Genomics of disease risk in globally diverse populations

Deepti Gurdasani[1,2], *Inês Barroso*[1,3], *Eleftheria Zeggini*[4] *and Manjinder S. Sandhu*[2]*

Abstract | Risk of disease is multifactorial and can be shaped by socio-economic, demographic, cultural, environmental and genetic factors. Our understanding of the genetic determinants of disease risk has greatly advanced with the advent of genome-wide association studies (GWAS), which detect associations between genetic variants and complex traits or diseases by comparing populations of cases and controls. However, much of this discovery has occurred through GWAS of individuals of European ancestry, with limited representation of other populations, including from Africa, The Americas, Asia and Oceania. Population demography, genetic drift and adaptation to environments over thousands of years have led globally to the diversification of populations. This global genomic diversity can provide new opportunities for discovery and translation into therapies, as well as a better understanding of population disease risk. Large-scale multi-ethnic and representative biobanks and population health resources provide unprecedented opportunities to understand the genetic determinants of disease on a global scale.

Genome-wide association studies
Hypothesis-free studies of association between genetic variants and quantitative traits or diseases; typically, associations are examined across the whole genome using genotype array or sequencing approaches.

[1]*Department of Human Genetics, Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK.*

[2]*Department of Medicine, University of Cambridge, Cambridge, UK.*

[3]*MRC Epidemiology Unit, University of Cambridge, Cambridge, UK.*

[4]*Institute of Translational Genomics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany.*

**e-mail: mss31@cam.ac.uk*

https://doi.org/10.1038/s41576-019-0144-0

Global differences in the prevalence and distribution of diseases and their risk factors are a complex phenomenon determined by environmental, social, demographic, cultural and genetic factors. Genetic variation at the population level is itself shaped by population history, demography, regional environments and adaptive evolution. Understanding global genetic diversity and its impact on human health and disease has the potential to provide additional insights into the biological mechanisms underlying disease risk and can help quantify the impact of the interplay between genetic and environmental variation on population-level disease risk[1]. As such, conducting genomic research in diverse populations across the globe can inform therapeutic development, public health and precision medicine initiatives as well as facilitate global equity in the benefits of genomics[1]. Here, we define diverse populations as genetically heterogeneous populations that include ethno-linguistically and geographically diverse individuals.

Studies assessing genetic diversity among global populations[2–4] have laid the framework for understanding the impact of genetic variation on disease risk in a global context. However, although the proportion of individuals of non-European ancestry represented in genome-wide association studies (GWAS) has increased over the past 5 years, the number and scale of GWAS in European populations still far exceed those in non-Europeans[5,6] (FIG. 1). Despite the number of individuals of African and Hispanic or Latin American ancestry in GWAS being smaller, evidence suggests that these individuals contribute disproportionately to genome-wide significant associations and thus may have a greater impact on discovery compared with European or Asian populations[5,7] (FIG. 1). This observation is consistent with the higher level of genetic variation among African populations relative to European or Asian populations[2], which suggests greater opportunities for discovery per individual among populations with African ancestry relative to studies of Europeans or Asians[7].

The recent development of larger and more globally diverse whole-genome sequence resources and imputation reference panels is greatly improving our understanding of genetic susceptibility to disease by increasing the power of GWAS based on single-nucleotide polymorphism (SNP) arrays[8]. Complemented by the development of large whole-genome and whole-exome sequencing resources (for example, gnomAD[9]), this approach is yielding a better understanding of population differences in the distribution of common (minor allele frequency (MAF) >5%) and rare (MAF <5%) genetic variation, deleterious mutations and their association with disease risk[9].

In this Review, we first consider our understanding of genetic determinants of disease risk among global populations and the extent to which these are likely to be shared between or specific to populations. We then discuss the value of examining diverse populations to better understand genetic contributors to disease risk and variation in traits, including how to leverage this diversity to increase power for discovery. We conclude with considering the implications of these research approaches for

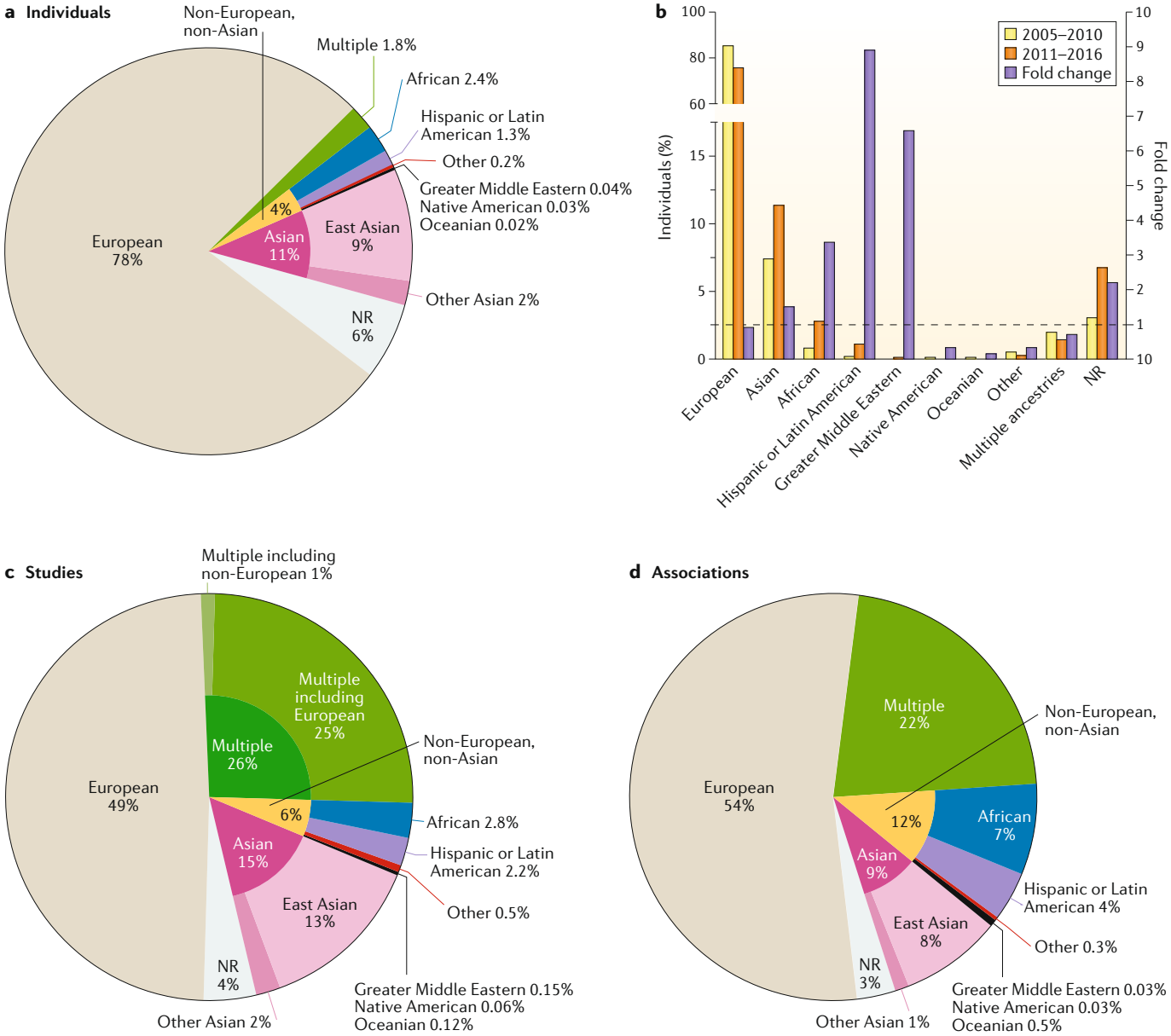


Fig. 1 | **Representation of different ethnic groups in genome-wide association studies.** This figure summarizes the distribution of ancestry categories (in percentages) of individuals (n = 110,291,046; part **a**), individuals over time (n = 110,291,046; part **b**), studies (n = 4,655; part **c**) and associations (n = 60,970; part **d**). The largest category in all panels is European (grey). At the level of individuals (part **a**), the largest non-European category is Asian (bright pink), with East Asian (light pink) accounting for the majority. The non-European, non-Asian category (yellow) comprises 4% of individuals, and there are 6% (white) of samples for which an ancestry category could not be specified (NR). Part **b** displays the distribution of individuals (in percentages) included in the 915 studies published between 2005 and 2010 compared to the distribution of individuals included in the 2,905 studies published between 2011 and 2016. Part **d** demonstrates the disproportionate contribution of associations from the African (blue) and Hispanic or Latin American (purple) categories, when compared to the percentages of individuals (part **a**, blue and purple, respectively) and studies (part **c**, blue and purple, respectively). Reproduced from REF.[7], CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

medical genetics and future directions to advance our understanding of global disease risk.

Shared genetic susceptibility to disease

The extent to which underlying genetic risk factors for disease are similar or shared among different populations is not fully understood. To some extent, this can be inferred from the genetic architecture of disease. In other words, are most genetic variants that confer risk common and therefore likely to be shared among populations, or are these variants rare and specific to a given population? An empirical measure that can be used to assess shared genetic risk is reproducibility, that is, the extent to which variants associated with disease in one population are also observed to be associated with disease in another population. Next, we examine our understanding of the genetic architecture of disease among populations and then review our understanding of reproducibility as a marker of shared genetic risk factors underlying disease among populations.

Imputation
Statistical inference of unobserved genotypes in individuals based on a collection of observed haplotypes among another set of individuals (usually referred to as the reference panel).

The genetic architecture of polygenic traits. The polygenic model of complex traits and diseases suggests that their genetic variance is a composite of small effects of multiple variants spread across the allele frequency spectrum[10]. Although early GWAS explained only a small proportion of genetic variance, subsequent studies showed that increasing statistical power and the subsequent discovery of additional variants with weak effects could explain a substantial proportion of the genetic variance of a given trait or disease[11,12]. Less common or rare variants may also contribute to complex diseases or traits[13]. Increasing sample sizes and better imputation of rare variants have resulted in genetic differences explaining more variation in traits at a population level. For example, the variation in height explained by additive genetic effects increases substantially when common variation[11] and rare variation are included as explanatory variables[10,14]. Family-based designs can also be efficacious in identifying rare genetic variants with large effect sizes associated with monogenic or Mendelian diseases[15–17]. However, the much greater statistical power required and the challenges of reliably capturing less common and rare variants in relevant populations have largely limited our understanding of genetic determinants of most disease traits to common genetic variation (MAF >5%). Equally, statistical models and empirical analyses suggest that larger-scale GWAS for disease traits predominantly identify common variants with weak effects (OR ~1.05–1.2) that are shared across populations[18].

Evolutionary theory suggests that common variants are fairly old, with many of these mutations having occurred >100,000 years ago, before human migration out of Africa[13]. With much of the common variation predating the divergence of modern populations, these variants and the genetic risk they confer are likely to be shared across populations. While the full extent of sharing genetic risk factors for disease among populations is unclear, consistency (similar direction and effect size) and reproducibility of GWAS association signals across different populations can provide broad insights into the sharing of risk loci.

