biobank samples. E) Genetic prediction accuracy for the same 17 anthropometric and blood panel traits in independent British individuals using summary statistics from GWAS of independent BioBank Japan versus the UK Biobank samples. F) Genetic prediction accuracy for 17 anthropometric and blood panel traits in African descent individuals in the UK Biobank using summary statistics from GWAS of independent BioBank Japan versus UK Biobank samples.

[**] Thalassemia and sickle cell disease are unlikely to explain a significant fraction of these prediction differences, as few individuals have been diagnosed with these disorders via ICD-10 codes (**Table S11**).

References

1. Knowles, J. W. & Ashley, E. A. Cardiovascular disease: The rise of the genetic risk score. *PLoS Med* **15,** e1002546–7 (2018).
2. Maas, P. *et al.* Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States. *JAMA Oncol* **2,** 1295–8 (2016).
3. Mavaddat, N. *et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *The American Journal of Human Genetics* 1–42 (2018). doi:10.1016/j.ajhg.2018.11.002
4. Schumacher, F. R. *et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat Genet* 1–13 (2018). doi:10.1038/s41588-018-0142-8
5. Khera, A. V. *et al.* Genome-wide polygenic score to identify a monogenic risk-equivalent for coronary disease. (2017). doi:10.1101/218388
6. Kullo, I. J. *et al.* Incorporating a Genetic Risk Score Into Coronary Heart Disease Risk Estimates: Effect on Low-Density Lipoprotein Cholesterol Levels (the MI-GENES Clinical Trial). *Circulation* **133,** 1181–1188 (2016).
7. Paquette, M. *et al.* Polygenic risk score predicts prevalence of cardiovascular disease in patients with familial hypercholesterolemia. *Journal of Clinical Lipidology* **11,** 725–732.e5 (2017).
8. Natarajan, P. *et al.* Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention SettingClinical Perspective. *Circulation* **135,** 2091–2101 (2017).
9. Tikkanen, E., Havulinna, A. S., Palotie, A., Salomaa, V. & Ripatti, S. Genetic risk prediction and a 2-stage risk screening strategy for coronary heart disease. *Arterioscler. Thromb. Vasc. Biol.* **33,** 2261–2266 (2013).
10. Frieser, M. J., Wilson, S. & Vrieze, S. Behavioral Impact of Return of Genetic Test Results for Complex Disease: Systematic Review and Meta-Analysis. *Health Psychol* **37,** 1134–1144 (2018).
11. Khera, A. V. *et al.* Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. *N Engl J Med* **375,** 2349–2358 (2016).
12. Khera, A. V. & Kathiresan, S. Genetics of coronary artery disease: discovery, biology and clinical translation. *Nat Rev Genet* **18,** 331–344 (2017).
13. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100,** 635–649 (2017).
14. Scutari, M., Mackay, I. & Balding, D. Using Genetic Distance to Infer the Accuracy of Genomic Prediction. *PLoS Genet* **12,** e1006288 (2016).
15. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *The American Journal of Human Genetics* **97,** 576–592 (2015).
16. Ware, E. B. *et al.* Heterogeneity in polygenic scores for common human traits. 1–13 (2017). doi:10.1101/106062

