

# Ancestral diversity in complex disease genetics: from discovery to translation

Karoline Kuchenbaecker <sup>1,2,3</sup> ✉ & Georgina Navoly<sup>1</sup>

## Abstract

The expansion of ancestrally diverse genetic cohorts has altered the landscape of complex disease genetics. Differences in linkage disequilibrium across populations have improved fine-mapping and the identification of target genes – key steps for translating findings from genome-wide association studies into biological understanding. Whilst there is widespread sharing of genetic architecture across ancestries, loci that display heterogeneity in causal genetic effects across populations can offer unique biological insights. Here, we review how ancestral and global diversity shape genetic discoveries. As we advance towards global precision medicine, integrating genomic data with diverse environmental and social factors will be crucial to account for population-specific contexts that can influence disease risk or treatment response.

## Sections

Introduction

Human evolution and genetic diversity

From ancestral diversity to genomic discovery

Comparisons of genetic associations across populations

Conclusions and future perspectives

<sup>1</sup>Division of Psychiatry, University College London, London, UK. <sup>2</sup>UCL Genetics Institute, University College London, London, UK. <sup>3</sup>Genomics England, London, UK. ✉ e-mail: [k.kuchenbaecker@ucl.ac.uk](mailto:k.kuchenbaecker@ucl.ac.uk)

## Introduction

Common, complex diseases arise from a combination of genetic and non-genetic risk factors, such as environmental exposures, lifestyle (diet, physical activity, smoking) or socioeconomic factors, that can interact to influence disease susceptibility. Disease prevalence, that is, the proportion of individuals in a population with a disease at a given time, can vary by geographic location. For example, the prevalence of obesity is considerably higher in the USA (42% in adults) than in Japan (5%), largely due to differences in dietary patterns such as higher consumption of sugar-sweetened beverages in the USA<sup>1</sup>, which increases the risk of obesity-associated diseases such as type 2 diabetes and cardiovascular disease<sup>2</sup>. Air pollution and insufficient physical activity are other geographically stratified exposures that have a strong impact across several complex diseases, including coronary artery disease and stroke<sup>3,4</sup>. Prevalence rates for some diseases or specific subtypes can also differ for ethnic groups within a country. For example, in England, overall cancer incidence rates are lower in Black and Asian ethnic groups than in the white ethnic group; however, the incidence rate of prostate cancer is 2.1-fold higher in Black men than in white men<sup>5</sup>.

Over the past two decades, genome-wide association studies (GWAS) have uncovered many genetic variants linked to complex traits or diseases. However, the vast majority (86%) of published studies were based on data for individuals with European ancestries<sup>6</sup>. This Eurocentric bias has ethical, scientific and clinical consequences as the generalizability of genetic findings is uncertain, which could deepen existing health disparities across global populations<sup>7</sup>. Polygenic risk scores (PRS) are an important example that illustrates how a lack of ancestral diversity in genetic studies hinders discovery and inclusive research. PRS are derived by aggregating the effects of genetic variants associated with disease and can estimate part of an individual's susceptibility to diseases<sup>8</sup>. PRS have shown utility in research settings and, despite limited predictive accuracy, are increasingly incorporated into clinical practice, for example, to identify individuals at higher lifetime risk<sup>9–11</sup>. The predictive accuracy of PRS developed using data from individuals of European ancestry can be up to 4.9-fold lower in non-European populations than in European populations, raising concerns that their clinical use may exacerbate disparities in health outcomes<sup>12</sup>.

A lack of ancestral diversity in genetic research also limits the discovery of risk alleles that are common or have stronger effects in specific populations. For example, a damaging missense variant (*rs730881101*) in *TNNT2*, which is associated with lower heart function and increased risk of heart failure, was identified in a GWAS of more than 260,000 Japanese individuals and would have been undetectable in European-ancestry-only cohorts<sup>13</sup>. As another example, an intronic variant (*rs77408001*) in the *ELN* gene, associated with kidney function, was identified in a meta-analysis of approximately 80,000 individuals of African ancestries and provided new insights into the pathogenesis of chronic kidney disease<sup>14</sup>. Addressing imbalanced representation in genetic research is therefore essential to ensuring that advances in genomics benefit all equitably.

The past few years have seen major developments in the genetic and genomic data landscape, with large biobanks emerging outside Western countries as well as large studies focusing on underrepresented groups within Western countries<sup>15–18</sup>. However, recent political developments in the USA pose emerging challenges to equity in biomedical research<sup>19,20</sup>. Reducing funding and support for research focusing on diversity and equity in genomics risks undermining efforts to understand complex diseases in diverse populations<sup>7,21</sup>.

Here, we review how key demographic events and evolutionary forces, including migration, genetic drift and natural selection, contribute to population genetic variation by influencing allele frequency differences, patterns of linkage disequilibrium and population structure. We then discuss how increased ancestral and geographic diversity in genomic resources has contributed to novel genetic discoveries, including population-specific variants, new trait associations and improved fine-mapping resolution. Diverse data resources also offer new insights on the generalizability of genetic findings based largely on one ancestry group across diverse ancestries and settings. Finally, we address how genetic ancestry can influence patterns of disease prevalence.