Reproducibility as a marker of shared genetic susceptibility to disease. Many loci discovered by GWAS are shared across diverse populations and may be readily reproducible or transferable among populations. For example, a multi-ethnic case–control study, which included 6,142 cases and 7,403 controls, analysed 19 common genetic risk markers validated for type 2 diabetes mellitus (T2DM) in European populations. This study showed broadly consistent direction of effects across ethnic groups, with the majority of these variants nominally significant in their association with diabetes risk across ethnic groups[19]. Analyses of 16,235 multi-ethnic diabetes cases and 46,122 controls from the Population Architecture using Genomics and Epidemiology (PAGE) consortium recapitulated these findings, showing broad consistency in the direction of effect across different ethnic groups[20].

The DIAGRAM study, which included 26,488 cases of T2DM and 83,964 controls, corroborated these findings, showing statistically significant enrichment for directionally consistent effects across multiple ethnic populations[21] and high correlation of effect estimates across populations. Where studies have shown poorer reproducibility across populations, differences in linkage disequilibrium (LD) have been found to explain why many associations do not replicate directly across populations[18]. This phenomenon can occur when different observed variants tag the same causal variant in different populations, which can lead to the appearance of non-reproducibility when examining the variant (rather than locus). Limited reproducibility for traits can also reflect other factors, including false association signals in discovery studies (or false-negatives in other, less-powered studies), sparse data among diverse ethnic populations, differences in allele frequency, heterogeneity of effect due to gene–gene or gene–environment interactions, or allelic heterogeneity. Understanding the impact of these factors on reproducibility, and on susceptibility to disease, will require a comprehensive understanding of genetic diversity among populations, as well as large-scale and well-powered studies undertaken among genetically diverse populations.

Genomic diversity among populations

Although the vast majority of underlying genetic risk factors for disease are likely shared among populations, genomic diversity among populations can provide new opportunities for discovery. Where genetic risk factors differ among populations (for example, in the case of population-specific variants or because of gene–gene or gene–environment interactions), studying diverse populations can help understand differences in susceptibility, which may not be apparent in studies of more homogeneous populations.

Genetic diversity varies globally among populations. Genetic variation in populations arises from new mutations occurring in each generation, random changes in allele frequencies due to genetic drift and non-random changes in allele frequencies owing to differences in fitness levels conferred by different alleles in the presence of certain environments (selection). The relative contribution and impact of the above factors on genetic variation in a population over time can depend on the demographic history of populations. For example, a population bottleneck (that is, lower effective population size) can lead to rapid changes in allele frequencies of background variation as a result of increased genetic drift acting on the population (FIG. 2). Similarly, migration can expose the human genome to different environments, leading to adaptive changes and regional differentiation that can differentially influence genetic disease risk. That is, natural selection may favour certain alleles in populations exposed to specific environments, leading to an increase in allele frequencies that may influence disease risk. Population expansions can lead to increased variation within populations due to new mutations occurring within offspring at each generation. The genetic diversity observed in populations globally today is a result of these complex forces that have shaped the genetic structure of populations over tens of thousands of years.

Minor allele frequency
The frequency of the less common allele at a site of genetic variation across a sample of individuals or a population.

Genetic variance
The contribution of genetic variation among individuals to phenotypic variation.

Linkage disequilibrium
The non-random association of alleles at loci along the genome in a given population.

Heterogeneity of effect
Statistically significant differences in effect size observed for associations between genetic variants and traits or disease among different studies or populations.

Allelic heterogeneity
The phenomenon whereby multiple causal variants within a given gene can be associated with the same trait or disease.

Genetic drift
A process by which frequencies of alleles in a given population change over time due to random sampling of individuals who may reproduce at every generation.

Selection
A process in which environmental or genetic influences determine which types of organism thrive better than others. Regarded as a factor in evolution.

Population bottleneck
An event that drastically reduces the size of a population. Such events can greatly reduce the genetic diversity of a population and make the population more susceptible to the influence of genetic drift.

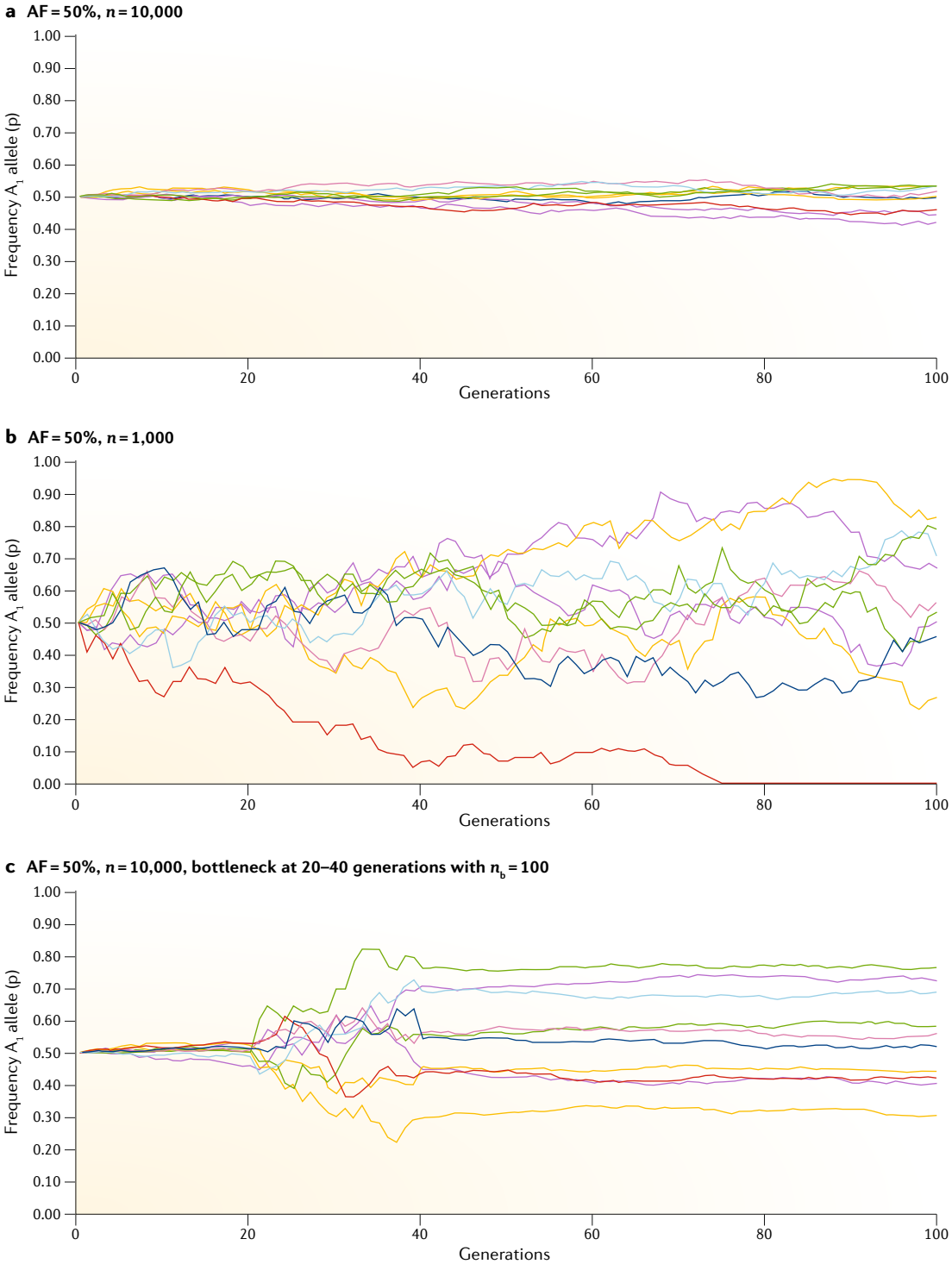


Fig. 2 | **Genetic drift and changes in allele frequency as a function of population size.** Simulations examining the changes in allele frequency (AF) of a given allele over time as a function of population size. Small population sizes and population bottlenecks can lead to rapid changes in AF over time. Founder populations can show markedly different AFs relative to their ancestral population and can therefore provide opportunities for the study of variants that were rare in the ancestral population but have increased to higher frequencies among these groups. **a** | 10 simulations on an allele with a frequency of 50% in a given population of size 10,000 over 100 generations. Frequencies are seen to remain stable over time. **b** | 10 simulations on an allele with a frequency of 50% in a given population of size 1,000 over 100 generations. Relative to part **a**, marked changes in allele frequencies are seen over time with a founding population size of 1,000, suggesting that low founder population sizes can be associated with increased genetic drift over time. **c** | Simulated changes in AF on an allele with a frequency of 50% and a population size of 10,000, which then underwent a population bottleneck to a size of 100 between 20 and 40 generations. The observed AFs of the variant are substantially changed following the bottleneck in simulations. Simulations carried out using Web PopGen simulator. *n* and n_b refer to the population size for the models.

It is well known that genetic variation is greater in African populations relative to non-African populations[2,3]. This is expected given that all modern humans arose from an ancestral African population. Populations that migrated out of Africa were subject to a prolonged population bottleneck, leading to lower genetic variation among these founder individuals and lower effective population sizes[22]. Although subsequent population expansions increased variation within these populations as a result of new mutations, the amount of variation among modern non-African populations remains lower than that of modern African populations.

In the following subsections, we discuss different sources of genetic diversity among populations and how these can influence genetic variation and disease risk. Here, the term 'genetic diversity' refers to structural differences (genetic variation and differences in allele frequency), differences in genetic effects on traits (heterogeneity of effect) as well as epigenetic differences among populations.

Heterogeneity in allele frequency. Differentiation in allele frequencies following divergence between African populations and populations that migrated out of Africa has been shaped by a combination of random drift and selection due to adaptive forces. Population bottlenecks, as undergone by modern non-African populations during out-of-Africa migrations, increase the impact of random drift, leading to genetic variants more rapidly drifting up or down in frequencies, relative to the ancestral population. This substantial genetic drift has resulted in non-African populations exhibiting allele frequencies with greater divergence (on average) from Africans[23]. This model of population demography is recapitulated by several studies of GWAS associations, which have shown marked heterogeneity in allele frequencies of risk variants for diseases among populations[24,25]. Differences in allele frequencies among populations have also been shaped by adaptive forces. An example of this is the sickle cell variant, which has increased to fairly high frequencies in malaria-endemic regions within Africa, while being rare or absent in other populations in non-endemic regions[26]. The presence of this allele is protective against severe malaria and thus confers a survival advantage.

More recently, it has become clear that the global population history is far more elaborate than previously thought, with analyses of ancient DNA revealing complex migrations and replacement of populations in different regions[27–30]. Furthermore, European and Asian populations have come into contact with archaic populations (Neanderthal and Denisovan), with interbreeding resulting in small introgressed genomic segments evident in modern humans in these regions[31,32]. Similarly, there is evidence to suggest contact with archaic or basal populations (a lineage that split even earlier than the oldest known modern African lineage) among Africans[30,33,34]. These introgressed genomic regions among different ancestral groups have provided important insights about historical adaptation and selection events that resulted in segments of archaic ancestry persisting in modern humans, potentially because they afford a survival advantage in certain environments[31,35,36].