17. Curtis, D. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *bioRxiv* (2018). doi:https://doi.org/10.1101/287136
18. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–427 (2014).
19. Belsky, D. W. *et al.* Development and Evaluation of a Genetic Risk Score for Obesity. *Biodemography and Social Biology* **59,** 85–100 (2013).
20. Domingue, B. W., Belsky, D. W., Conley, D., Harris, K. M. & Boardman, J. D. Polygenic Influence on Educational Attainment. *AERA Open* **1,** 233285841559997–13 (2015).
21. Wedow, R. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat Genet* **92,** 109 (2018).
22. Vassos, E. *et al.* An Examination of Polygenic Score Risk Prediction in Individuals With First-Episode Psychosis. *Biological Psychiatry* **81,** 470–477 (2017).
23. Akiyama, M. *et al.* Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nat Genet* **49,** 1458–1467 (2017).
24. Li, Z. *et al.* Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nature Publishing Group* **49,** 1576–1583 (2017).
25. Need, A. C. & Goldstein, D. B. Next generation disparities in human genomics: concerns and remedies. *Trends in Genetics* **25,** 489–494 (2009).
26. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538,** 161–164 (2016).
27. Morales, J. *et al.* A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. 1–10 (2018). doi:10.1186/s13059-018-1396-2
28. Rosenberg, N. A. *et al.* Genome-wide association studies in diverse populations. *Nat Rev Genet* **11,** 356–366 (2010).
29. Sham, P. C., Cherny, S. S., Purcell, S. & Hewitt, J. K. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *The American Journal of Human Genetics* **66,** 1616–1630 (2000).
30. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
31. Consortium, T. S. 2. D. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature* **506,** 97–101 (2014).
32. Estrada, K. *et al.* Association of a Low-Frequency Variant in HNF1A With Type 2 Diabetes in a Latino Population. *JAMA* **311,** 2305–2314 (2014).
33. Haiman, C. A. *et al.* Genome-wide association study of prostate cancer in men of African ancestry identifies a susceptibility locus at 17q21. *Nat Genet* **43,** 570–573 (2011).
34. Genovese, G. *et al.* Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science* **329,** 841–845 (2010).
35. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* **47,** 979–986 (2015).

36. Carlson, C. S. *et al.* Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study. *PLoS Biol* **11,** e1001661 (2013).
37. Easton, D. F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447,** 1087–1093 (2007).
38. Mahajan, A. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* **46,** 234–244 (2014).
39. Waters, K. M. *et al.* Consistent Association of Type 2 Diabetes Risk Variants Found in Europeans in Diverse Racial and Ethnic Groups. *PLoS Genet* **6,** e1001078 (2010).
40. Lam, M. *et al.* Comparative genetic architectures of schizophrenia in East Asian and European populations. 1–41 (2018). doi:10.1101/445874
41. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48,** 1279–1283 (2016).
42. Huang, H. *et al.* Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547,** 173–178 (2017).
43. Berg, J. J. *et al.* Reduced signal for polygenic adaptation of height in UK Biobank. 1–44 (2018). doi:10.1101/354951
44. Sohail, M. *et al.* Signals of polygenic adaptation on height have been overestimated due to uncorrected population structure in genome-wide association studies. 1–44 (2018). doi:10.1101/355057
45. Kerminen, S. *et al.* Geographic variation and bias in polygenic scores of complex diseases and traits in Finland:. (2018). doi:10.1101/485441
46. Ge, T., Chen, C.-Y., Neale, B. M., Sabuncu, M. R. & Smoller, J. W. Phenome-wide heritability analysis of the UK Biobank. *PLoS Genet* **13,** e1006711–21 (2017).
47. Inouye, M. *et al.* Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. *J. Am. Coll. Cardiol.* **72,** 1883–1893 (2018).
48. Pourcain, B. S. *et al.* ASD and schizophrenia show distinct developmental profiles in common genetic overlap with population-based social communication difficulties. *Molecular Psychiatry* 1–8 (2016). doi:10.1038/mp.2016.198
49. Novembre, J. & Barton, N. H. Tread Lightly Interpreting Polygenic Tests of Selection. *Genetics* **208,** 1351–1355 (2018).
50. Henn, B. M., Botigué, L. R., Bustamante, C. D., Clark, A. G. & Gravel, S. Estimating the mutation load in human genomes. *Nat Rev Genet* **16,** 333–343 (2015).
51. Li, D., Zhao, H. & Gelernter, J. Strong protective effect of the aldehyde dehydrogenase gene (ALDH2) 504lys (*2) allele against alcoholism and alcohol-induced medical diseases in Asians. *Human Genetics* **131,** 725–737 (2011).
52. Zhu, Z. *et al.* Dominance Genetic Variation Contributes Little to the Missing Heritability for Human Complex Traits. *The American Journal of Human Genetics* **96,** 377–385 (2015).
53. Paré, G., Mao, S. & Deng, W. Q. A machine-learning heuristic to improve gene score prediction of polygenic traits. *Sci Rep* **7,** 12665 (2017).