Understanding these differences in allele frequencies of variants among populations is important, as this can influence our ability to identify GWAS association signals (FIG. 2). Because the power to detect associations is generally greater for common variation in a given population, most risk alleles identified to date have relatively high MAFs among European populations, having been discovered in European GWAS[23,37]. As a result, risk alleles in current GWAS databases are likely enriched for variants common among Europeans and depleted for variants that are rare among Europeans[23].

In this context, studying more diverse populations would provide opportunities for identification of new associations of variants with disease, as these studies would be better powered to detect disease associations of variants that are common in these populations but rare in other populations[23]. An example of associations that have been identified in a non-European population on the basis of higher allele frequencies in these populations includes variants in the gene *KCNQ1* and T2DM[38]; the SNPs rs2237897 and rs2237892 in the *KCNQ1* gene, which have been associated with T2DM in the Japanese population, have a much higher MAF in South East Asia (0.39 and 0.38, respectively) relative to Europe (0.04 and 0.06, respectively)[39,40]. Identifying these associations in European populations would have required much larger sample sizes given the lower frequency of these variants, as has been discussed previously[38].

Nevertheless, it is important to recognize the trade-off between opportunities for novel discovery and loss of power that may occur due to heterogeneity of genetic structure or allele frequencies among populations in multi-ethnic studies. Heterogeneity in allele frequencies is likely to increase power in multi-ethnic studies in the context of genetic variants that are rare or absent in one population but more common in other ethnic groups, providing novel opportunities for discovery that would not exist if only the former population was sampled[41]. However, heterogeneity with regard to allele frequencies and population structure (differences in LD) can also reduce power in the context of common and shared genetic variation, where associations with traits are detected with higher statistical resolution in more homogeneous populations. For example, a simulation study showed that including an additional 10,000 individuals from an African population in a GWAS of 10,000 individuals from a European population substantially improved power to detect associations for variants that were of low frequency in Europeans, relative to conducting a GWAS of all 20,000 individuals from a population of European ancestry; power to detect an association increased by 40%, with a relative risk of 1.3 for variants with frequencies between 1% and 5%[41]. This increase in power was largely driven by variants that had allele frequencies that were 15–40% higher in Africans relative to Europeans. However, for variants that were common in Europeans, a loss of power was observed when African individuals were included in the second stage of the GWAS; nevertheless, this decrease was marginal, as power to identify

associations for common variants was already high among Europeans[41].

Population-specific variants. Population specificity is a special case of allele frequency differentiation, whereby a given variant is present in a specific population but absent in another. This concept must be understood as a dynamic concept, because whether a variant seems to be specific to a given population depends on the sampling frame. For example, a variant observed in population A can be defined as specific to population A if not observed in a finite sample within population B, although it is possible that it may be present, albeit rare, in population B and could be observed if all individuals within the population were sampled. This concept must therefore be treated as contingent, to some extent, on the observed sample of sequences from different populations at a given time. Truly population-specific variants are more likely to be rare variants and recent in origin relative to common ancestral variants, having occurred in a specific lineage of a population following divergence from other populations[13]. These variants often track with recent demographic changes among populations, including rapid population expansions[22]. Rapid population expansions lead to an increase in rare variants due to the entry of new mutations into the population with every generation. Rare variants may also cluster geographically[42]. With the development of large-scale whole-genome sequence resources, such as the 1000 Genomes Project or the African Genome Variation Project, it has become clear that a substantial proportion (between 10 and 23%) of non-reference alleles observed in individuals within a given population may be defined as specific to that population[2,3].

There are several examples of population-specific variants that have been implicated in infectious and non-communicable diseases. Important infectious disease susceptibility loci such as the *CCR5* Δ32 variant, which confers protection against HIV transmission and disease progression, are found principally in Europe and West Asia and are absent in sub-Saharan Africa; this allele has been shown to be under strong selection potentially relating to smallpox, with long-range dispersal and selection gradients explaining the differences in allele frequency observed within Europe[43]. Furthermore, studies examining differences in hypertension susceptibility across global populations have cited population-specific variants at several loci, including *ALDH2*, which is associated with hypertension[44,45]. Specific loss-of-function mutations in *PCSK9* that are associated with substantially lowered LDL-cholesterol levels and risk of heart disease were found to be more common in populations of African descent and are rare or absent in European populations[46,47]. A population-specific variant in *MYBPC3* associated with cardiomyopathy has a frequency of ~4% in the Indian subcontinent but is rare or absent elsewhere[48].

In addition to SNPs, copy number variations (CNVs) and structural variants can also be functionally important, with many shown to be associated with specific diseases, including autism spectrum disorders, neuro-developmental disorders[49] and congenital abnormalities[50]. Substantial differences in the distribution of CNVs among ethnic groups have been described, with one study showing only 15% of overlap in genotyped CNVs between European and East Asian populations[51].

Heterogeneity in variant effects. In addition to the heterogeneity in allele frequencies of risk variants, the observed effects of risk alleles on complex traits can also vary across populations. Observed heterogeneity in inferred effect sizes can arise from differences in LD patterns around the causal variant between populations, when the causal variant is unobserved (FIG. 3a). Other reasons for this variation include gene–environment (FIG. 3b) or gene–gene (FIG. 3c) interactions, which may be mediated by epigenetic factors that influence disease-related signalling pathways or gene expression. Heterogeneity of genetic effects can influence the power for discovery of associations across populations, with high levels of heterogeneity reducing power to detect associations[52]. Furthermore, understanding heterogeneity in effect across populations may provide important insights into the mechanisms underlying disease susceptibility in different contexts. The delineation of factors underlying the heterogeneity of variant effects has been limited in most studies; detection of gene–gene and gene–environment interactions would require much larger sample sizes than those needed for the identification of primary associations of interest[10], limiting the capacity to explore these interactions.

The extent to which common shared variants across populations have varying effects on disease susceptibility is unknown, although most studies suggest broad consistency in effect sizes of associations validated in replication GWAS or meta-analyses across populations for most variants. For example, one study examining 43 validated gene–disease associations across 697 study populations with different ancestry found broadly consistent effects at loci across populations, with only 14% of variants studied showing substantial heterogeneity in effect size (defined as I^2>75%, where I^2 is a statistical variable representing the difference in effects attributed to heterogeneity rather than chance)[24]. We note that the analysis of between-study heterogeneity in this study may have been limited by poor representation of non-European studies (with regard to sample size and number). Analysis of between-population heterogeneity can also be diluted by substantial within-population heterogeneity, as noted in the study that screened 134 meta-analyses to examine the genetic effects for 43 validated gene–disease associations across 697 study populations of various ancestries; 46% of the screened studies showed significant within-population heterogeneity[24]. A similar observation was made by another study[53], suggesting that a large proportion of statistical heterogeneity may be attributable to factors other than ancestry[24].

Similarly, a systematic review of meta-analyses of six cancer types observed generally consistent genetic effects across different ethnic groups, although the power to detect heterogeneity of effect was limited in many studies due to limited sample size[54]. Statistically

Non-reference alleles
An allele that is different from the allele in the human reference genome at a given position. The human reference genome is a curated human genome assembly that is based on existing knowledge about the human genome at a given time.