673 54. Martin, A. R. *et al.* An Unexpectedly Complex Architecture for Skin Pigmentation
674 in Africans. (2017). doi:10.1101/200139
675 55. Hindorff, L. A. *et al.* Prioritizing diversity in human genomics research. *Nature*
676 *Publishing Group* 1–11 (2017). doi:10.1038/nrg.2017.89
677 56. H3Africa Consortium *et al.* Research capacity. Enabling the genomic revolution in
678 Africa. *Science* **344,** 1346–1348 (2014).
679 57. Duncan, L. E. *et al.* Largest GWAS of PTSD (N=20□070) yields genetic overlap
680 with schizophrenia and sex differences in heritability. *Molecular Psychiatry* (2017).
681 doi:10.1038/mp.2017.77
682 58. Howrigan, D. Details and Considerations of the Uk Biobank GWAS.
683 *http://www.nealelab.is/blog/2017/9/11/details-and-considerations-of-the-uk-*
684 *biobank-gwas* (2017). Available at: (Accessed: 9 November 2017)
685 59. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population
686 links cell types to complex human diseases. *Nat Genet* 1–16 (2018).
687 doi:10.1038/s41588-018-0047-6
688 60. Liu, S. *et al.* Genomic Analyses from Non-invasive Prenatal Testing Reveal
689 Genetic Associations, Patterns of Viral Infections, and Chinese Population
690 History. *Cell* **175,** 347–359.e14 (2018).
691 61. Brown, B. C., Ye, C. J., Price, A. L. & Zaitlen, N. Transethnic Genetic-Correlation
692 Estimates from Summary Statistics. *The American Journal of Human Genetics*
693 **99,** 76–88 (2016).
694 62. Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nature*
695 *Publishing Group* **14,** 507–515 (2013).
696 63. Wray, N. R. *et al.* Research Review: Polygenic methods and their application to
697 psychiatric traits. *J Child Psychol Psychiatr* **55,** 1068–1087 (2014).
698 64. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of
699 polygenic risk scores. *Nat Rev Genet* 1–10 (2018). doi:10.1038/s41576-018-
700 0018-x
701 65. Manrai, A. K., Patel, C. J. & Ioannidis, J. P. A. In the Era of Precision Medicine
702 and Big Data, Who Is Normal? *JAMA* **319,** 1981–1982 (2018).
703 66. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets
704 through human genetics. *Nature Publishing Group* **12,** 581–594 (2013).
705 67. Carroll, M. D., Kit, B. K., Lacher, D. A., Shero, S. T. & Mussolino, M. E. Trends in
706 lipids and lipoproteins in US adults, 1988-2010. *JAMA* **308,** 1545–1554 (2012).
707 68. Rappoport, N. *et al.* Creating ethnicity-specific reference intervals for lab tests
708 from EHR data. *bioRxiv* (2017). doi:10.1101/213892
709 69. Lim, E., Miyamura, J. & Chen, J. J. Racial/Ethnic-Specific Reference Intervals for
710 Common Laboratory Tests: A Comparison among Asians, Blacks, Hispanics, and
711 White. *Hawaii J Med Public Health* **74,** 302–310 (2015).
712 70. Hero, J. O., Zaslavsky, A. M. & Blendon, R. J. The United States Leads Other
713 Nations In Differences By Income In Perceptions Of Health And Health Care.
714 *Health Aff* **36,** 1032–1040 (2017).
715 71. Gilly, A. *et al.* Very low depth whole genome sequencing in complex trait
716 association studies. 1–24 (2018). doi:10.1101/169789
717 72. Pasaniuc, B. *et al.* Extremely low-coverage sequencing and imputation increases
718 power for genome-wide association studies. *Nat Genet* **44,** 631–635 (2012).

73. Martin, A. R., Teferra, S., Möller, M., Hoal, E. G. & Daly, M. J. The critical needs and challenges for genetic architecture studies in Africa. *Current Opinion in Genetics & Development* **53,** 113–120 (2018).
74. Coles, E. & Mensah, G. A. Geography of Genetics and Genomics Research Funding in Africa. *Global Heart* **12,** 173–176 (2017).
75. Mulder, N. J. *et al.* Development of Bioinformatics Infrastructure for Genomics Research. *Global Heart* **12,** 91–98 (2017).