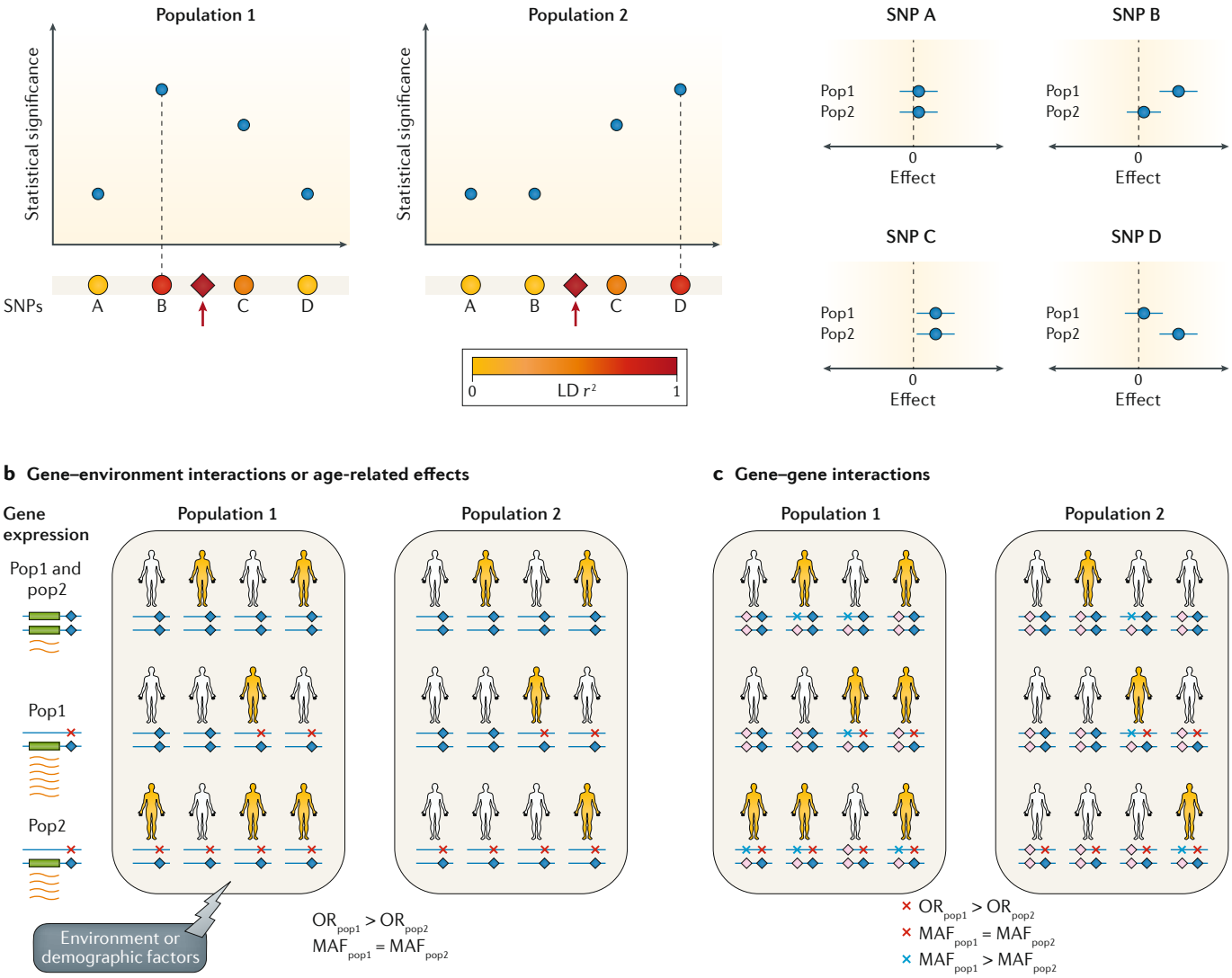


Fig. 3 | **Mechanisms for observed heterogeneity of effect size between populations. a** | Heterogeneity in observed effect size for specific markers between populations (populations 1 and 2) arising due to differences in linkage disequilibrium (LD) between the causal variant (unobserved) and observed variants between two populations. The grey box shown below the graphs represents a chromosomal locus, with the diamond representing the unobserved causal variant at the locus and the circles representing observed markers in the study. The colour of the circle represents the degree of correlation (LD) with the causal variant. In population 1, marker B is in high LD with the causal variant and marker D is in low LD with the causal variant, whereas in population 2 marker D is in high LD with the causal variant and marker B is in low LD with the causal variant. On comparing effect sizes of markers between populations (right panels), these differences in LD with the causal variant manifest as heterogeneity of effect size for markers B and D between populations 1 and 2. **b** | Gene–environment interaction, whereby the effect of the risk allele is amplified by the presence of a specific environmental factor. Yellow shapes represent diseased individuals. The two blue lines simplistically represent the diploid genome of each individual, with the allele represented by the red cross being the risk allele and the blue diamond being the non-risk allele. Although the same number of individuals carry the risk allele in both populations, more individuals carrying the risk allele develop disease in population 1. Of note, this scenario also applies in circumstances where the demographic patterns (for example, older age for age-related manifestation of genetic effects) in a certain population can influence the gene effect. These gene–environment interactions can be mediated through epigenetic differences that alter gene expression differentially in the two populations (shown on the left side of the figure). **c** | Gene–gene interactions, whereby the effect of the risk variant is amplified in the presence of another variant (interacting variant is represented by a blue cross; non-interacting variant is represented as a pink diamond). Differences in allele frequencies of the interacting variant, or differences in LD structure (influencing how often the risk variant and the interacting variant occur together), can lead to differences in effect sizes among different populations. In this case, the interacting variant occurs commonly along with the risk allele in population 1 but not in population 2, resulting in heterogeneity in genetic effects between the populations. OR, odds ratio; MAF, minor allele frequency; pop, population; r^2, Pearson's coefficient of correlation; SNP, single-nucleotide polymorphism.

significant heterogeneity in effect size was identified among 25% of associated SNPs, with levels of heterogeneity being associated with differences in LD structure at loci[54], suggesting that the heterogeneity was likely overestimated and attributable to differences in tagging of causal SNPs across different ethnic groups.

Another study[25] examined the heterogeneity of allele frequencies and effect sizes at 108 loci identified as

statistically significantly associated with complex traits in at least one ethnic group in GWAS from the NHGRI catalogue[6]. This study showed substantial heterogeneity in effect across ethnic groups. Among the eligible associations examined, significant heterogeneity in effect sizes was seen in European–Asian comparisons (22% of associations) and European–African comparisons (42% of associations)[25]; however, given that the interrogated variants were not necessarily causal, these observations may reflect heterogeneity in LD patterns across populations, where different variants tag the causal variant differentially[25], as suggested in previous work[54].

Despite assessment of heterogeneity in recent multi-ethnic studies[52,55], there have been few reproducible examples of heterogeneity of effect at known causal or functional loci associated with disease across populations. Understanding pathways mediating this association in cellular models from diverse populations could provide important insights into biological heterogeneity. One example of heterogeneity of effect with mechanistic evidence is the modification of the association between *MTHFR* 677C>T and stroke risk. *MTHFR* 677C>T has been shown to be associated with increased homocysteine levels (reduction in homocysteine: 3.12 μmol L^{-1}, 95% CI 2.23–4.01 in low-folate regions; 0.13 μmol L^{-1}, 95% CI –0.85 to 1.11 in areas with folate fortification) and increased risk of stroke (OR 1.68, 95% CI 1.44–1.97 in low-folate regions compared with OR 1.03, 95% CI 0.84–1.25 in regions of folate fortification), with effects on homocysteine reduction and stroke being larger in regions with low folate in the diet relative to regions where folate fortification is undertaken to increase dietary folate[56]. This hypothesis of gene–environment interaction for *MTHFR* and dietary folate is biologically compatible with the known metabolism of folate and homocysteine and the action of the enzyme produced by the *MTHFR* gene[56].

These findings collectively suggest that while statistical heterogeneity in multi-ethnic meta-analyses may indicate biological heterogeneity of effect, as a result of gene–environmental and gene–gene interactions, statistical heterogeneity can arise from artefactual and study design factors such as differences in LD structure around the causal variant — which result in differences in the efficiency of tagging in different populations; genotyping errors and differences in imputation accuracy across populations; and non-genetic factors such as differences in phenotype definitions and the demographic and disease profile of individuals included in different studies. Examining and reporting statistical heterogeneity remain pivotal and could provide important insights into the allelic architecture at a locus, fine-mapping of causal drivers of the association and identifying biological mechanisms underlying disease in larger studies[57]. However, substantial statistical heterogeneity can reduce the power to detect association signals in multi-ethnic GWAS[58,59]. Recently developed statistical methods[55,60] that allow exploration of heterogeneity combined with large-scale study resources in diverse populations provide new opportunities to identify and explore heterogeneity among populations.

Differences in functional elements. While clear differences in genome sequence are observed among different global populations, differences in genome function, for example, gene expression and epigenetic profiles, are poorly understood. Gene expression and functional annotation databases, such as ENCODE[61], Roadmap Epigenome[62] and the Genotype-Tissue Expression (GTEx) Project[63], are largely focused on cell lines and tissue originating from European ancestry samples. Limited exploration of differences in gene expression profiles from lymphoblastoid cell lines across populations of varying ancestry has suggested potentially important differences in the transcriptomic landscape between ethnic groups[64–66], with up to 25% of variation in gene expression among individuals attributable to population ancestry[64]. The contribution of differences in gene expression and alternative splicing to this variation has been inconsistent among studies[64,67,68], with some reporting primarily differences in transcript isoforms among genetically distant populations[68] and others reporting predominately differences in gene expression (transcript usage)[65]. Consistent with these differences observed among populations of different ancestries, prediction of gene expression in individuals using reference databases has been found to be dependent on the composition of ancestry within the database. For example, using European gene expression reference databases has been shown to be less accurate for the prediction of gene expression in individuals of African descent, with a true-positive rate >0.10 lower when using training data from European individuals compared with using data from African-American individuals for training[69].

Similarly, studies of DNA methylation have suggested differences among populations[70–72], providing early insights into potentially different regulation of genomic function across different populations. One study showed differential cytosine methylation at 13% of studied CpG sites between European and African populations[72]. Several genetic variants associated with these differential modifications, that is, methylation quantitative trait loci (mQTLs), were associated with cardiometabolic and respiratory traits in previous GWAS[72]. Another study examining DNA methylation variation at 552,141 CpG sites across the genome in monocytes across individuals of European and African ancestry identified 14.1% statistically significant differentially methylated sites[73]. mQTLs were found to account, on average, for ~58% of the observed population differences in DNA methylation, suggesting that a substantial proportion of population differences in methylation were driven by differences in allele frequency of mQTLs[73].

While there has been a recent increase in the representation of diverse non-European populations in GWAS, parallel development of functional resources such as tissue expression databases, well-characterized immortalized cell lines and induced pluripotent stem cell lines from diverse populations has been very limited[63]. Differences in the regulation of gene function among populations suggest that our understanding of genome function from European populations may not be directly transposable to other populations. This has implications not just for our understanding of disease susceptibility

and mechanisms underlying disease in different populations but also for the development of disease therapies that are applied globally. Understanding the impact of genetic variation and environmental factors on genome function across different populations will require parallel development of functional and cellular resources from diverse global populations.

From diversity to discovery

As we have discussed, genomic diversity — differences in the allele frequency spectra, LD structure and functional genomic elements — is thought to have arisen from complex demographic histories of populations (including population bottlenecks, migrations and expansions), genetic drift and adaptive forces relating to differences in environmental exposures. The genetic variability among populations provides important opportunities for discovery of new genetic loci associated with disease that may be evident in some populations and not others. In this context, populations with high levels of genetic variability provide greater opportunities to understand the impact of genetic variation on traits or diseases. Here, we provide examples of specific contexts in which discovery of associations can be enhanced when studying genetically diverse populations (TABLE 1).

Founder effects. Over the past decade, an increasing number of investigations have leveraged the known demographic history among populations to enhance novel discovery of variants associated with disease risk. An example of this approach is the study of population isolates with a limited number of founders[74]. Small founding population sizes in such isolates lead to drift forces having a higher effect on genomes within these populations, sometimes allowing variants that are rare in other populations to increase to higher frequencies in these populations (FIG. 2). The enrichment of rare variants is dependent on multiple factors, including the number of population bottlenecks and effective population size, leading to unique compositions of rare alleles in each population isolate[75]. This may present as differential susceptibility to some diseases. For example, the Pima Indians of Arizona have a very high prevalence of T2DM (~38%)[76–78] and a near absence of type 1 diabetes mellitus[74].

This pattern of enrichment of certain rare variants can enhance discovery of associations with disease. An example is the cardioprotective variant p.Arg19Ter in the *APOC3* gene (rs76353203), which is associated with reduced blood triglyceride levels and has drifted up in frequency in the Amish founder population and, independently, in an isolated population from Crete, Greece[79–81]. Another example is the nonsense p.Arg684Ter variant in the *TBC1D4* gene (rs61736969), which is found at high frequencies in the Greenlandic population and is associated with a substantially increased risk of T2DM among homozygotes[82].

As the power to detect variants associated with disease is highly dependent on allele frequency, the effects of rare variants associated with disease are much more likely to be detected in isolated populations (even in fairly modest sample sizes of a few thousand individuals) where these are common[80,83].

Selection. Selection can have an important role in the differentiation of functional variants across populations. For example, common variants associated with sickle cell anaemia, glucose-6-phosphate dehydrogenase

Table 1 | **Characteristics of specific populations and cohorts that facilitate genetic discovery**

Population	Characteristic	Opportunities
Genetically diverse populations (for example, African populations)	High levels of genetic variation among individuals in the population	Novel discovery of loci associated with traits relative to less diverse populations — for example, population-specific variants, variants common in these populations but rare in other well-studied populations
Population isolates, founder populations (for example, Amish populations, Greek isolates)	Low effective population size, relative genetic homogeneity, enrichment for some rare deleterious variants	Novel discovery among loci that have increased to high frequencies in these populations but are rare in most other global populations
Populations with high levels of consanguinity (for example, Middle-Eastern populations)	High levels of homozygosity	Assessment of pathogenic potential of rare variants in homozygous form and gene function by assessment of naturally occurring gene knockouts
Admixed populations (for example, African-Americans)	Genomes of individuals are a mosaic of haplotypes of different ancestral origin	Assessment of the association between local ancestry with disease (where disease susceptibility is known to vary among source populations). Cases with disease will be enriched for specific ancestry at loci associated with disease
Populations exposed to different environmental stimuli (for example, sub-Saharan African populations exposed to malaria)	Genetic adaptation in response to environment stimulus	Adaptation, including selective sweeps or balancing selection leading to certain alleles rare or absent in other populations reaching higher frequencies in these populations (for example, the sickle cell variant associated with malaria)
Multi-ethnic cohorts	High levels of differentiation between different ethnic groups studied and different linkage disequilibrium patterns	Better resolution of causal variants associated with traits or diseases
Family-based cohorts	Pedigrees with related individuals (diseased and healthy), with detailed phenotyping for each pedigree	• Assessment of loci associated with Mendelian disease; discovery of de novo mutations associated with disease • Assessment of heritability of complex traits, accounting for shared environment

(G6PD) deficiency or α-thalassaemia are thought to have increased to high frequencies in some African populations owing to the protection these variants confer against severe malaria[84–86]. These variants have been found to be associated with haematological traits in GWAS among individuals of African ancestry[87–89], although equivalent studies at a much larger scale have not detected these association signals in largely European cohorts, where these variants are relatively rare[90]. In this context, selection related to malaria has increased the allele frequencies of these otherwise deleterious mutations in endemic regions[91].

While selective sweeps such as those described above are most often described in the literature, more pervasive — and potentially more important, albeit fairly small — changes in allele frequencies at numerous loci are likely to have occurred through adaptive selection. Changes arising from environmental adaptation can be difficult to distinguish from random stochastic changes resulting from genetic drift. Parallel adaptation refers to independent multiple mutations that arise in parallel, at the same locus or different loci, and give rise to the same adaptive or advantageous phenotype[92]. In many cases of parallel adaptation, different alleles may reach intermediate frequencies in populations, not giving rise to an allele that reaches fixation in a given population (a hard selective sweep) but, rather, a 'soft selective sweep'[92]. Previous work has indicated substantial parallel adaptation across geographically distinct populations[93], suggesting that genetic architecture is shaped in parallel ways by the environment in different regions over thousands of years.

Genes that have adapted in a parallel manner across populations can have important pleiotropic functions[93], such that selection on one trait has effects on the susceptibility to other traits. For example, parallel divergence among populations was observed for the *IFIH1* gene[93], which has been associated with several traits, including antiviral defence[94], type 1 diabetes mellitus[95] and psoriasis[96]. Given the high polygenicity of most traits, these complex adaptive forces can have a collective impact on genetic susceptibility to disease among different populations.

Adaptive selection can thus lead to differentiation of functionally important alleles among populations, with some alleles reaching high frequencies in specific populations when they confer a selection advantage in the presence of specific environments. These differences can be leveraged for enhanced genomic discovery.

Admixture. Studies showing associations between the proportion of ancestry inherited from a given source population and disease suggest that differences between disease susceptibility among populations may be genetically determined[97–99]. Admixture mapping is an approach that leverages potential differences in genetic susceptibility to disease among different ethnic groups to examine the association between local genetic ancestry and disease across the genome. It relies on the principle that for diseases where associated genetic variants differ substantially in frequency between ancestral populations, admixed individuals with disease will be enriched for ancestry from the population with the higher proportion of risk alleles at loci associated with disease[100,101]. This approach has been used to identify genetic loci associated with hypertension[98,99], infectious disease susceptibility[101], prostate cancer[102], multiple sclerosis[103] and cardiometabolic diseases[101,104,105].

Admixture mapping can also facilitate discovery of loci associated with traits in the context of parallel adaptation, where variants at multiple loci that have entered populations through admixture can reach high frequencies (due to adaptive selection), thus manifesting as greater than expected local ancestry from one source at these loci. An example of this is the genetic adaptation to high altitude among Tibetans, which can be inferred through enrichment of high-altitude (Sherpa-like) ancestry at the *EPAS1* and *EGLN1* genes, which are known to be associated with haemoglobin concentration[106].

Endogamy and autozygosity. Cultural practices vary among different populations, and certain practices such as consanguinity and endogamy can influence disease susceptibility. This fact has been recognized in some populations, where the burden of recessive genetic disease is thought to be linked to these cultural practices. High rates of consanguineous marriage occur in North Africa, the Middle East and West, Central and South Asia[107]. Over generations, endogamy can lead to similarities between the inherited maternal and paternal chromosome segments, resulting in long segments of autozygosity. Autozygosity raises the probability that two deleterious alleles occur together (homozygosity), increasing the susceptibility to monogenic recessive disorders. Consanguinity has been associated with an increased prevalence of haematological disorders such as α-thalassaemia and β-thalassaemia in Middle Eastern countries[108].

Populations with high levels of autozygosity provide a unique opportunity to examine the functional impact of gene knockouts, owing to the higher probability of deleterious recessive mutations being homozygous. In addition to monogenic disorders, consanguinity has also been associated with an increased susceptibility to complex traits such as tuberculosis and chronic hepatitis B infection in West African individuals[109]. Studies in well-characterized population isolates have suggested that the risk of hypertension is correlated with the inbreeding coefficient[110], which indicates that part of this raised risk is attributable to an increase in deleterious recessive mutations associated with disease.

Delineating genetic effects from shared environmental effects in large-scale studies of populations with high levels of consanguinity, using variance partitioning approaches[111], will be vital to understanding the impact of autozygosity on complex diseases and may provide novel insights into the genetic architecture of complex traits. Recent efforts such as the East London Genes and Health Study, the Saudi Genomes Project[112] and the Qatar Genomes Project provide unique opportunities to study the impact of autozygosity on disease.

Improved resolution of causal variants. Although GWAS have continuously increased the number of loci associated with complex diseases or traits over the past decade, the resolution of causal variants driving these

Adaptive selection
Evolutionary changes to the genome that occur due to selection and are adaptive to the given environment.

Fixation
The change in the genetic pool of a population from the presence of two alleles at a given locus to only one allele being present; this allele is said to be fixed.

Admixture
Interbreeding or mixing of two or more populations that were previously isolated.

Consanguinity
The state of being closely related to someone by sharing a recent ancestor; in genetics, commonly used to refer to mating with close relatives.

Endogamy
The practice of marrying only within the limits of a local community, clan or tribe.

Autozygosity
Stretches of the two homologous chromosomes within the same individual that are identical by descent; occurs when there is non-random mating.

Inbreeding coefficient
The probability that two alleles at a locus in an individual are identical by descent from a common ancestor, that is, the chance that an individual is homozygous for an ancestral allele by inheritance (not by mutation).

associations has lagged behind. Although LD increases the power to detect associations in populations by facilitating imputation and allowing tagging of causal variation, LD can also limit the identification of causal drivers of the association signal when multiple variants at a given locus attain equivalent statistical significance for association with the trait or disease in question. Differences in LD structure among populations can provide opportunities to resolve causal associations at such loci. This is particularly true for African populations, where LD is weaker across the genome and decays faster[2], allowing better resolution of associated loci in GWAS (BOX 1). Indeed, studies have shown that addition of modest samples from African populations can greatly improve resolution at loci, relative to adding large numbers of samples from a homogeneous population[113]. Approaches that leverage heterogeneity in allele frequencies and LD can facilitate novel discovery in multi-ethnic meta-analyses and allow better resolution of identified signals[52,55,113–115].

Implications for medical genetics

Knowledge of the genetic susceptibility to disease in globally diverse populations and of the interplay of genetic and environmental factors contributing to disease is important to understand population disease risk and to inform preventive, diagnostic or therapeutic strategies.

Implications for risk prediction. Understanding the risk associated with specific loci can facilitate the direct development of risk scores that, in combination with clinical risk factors, can be used to predict the likelihood of developing a given disease[116]. Many studies have highlighted the limitations of applying polygenic risk scores (also known as genome-wide polygenic scores) that have been ascertained from European cohorts to other populations, as these are likely to be biased and reduce predictive accuracy[117–120]. These biases are thought to relate to several factors, including biases in the allele frequency spectrum of risk variants ascertained in European GWAS, with undiscovered associated variants that are common in non-European populations but rare among Europeans not included in scores[37]; differences in LD structure around the causal variant among populations, leading to error in assignment of appropriate risk scores to the causal allele which may be unknown; and heterogeneity in effect sizes across populations. Given these caveats, understanding and characterizing genetic risk of disease among diverse populations is essential for the successful application of risk prediction scores among these populations. Indeed, even inclusion of data from modest-sized studies from the target population to European-ascertained GWAS data can substantially improve prediction of risk and reduce bias[120].

Implications for screening and diagnostics. Differences in the spectrum and frequency of mutations across populations are likely to have an impact on screening initiatives when genetic tests have been designed based on the mutational spectrum in a specific ethnic group. That is, the majority of variants in databases for clinically significant or pathogenic genetic mutations have been ascertained in European individuals and may not be representative of other less studied population groups[2,121]. This has implications for clinical genetics diagnostics and precision medicine initiatives.

For example, cystic fibrosis is a recessively inherited disease caused by mutations in the *CFTR* gene. The spectrum and frequency of individual *CFTR* variants varies among ethnic groups and geographical locations, with the p.Phe508del mutation identified in 90% of white patients with cystic fibrosis, whereas this mutation is absent in 17%, 30%, 38% and 40% of those of Native American, Hispanic, African and Asian ethnicities with cystic fibrosis, respectively[122]. While the American College of Medical Genetics 23-mutation panel for cystic fibrosis screening reportedly identifies the majority of white and Native American patients with cystic fibrosis by identifying two copies of causal mutations within the gene, less than half of the patients with cystic fibrosis of other ethnicities would have causal variants discovered based on these tests[122]. Hence, allelic heterogeneity at loci associated with syndromic diseases among different populations, as identified for cystic fibrosis[123], has direct implications for screening and diagnosis of individuals with disease in ethnically diverse populations.

As another example, recent work identifying associations between a common α-thalaessemia variant (22% frequency among African populations) and *G6PD* variants with glycated haemoglobin levels suggest that these effects must be taken into account when considering the use of glycated haemoglobin as a marker for the diagnosis of diabetes mellitus in African populations in whom these mutations are common[124].

Implications for therapeutic strategies. Genetic diversity among populations has been shown to influence responses to drugs, which has important implications for precision medicine and pharmacogenomics. The necessity for different dosages of the anti-clotting drug warfarin to maintain therapeutic effects across different ethnic groups[125,126] illustrates the need for inclusion of diverse population groups in pharmacogenomic investigations and trials. Understanding pharmacogenomic differences among populations has the potential to directly inform clinical care and potentially avert adverse events[126]. Initiatives such as Human Heredity and Health Africa (H3Africa) have established local networks and capacity for large-scale GWAS across Africa, providing exciting opportunities for drug discovery[127].

Conclusions and outlook

Inclusion of diverse populations in studies of genetic determinants of disease has to date been fairly limited due to several challenges, including the need for building close partnerships with local communities and governments, regional collaborators and academic universities, as well as the need for building human and infrastructure research capacity in a sustainable way. These efforts require substantive financial and time investments, which would benefit from strategic support from funding bodies outside mainstream research grant systems. Recent funding initiatives, for example,

Box 1 | **Linkage disequilibrium**

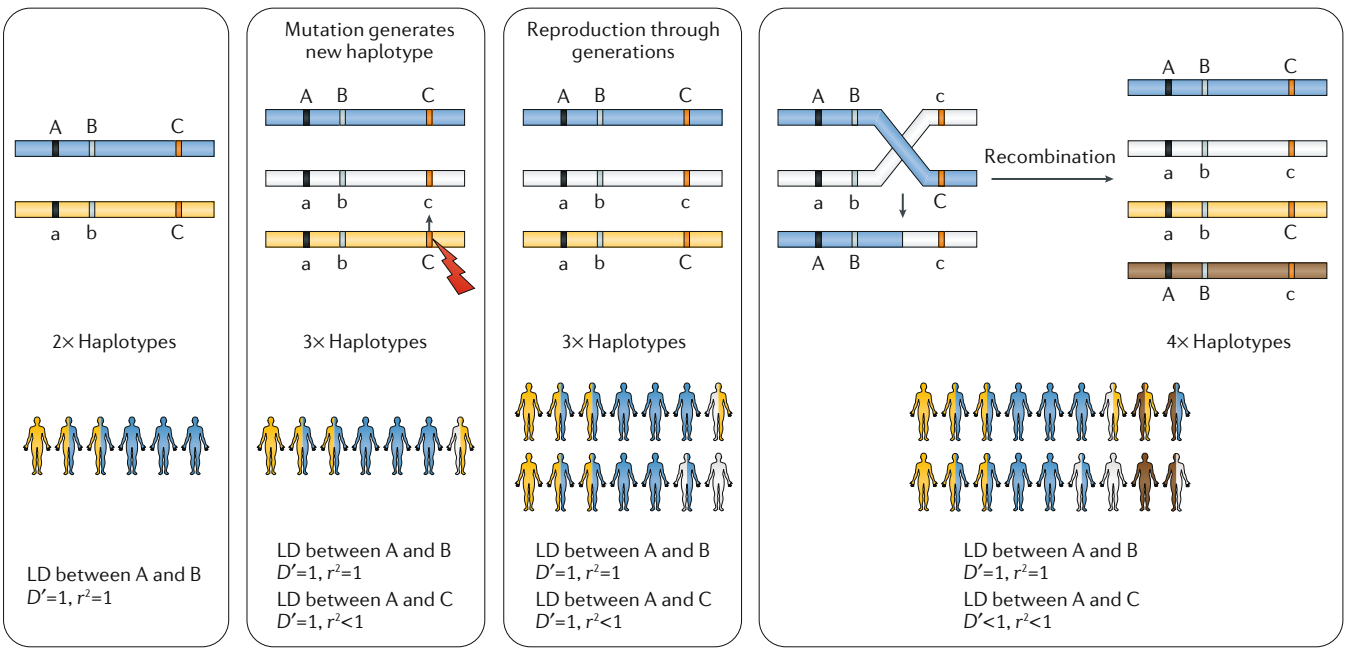
Linkage disequilibrium (LD) refers to the non-random correlation between markers along a chromosomal segment. LD is defined as the difference between the observed frequency of a particular combination of alleles at two loci and the frequency expected if assortment was random. Initially, when a new mutation arises on a haplotype, it is in perfect LD with the markers on the haplotype. As recombination between markers occurs, linkage decays over time. Recombination depends on the genetic distance and the number of generations — it occurs with higher probability the further alleles are from each other and with each generation, leading to decay in LD over time. The recombination rates across the human genome lead to specific patterns of LD observed in humans. The human genome is made up of segment blocks of loci in strong LD with each other. These are referred to as 'haplotype blocks', and boundaries are usually associated with hot spots of recombination. Haplotype blocks in humans can vary in size from a few kilobases to more than 100 kb. These blocks can make it difficult to delineate causal variants associated with disease in genome-wide association studies (GWAS).

Factors influencing LD

LD is affected by population demographic forces, such as genetic drift, mutation, selection and admixture. Population bottlenecks increase LD between markers, and slower decay may be observed over distance. The decay of LD has been shown to be lower for European populations and other populations that have been through population bottlenecks, relative to African populations[2]. Selection forces can also lead to increased LD between markers if a given haplotype with specific combinations of markers is selected for and increases in frequency relative to other haplotypes. Genetic drift can act in conjunction with population bottlenecks to increase LD between markers. Mixing between populations can also lead to strong LD between markers, especially for markers that are highly differentiated between the mixing populations[130]. This is referred to as admixture LD. Admixture LD can provide important information about the time and complexity of admixture between source populations.

What information does LD give us? Differences in patterns of LD and rates of LD decay among populations have been used to better resolve causal variants associated with traits in GWAS. Due to differences in population history, different observed markers may be in strong LD with (that is, 'tag') the unobserved causal variant in different populations. This can provide clues as to which variant is likely to be causal. Due to lower LD between variants in African populations, haplotype blocks are shorter, making it easier to resolve regions within which causal variants lie. Using these differences in LD to better resolve the true causal variant driving associations in GWAS is referred to as fine-mapping. LD between markers that does not decay over relatively long distances may be indicative of selection forces that have led to specific haplotypes reaching high frequencies in a given population. LD patterns have therefore been used to identify regions of the genome under natural selection[131,132].

As LD decays over time, it can be used to estimate the age of a haplotype. For example, markers in high LD with each other, where recombination has not occurred as expected based on local recombination rates, indicate a potential recent increase in haplotype frequency that is indicative of recent selection. The figure represents the influence of recombination on LD over time. Two haplotypes (blue and yellow) differ in alleles at the single-nucleotide polymorphisms A and B. The alleles A and B lie on one haplotype, and a and b lie on another haplotype. These two markers are in complete LD, as the combination of A and b or a and B does not exist within the population. Only a single allele exists in the population at position C at this point in time (that is, the site is monomorphic in the population). However, a mutation occurs at this site, with C mutating into c in one haplotype, producing a new haplotype abc (white). At this point, three haplotypes exist within the population. As the population expands, and reproduces, this haplotype is also observed in a proportion of the population. However, there is strong LD ($D' = 1$) between markers A and C as only three haplotypes exist. The haplotype Ac does not exist at this point. During reproduction, recombination occurs between two haplotypes, so that all four possible haplotypes between A and C are now seen within the population (Ac, aC, AC and ac). LD has decayed due to recombination between markers A and C. However, markers A and B remain in complete LD as no recombination has occurred between these markers. The recombination probability depends on genetic distance, and markers that are genetically less distant (such as A and B) are more likely to remain in high LD relative to markers that are genetically distant (such as A and C), as shown in the example here.



D', relative measure of disequilibrium (D) compared to its maximum; r^2, Pearson's coefficient of correlation.

the National Institutes of Health's Trans-Omics for Precision Medicine (TOPMed) programme, have begun to enable such activities. Including genetically diverse populations in GWAS is a pragmatic strategy to enhance novel discovery of associations with diseases or traits and provide better resolution of causal variation and allow translation of discoveries into clinical practice.

Understanding genetic contributions to disease susceptibility across populations will be vital for precision medicine initiatives, including for risk prediction and the development and evaluation of therapies, enabling global equity in the benefits of genomics[1,128]. Comprehensively understanding the contribution of rare and common variation to different diseases among populations will require the development of multi-ethnic large-scale population resources with whole-genome sequencing or whole-exome sequencing data, along with relevant clinical data. Inclusion of functional annotations in GWAS can also provide important information about the contribution of different genetic and epigenetic factors to traits[129]. This, however, requires more comprehensive characterization of molecular quantitative trait loci in larger sample sizes and across diverse human populations[129]. Collection of integrated electronic health-record real- world resources, in conjunction with genetic sampling, provides a cost-effective strategy for the simultaneous examination of multiple complex disease traits and treatment phenotypes.

Future work in diverse populations should focus on using unbiased approaches, including unbiased variant discovery and genome references, as well as study designs that incorporate globally diverse whole-genome or whole-exome sequence data and genotyping using arrays that enable efficient genomic coverage for diverse or specific populations. Unbiased design of GWAS in globally diverse populations will provide vital information to better understand reproducibility and heterogeneity of effects among populations, as well as important resources for fine-mapping of causal variants. In parallel, the development of methods that leverage differences in genetic architecture among populations and better characterize heterogeneity among populations[52,55,114,115] will improve the power to identify causal drivers of association signals, including at complex loci where effects are a composite of multiple drivers. With larger scale and more diverse populations in GWAS, along with the ability to better capture rarer genetic association signals, we may observe increasing population differences in the structure and shape of genetic association signals globally, providing much finer-scale insights into the genetic patterns of disease risk.

A more comprehensive understanding of genetic determinants of disease susceptibility worldwide will require moving from GWAS to understanding biological mechanisms underlying associations and functional validation in global populations. This shift will involve the development of globally relevant functional resources, including tissue biobanks and transcriptomic resources across global populations, to better understand the impact of population-specific variation and heterogeneity in variant effects at the transcriptomic level. Functional validation will also require the development of cellular models from genetically diverse populations in order to directly observe in vitro effects in relevant cell types with specific genetic profiles. New large-scale GWAS based on population-specific and multi-ethnic bioresources will provide unprecedented opportunities to understand genetic susceptibility to disease globally. Mega-biobank initiatives such as the China Kadoorie Biobank, BioBank Japan, H3Africa, the NIH All of Us, the Finnish Biobanks and the Million Veterans Program (TABLE 2) are expected to enable the unprecedented

Table 2 | **Large-scale resources for genomic studies of diverse populations**

Resource	Ethnicity	Data collected	Description
China Kadoorie Biobank	Chinese	Lifestyle data, measurements, death and health-related data; 25,000 surveyed repeatedly for follow up	>510,000 individuals recruited
BioBank Japan[133]	Japanese	Questionnaires, measurements, laboratory tests, imaging, serial review of medical records, disease codes and survival data	200,000 participants recruited, hospital-based sampling
NIH All of Us	Multi-ethnic (oversampling from diverse communities, 50% from ethnic minorities)	Lifestyle data, blood tests, electronic health records	>210,000 participants recruited in year 1; aim to recruit 1 million participants
H3Africa	African	Genotyping, whole-genome sequencing data, biomarkers, clinical data	>54,000 participants recruited to multiple projects
Million Veteran Program	Multi-ethnic	Blood biomarkers, electronic health record data	>730,000 participants recruited; 450,000 samples genotyped; 10,000 samples whole-genome sequenced
Finnish Biobanks	Finnish	National health register, blood samples	Cooperative of six hospital-based biobanks; aim to recruit 500,000 Finns (10% of population)
TOPMed Programme	Multi-ethnic (40% European, 32% African, 16% Hispanic/Latino, 10% Asian, 2% Other)	Risk factors, subclinical disease measures and incident disease (heart, lung and blood disorders)	• ~144,000 participants • Sequencing of patients with heart, lung, blood and sleep disorders (National Heart, Lung, and Blood Institute); >90,000 participants sequenced

characterization of fine-scale genetic architecture of disease among diverse populations. Further representative population-specific resources from other ancestrally diverse populations across the world are needed to inform our understanding of genetic susceptibility to disease in the global context. This will require a concerted effort of sustained long-term investment in capacity-building in the fields of epidemiology, statistical genetics and bioinformatics, supporting local researchers and infrastructure to facilitate the development of such resources from understudied populations[1].

Published online: 24 June 2019

1. Hindorff, L. A. et al. Prioritizing diversity in human genomics research. *Nat. Rev. Genet.* **19**, 175–185 (2018).
This article presents an insightful review focusing on the need for increased diversity in human genetics research, and efforts by the NHGRI to increase diversity in participants as well as researchers.
2. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
A landmark study from the 1000 Genomes Project Consortium outlining the first whole-genome sequencing study of multiple diverse ethnic groups providing novel insights into differences in genomic variation across different populations.
3. Gurdasani, D. et al. The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**, 327–332 (2014).
This study is one of the first comprehensive evaluations of genetic diversity among different ethno-linguistic groups within Africa based on genotyping data.
4. Mallick, S. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
A study of highly genetically diverse populations across the globe using deep whole-genome sequencing approaches.
5. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
6. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
This article provides a summary of data within the GWAS catalogue (a collection of all GWAS study data deposited to date), including the ethnic distribution of existing studies.
7. Morales, J. et al. A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biol.* **19**, 21 (2018).
8. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
9. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
10. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
This article presents an excellent overview of the history of GWAS and their role in discovery of genetic determinants of disease.
11. Wood, A. R. et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
12. Speliotes, E. K. et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937–948 (2010).
13. Dries, D. L. Genetic ancestry, population admixture, and the genetic epidemiology of complex disease. *Circ. Cardiovasc. Genet.* **2**, 540–543 (2009).
14. Yang, J. et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015).
15. Ruhle, F. et al. Rare genetic variants in SMAP1, B3GAT2, and RIMS1 contribute to pediatric venous thromboembolism. *Blood* **129**, 783–790 (2017).
16. Ng, S. B. et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.* **42**, 790–793 (2010).
17. Auer, P. L. & Lettre, G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med.* **7**, 16 (2015).
18. Marigorta, U. M. & Navarro, A. High trans-ethnic replicability of GWAS results implies common causal variants. *PLOS Genet.* **9**, e1003566 (2013).
19. 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).
20. Haiman, C. A. et al. Consistent directions of effect for established type 2 diabetes risk variants across populations: the Population Architecture using Genomics and Epidemiology (PAGE) Consortium. *Diabetes* **61**, 1642–1647 (2012).
21. DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium 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).
22. Gravel, S. et al. Demographic history and rare allele sharing among human populations. *Proc. Natl Acad. Sci. USA* **108**, 11983–11988 (2011).
23. Kim, M. S., Patel, K. P., Teng, A. K., Berens, A. J. & Lachance, J. Genetic disease risks can be misestimated across global populations. *Genome Biol.* **19**, 179 (2018).
This article presents an important study examining the transferability of polygenic risk scores across different ethnic groups.
24. Ioannidis, J. P., Ntzani, E. E. & Trikalinos, T. A. 'Racial' differences in genetic effects for complex diseases. *Nat. Genet.* **36**, 1312–1318 (2004).
25. Ntzani, E. E., Liberopoulos, G., Manolio, T. A. & Ioannidis, J. P. Consistency of genome-wide associations across major ancestral groups. *Hum. Genet.* **131**, 1057–1071 (2012).
26. Kwiatkowski, D. P. How malaria has affected the human genome and what human genetics can teach us about malaria. *Am. J. Hum. Genet.* **77**, 171–192 (2005).
27. Lipson, M. et al. Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science* **361**, 92–95 (2018).
28. Pickrell, J. K. & Reich, D. Toward a new history and geography of human genes informed by ancient DNA. *Trends Genet.* **30**, 377–389 (2014).
29. Posth, C. et al. Reconstructing the deep population history of Central and South America. *Cell* **175**, 1185–1197 (2018).
30. Skoglund, P. et al. Reconstructing prehistoric African population structure. *Cell* **171**, 59–71 (2017).
31. Prufer, K. et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
32. Meyer, M. et al. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).
33. Hammer, M. F., Woerner, A. E., Mendez, F. L., Watkins, J. C. & Wall, J. D. Genetic evidence for archaic admixture in Africa. *Proc. Natl Acad. Sci. USA* **108**, 15123–15128 (2011).
34. Lachance, J. et al. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* **150**, 457–469 (2012).
35. Sankararaman, S., Mallick, S., Patterson, N. & Reich, D. The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Curr. Biol.* **26**, 1241–1247 (2016).
36. Xu, D. et al. Archaic hominin introgression in Africa contributes to functional salivary MUC7 genetic variation. *Mol. Biol. Evol.* **34**, 2704–2715 (2017).
37. Martin, A. R. et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
This article presents an important study examining the transferability of polygenic risk scores across different ethnic groups.
38. Rosenberg, N. A. et al. Genome-wide association studies in diverse populations. *Nat. Rev. Genet.* **11**, 356–366 (2010).
39. Yasuda, K. et al. Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. *Nat. Genet.* **40**, 1092–1097 (2008).
40. Unoki, H. et al. SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nat. Genet.* **40**, 1098–1102 (2008).
41. Pulit, S. L., Voight, B. F. & de Bakker, P. I. Multiethnic genetic association studies improve power for locus discovery. *PLOS ONE* **5**, e12600 (2010).
This article presents an important study of how inclusion of multi-ethnic populations influences power for discovery in GWAS, in comparison with inclusion of homogeneous populations.
42. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44**, 243–246 (2012).
43. Novembre, J., Galvani, A. P. & Slatkin, M. The geographic spread of the CCR5 Delta32 HIV-resistance allele. *PLOS Biol.* **3**, e339 (2005).
44. Franceschini, N., Reiner, A. P. & Heiss, G. Recent findings in the genetics of blood pressure and hypertension traits. *Am. J. Hypertens.* **24**, 392–400 (2011).
45. Yasukochi, Y. et al. Longitudinal exome-wide association study to identify genetic susceptibility loci for hypertension in a Japanese population. *Exp. Mol. Med.* **49**, e409 (2017).
46. Kent, S. T. et al. PCSK9 loss-of-function variants, low-density lipoprotein cholesterol, and risk of coronary heart disease and stroke: data from 9 studies of blacks and whites. *Circ. Cardiovasc. Genet.* **10**, e001632 (2017).
47. Cohen, J. C., Boerwinkle, E., Mosley, T. H. Jr & Hobbs, H. H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* **354**, 1264–1272 (2006).
48. Dhandapany, P. S. et al. A common MYBPC3 (cardiac myosin binding protein C) variant associated with cardiomyopathies in South Asia. *Nat. Genet.* **41**, 187–191 (2009).
49. Mefford, H. C. et al. Rare copy number variants are an important cause of epileptic encephalopathies. *Ann. Neurol.* **70**, 974–985 (2011).
50. Miller, D. T. et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am. J. Hum. Genet.* **86**, 749–764 (2010).
51. Li, J. et al. Whole genome distribution and ethnic differentiation of copy number variation in Caucasian and Asian populations. *PLOS ONE* **4**, e7958 (2009).
52. Cook, J. P. & Morris, A. P. Multi-ethnic genome-wide association study identifies novel locus for type 2 diabetes susceptibility. *Eur. J. Hum. Genet.* **24**, 1175–1180 (2016).
53. Ioannidis, J. P., Patsopoulos, N. A. & Evangelou, E. Heterogeneity in meta-analyses of genome-wide association investigations. *PLOS ONE* **2**, e841 (2007).
54. Jing, L., Su, L. & Ring, B. Z. Ethnic background and genetic variation in the evaluation of cancer risk: a systematic review. *PLOS ONE* **9**, e97522 (2014).
55. Magi, R. et al. Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Hum. Mol. Genet.* **26**, 3639–3650 (2017).
56. Holmes, M. V. et al. Effect modification by population dietary folate on the association between MTHFR genotype, homocysteine, and stroke risk: a meta-analysis of genetic studies and randomised trials. *Lancet* **378**, 584–594 (2011).
57. Helgason, A. et al. Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. *Nat. Genet.* **39**, 218–225 (2007).
58. Moonesinghe, R., Khoury, M. J., Liu, T. & Ioannidis, J. P. Required sample size and nonreplicability thresholds for heterogeneous genetic associations. *Proc. Natl Acad. Sci. USA* **105**, 617–622 (2008).

59. Kulminski, A. M. et al. Explicating heterogeneity of complex traits has strong potential for improving GWAS efficiency. *Sci. Rep.* **6**, 35390 (2016).
60. Lee, C. H., Eskin, E. & Han, B. Increasing the power of meta-analysis of genome-wide association studies to detect heterogeneous effects. *Bioinformatics* **33**, i379–i388 (2017).
This study outlines an important meta-analytic method to maximize power to detect associations in multi-ethnic GWAS with heterogeneous effects.
61. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
62. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
63. GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
64. Martin, A. R. et al. Transcriptome sequencing from diverse human populations reveals differentiated regulatory architecture. *PLOS Genet.* **10**, e1004549 (2014).
65. Mele, M. et al. Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).
66. Tian, L. et al. Genome-wide comparison of allele-specific gene expression between African and European populations. *Hum. Mol. Genet.* **27**, 1067–1077 (2018).
67. Kelly, D. E., Hansen, M. E. B. & Tishkoff, S. A. Global variation in gene expression and the value of diverse sampling. *Curr. Opin. Syst. Biol.* **1**, 102–108 (2017).
68. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
Thie study is one of the first investigations of the transcriptome across diverse populations from the 1000 Genomes Project, examining the key differences in gene expression and transcriptome structure among populations.
69. Mogil, L. S. et al. Genetic architecture of gene expression traits across diverse populations. *PLOS Genet.* **14**, e1007586 (2018).
70. Giuliani, C. et al. Epigenetic variability across human populations: a focus on DNA methylation profiles of the KRTCAP3, MAD1L1 and BRSK2 genes. *Genome Biol. Evol.* **8**, 2760–2773 (2016).
71. Fraser, H. B., Lam, L. L., Neumann, S. M. & Kobor, M. S. Population-specificity of human DNA methylation. *Genome Biol.* **13**, R8 (2012).
72. Moen, E. L. et al. Genome-wide variation of cytosine modifications between European and African populations and the implications for complex traits. *Genetics* **194**, 987–996 (2013).
73. Husquin, L. T. et al. Exploring the genetic basis of human population differences in DNA methylation and their causal impact on immune gene regulation. *Genome Biol.* **19**, 222 (2018).
74. Hatzikotoulas, K., Gilly, A. & Zeggini, E. Using population isolates in genetic association studies. *Brief Funct. Genomics* **13**, 371–377 (2014).
75. Kristiansson, K., Naukkarinen, J. & Peltonen, L. Isolated populations and complex disease gene identification. *Genome Biol.* **9**, 109 (2008).
This review provides an overview of how studying isolated populations has enhanced discovery through GWAS.
76. Dabelea, D. et al. Increasing prevalence of type II diabetes in American Indian children. *Diabetologia* **41**, 904–910 (1998).
77. Knowler, W. C., Bennett, P. H., Hamman, R. F. & Miller, M. Diabetes incidence and prevalence in Pima Indians: a 19-fold greater incidence than in Rochester, Minnesota. *Am. J. Epidemiol.* **108**, 497–505 (1978).
78. Schulz, L. O. et al. Effects of traditional and western environments on prevalence of type 2 diabetes in Pima Indians in Mexico and the US. *Diabetes Care* **29**, 1866–1871 (2006).
79. Pollin, T. I. et al. A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. *Science* **322**, 1702–1705 (2008).
80. Tachmazidou, I. et al. A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. *Nat. Commun.* **4**, 2872 (2013).
81. Gilly, A. et al. Very low-depth sequencing in a founder population identifies a cardioprotective APOC3 signal missed by genome-wide imputation. *Hum. Mol. Genet.* **25**, 2360–2365 (2016).
82. Moltke, I. et al. A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512**, 190–193 (2014).
83. Laitinen, T. et al. Characterization of a common susceptibility locus for asthma-related traits. *Science* **304**, 300–304 (2004).
84. Mockenhaupt, F. P. et al. Alpha(+)-thalassemia protects African children from severe malaria. *Blood* **104**, 2003–2006 (2004).
85. Elguero, E. et al. Malaria continues to select for sickle cell trait in Central Africa. *Proc. Natl Acad. Sci. USA* **112**, 7051–7054 (2015).
86. Luzzatto, L. G6PD deficiency: a polymorphism balanced by heterozygote advantage against malaria. *Lancet Haematol.* **2**, e400–e401 (2015).
87. Chen, Z. et al. Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network. *Hum. Mol. Genet.* **22**, 2529–2538 (2013).
88. Hodonsky, C. J. et al. Genome-wide association study of red blood cell traits in Hispanics/Latinos: the Hispanic Community Health Study/Study of Latinos. *PLOS Genet.* **13**, e1006760 (2017).
89. Malaria Genomic Epidemiology Network. Reappraisal of known malaria resistance loci in a large multicenter study. *Nat. Genet.* **46**, 1197–1204 (2014).
90. Soranzo, N. et al. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat. Genet.* **41**, 1182–1190 (2009).
91. Hedrick, P. W. Resistance to malaria in humans: the impact of strong, recent selection. *Malar. J.* **11**, 349 (2012).
92. Ralph, P. & Coop, G. Parallel adaptation: one or many waves of advance of an advantageous allele? *Genetics* **186**, 647–668 (2010).
93. Tennessen, J. A. & Akey, J. M. Parallel adaptive divergence among geographically diverse human populations. *PLOS Genet.* **7**, e1002127 (2011).
94. Meylan, E., Tschopp, J. & Karin, M. Intracellular pattern recognition receptors in the host response. *Nature* **442**, 39–44 (2006).
95. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J. A. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**, 387–389 (2009).
96. Li, Y. et al. Carriers of rare missense variants in IFIH1 are protected from psoriasis. *J. Invest. Dermatol.* **130**, 2768–2772 (2010).
97. Manolio, T. A. et al. Ethnic differences in the relationship of carotid atherosclerosis to coronary calcification: the Multi-Ethnic Study of Atherosclerosis. *Atherosclerosis* **197**, 132–138 (2008).
98. Zhu, X. et al. Admixture mapping for hypertension loci with genome-scan markers. *Nat. Genet.* **37**, 177–181 (2005).
99. Zhu, X. & Cooper, R. S. Admixture mapping provides evidence of association of the VNN1 gene with hypertension. *PLOS ONE* **2**, e1244 (2007).
100. Darvasi, A. & Shifman, S. The beauty of admixture. *Nat. Genet.* **37**, 118–119 (2005).
101. Cyr, D. D. et al. Evaluating genetic susceptibility to *Staphylococcus aureus* bacteremia in African Americans using admixture mapping. *Genes Immun.* **18**, 95–99 (2017).
102. Freedman, M. L. et al. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc. Natl Acad. Sci. USA* **103**, 14068–14073 (2006).
103. Reich, D. et al. A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. *Nat. Genet.* **37**, 1113–1118 (2005).
104. Scherer, M. L. et al. Admixture mapping of ankle–arm index: identification of a candidate locus associated with peripheral arterial disease. *J. Med. Genet.* **47**, 1–7 (2010).
105. Elbein, S. C., Das, S. K., Hallman, D. M., Hanis, C. L. & Hasstedt, S. J. Genome-wide linkage and admixture mapping of type 2 diabetes in African American families from the American Diabetes Association GENNID (Genetics of NIDDM) Study Cohort. *Diabetes* **58**, 268–274 (2009).
106. Jeong, C. et al. Admixture facilitates genetic adaptations to high altitude in Tibet. *Nat. Commun.* **5**, 3281 (2014).
107. Bittles, A. H. & Black, M. L. Evolution in health and medicine Sackler colloquium: consanguinity, human evolution, and complex diseases. *Proc. Natl Acad. Sci. USA* **107** (Suppl. 1), 1779–1786 (2010).
108. Weatherall, D. J. The inherited diseases of hemoglobin are an emerging global health burden. *Blood* **115**, 4331–4336 (2010).
109. Lyons, E. J., Frodsham, A. J., Zhang, L., Hill, A. V. & Amos, W. Consanguinity and susceptibility to infectious diseases in humans. *Biol. Lett.* **5**, 574–576 (2009).
110. Rudan, I. et al. Inbreeding and the genetic complexity of human hypertension. *Genetics* **163**, 1011–1021 (2003).
111. Heckerman, D. et al. Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proc. Natl Acad. Sci. USA* **113**, 7377–7382 (2016).
112. Saudi Mendeliome Group. Comprehensive gene panels provide advantages over clinical exome sequencing for Mendelian diseases. *Genome Biol.* **16**, 134 (2015).
113. Asimit, J. L., Hatzikotoulas, K., McCarthy, M., Morris, A. P. & Zeggini, E. Trans-ethnic study design approaches for fine-mapping. *Eur. J. Hum. Genet.* **24**, 1330–1336 (2016).
This study assesses the impact of inclusion of populations of different ancestries on resolution of causal loci and shows that fine-mapping is greatly improved by inclusion of individuals of African ancestry.
114. Evangelou, E. & Ioannidis, J. P. Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* **14**, 379–389 (2013).
115. Hong, J., Lunetta, K. L., Cupples, L. A., Dupuis, J. & Liu, C. T. Evaluation of a two-stage approach in trans- ethnic meta-analysis in genome-wide association studies. *Genet. Epidemiol.* **40**, 284–292 (2016).
116. Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
117. International Consortium for Blood Pressure Genome-Wide Association Studies. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478**, 103–109 (2011).
118. International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
119. Brown, B. C., Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, Ye, C. J., Price, A. L. & Zaitlen, N. Transethnic genetic-correlation estimates from summary statistics. *Am. J. Hum. Genet.* **99**, 76–88 (2016).
120. Marquez-Luna, C., Loh, P. R., South Asian Type 2 Diabetes Consortium, The SIGMA Type 2 Diabetes Consortium & Price, A. L. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.* **41**, 811–823 (2017).
121. Popejoy, A. B. et al. The clinical imperative for inclusivity: race, ethnicity, and ancestry (REA) in genomics. *Hum. Mutat.* **39**, 1713–1720 (2018).
122. Schrijver, I. et al. The spectrum of CFTR variants in nonwhite cystic fibrosis patients: implications for molecular diagnostic testing. *J. Mol. Diagn.* **18**, 39–50 (2016).
123. Rohlfs, E. M. et al. Cystic fibrosis carrier testing in an ethnically diverse US population. *Clin. Chem.* **57**, 841–848 (2011).
124. Wheeler, E. et al. Impact of common genetic determinants of hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: a transethnic genome-wide meta-analysis. *PLOS Med.* **14**, e1002383 (2017).
125. Johnson, J. A. Ethnic differences in cardiovascular drug response: potential contribution of pharmacogenetics. *Circulation* **118**, 1383–1393 (2008).
126. Caraco, Y., Blotnick, S. & Muszkat, M. CYP2C9 genotype-guided warfarin prescribing enhances the efficacy and safety of anticoagulation: a prospective randomized controlled study. *Clin. Pharmacol. Ther.* **83**, 460–470 (2008).
127. H3Africa Consortium. Research capacity. Enabling the genomic revolution in Africa. *Science* **344**, 1346–1348 (2014).
This work is an important article outlining the H3Africa initiative joint funded through the National Institutes of Health–Wellcome to facilitate genomics research in Africa, with a focus on capacity-building.
128. Petrovski, S. & Goldstein, D. B. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biol.* **17**, 157 (2016).
129. Hormozdiari, F. et al. Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat. Genet.* **50**, 1041–1047 (2018).
130. Smith, M. W. & O'Brien, S. J. Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. *Nat. Rev. Genet.* **6**, 623–632 (2005).

131. Kim, Y. & Nielsen, R. Linkage disequilibrium as a signature of selective sweeps. *Genetics* **167**, 1513–1524 (2004).
132. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLOS Biol.* **4**, e72 (2006).
133. Nagai, A. et al. Overview of the BioBank Japan Project: study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).

Acknowledgements
I.B. acknowledges funding from Wellcome (WT206194). M.S. acknowledges funding from the Wellcome Sanger Institute (WT098051), the UK Medical Research Council (G0901213-92157, G0801566 and MR/K013491/1) and the National Institute for Health Research Cambridge Biomedical Research Centre.

Author contributions
D.G. and M.S.S. researched the literature and wrote the manuscript. All authors substantially contributed to discussions of the content, and reviewed and/or edited the manuscript before submission.

Competing interests
The authors declare no competing interests.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Reviewer information
Nature Reviews Genetics thanks H. Hakonarson, T. Manolio and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

RELATED LINKS
China Kadoorie Biobank: http://www.ckbiobank.org/
East London Genes and Health Study: http://www.genesandhealth.org/
Finnish Biobanks: https://www.biopankki.fi/en/finnish-biobanks
gnomAD: https://gnomad.broadinstitute.org/
H3Africa: https://h3africa.org/
Million Veteran Program: https://www.research.va.gov/mvp
NIH All of Us: https://allofus.nih.gov/
Qatar Genomes Project: https://qatargenome.org.qa/
TOPMed Programme: https://www.nhlbi.nih.gov/science/trans-omics-precision-medicine-topmed-program
Web PopGen simulator: https://www.radford.edu/~rsheehy/Gen_flash/popgen