## Human evolution and genetic diversity

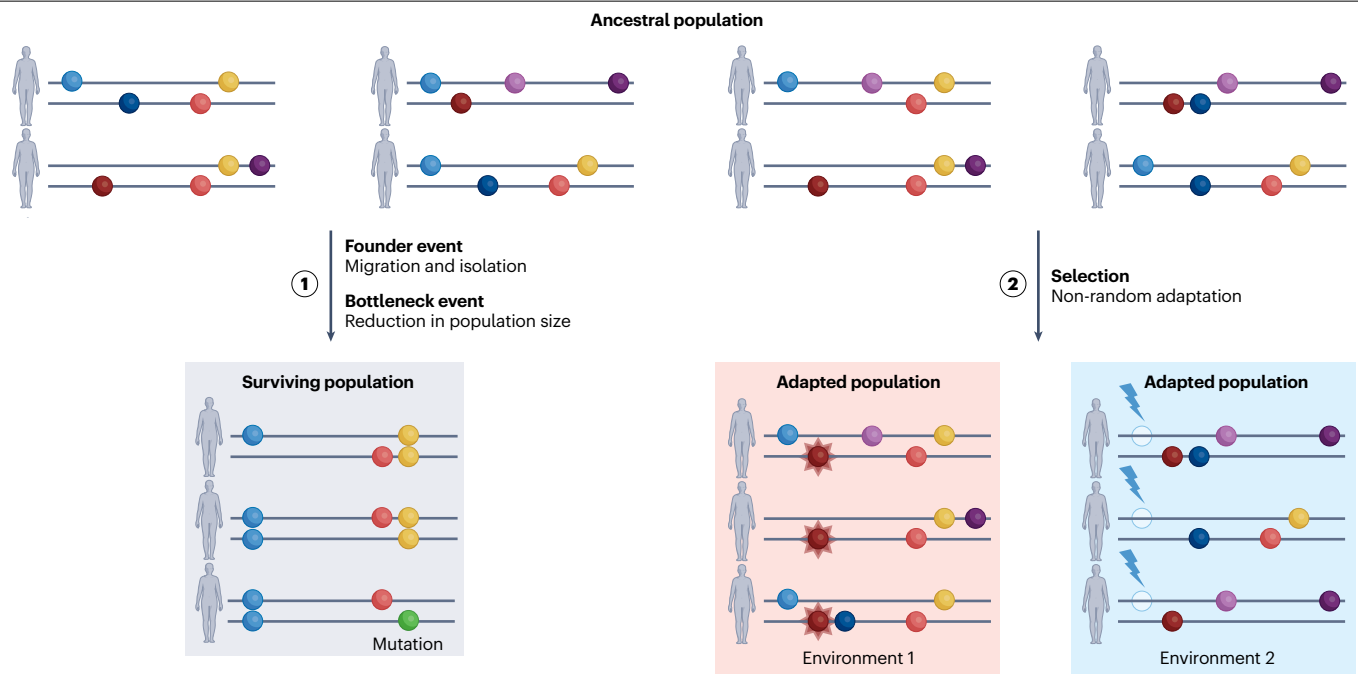
Applying a population genetics framework can help understand observed genetic differences in the context of the demographic events and evolutionary forces that act on allele frequencies over time and across populations<sup>24</sup> (Fig. 1). In this section, we introduce the key concepts of population and ancestry before outlining how demographic events, such as migration and archaic introgression, and evolutionary forces, such as genetic drift and natural selection, contribute to variation and why these patterns are relevant for genetic research.

'Human populations' refers to groups of people who live in a specific geographic area and share certain characteristics<sup>25</sup>, including genetics, location, and demographic and cultural aspects<sup>26</sup>. An individual's 'genetic ancestry' refers to a person's family origins and genetic lineage, tracing back through generations<sup>25</sup>. Genetic ancestry is a concept that is relative to time, geographic context, reference population data and the analytical methods used for ancestry inference<sup>27</sup>. In this Review, we mean 'genetically inferred ancestry' when discussing diversity in genetic research. Of note, when estimated using genetic data, ancestry is a continuous measure shaped by historical patterns of admixture and migration, and no clear line of distinction can be drawn between populations or ancestry groups<sup>25</sup>. When comparing genetic findings across populations or ancestry groups, it is important to carefully disentangle the social and genetic constructs that can influence them.

Most genetic association studies use methods such as principal component analysis to assign individuals to genetically inferred ancestry groups. This means that an individual's sample is assigned to a genetically similar reference group and labelled accordingly. The choice of reference groups and the boundaries for assignment are somewhat arbitrary. We use the term ancestry group throughout this Review to refer to the genetic similarity-based groupings that were used by the primary research studies and country-based ancestry (for example, labelling Biobank Japan as East Asian). However, we acknowledge that assigning individuals to discrete ancestry groups, although useful for comparisons, as done in this Review, oversimplifies the continuous nature of genetic variation<sup>28</sup> and is shaped by social and historical context (Box 1). Moreover, individuals do not derive from a single ancestry in any meaningful sense. Given the long history of race concepts being weaponized, it is crucial to distinguish these ancestry groups used in genetic research from socially constructed racial hierarchies<sup>29</sup>.

## Migration

Africa is the origin of all modern humans<sup>30,31</sup>. Most human genetic variation (about 99.9% at the DNA level) is shared across populations<sup>32</sup>. Approximately 50,000–100,000 years ago, modern humans began dispersing from Africa<sup>33,34</sup>, resulting in a severe population bottleneck (a sharp reduction in population size<sup>35</sup>) that reduced genetic diversity



**Fig. 1 | The impact of evolutionary forces on genetic variation.** **a**, The ancestral population (top) contains a high level of genetic diversity (represented by a wide variety of coloured alleles). During a bottleneck event, a sharp reduction in population size occurs, allowing only a small, non-representative subset of genetic variants to be passed on in the surviving population (1). This results in a population with reduced genetic diversity and increased linkage disequilibrium

across the genome relative to the ancestral population. Under selective pressure (for example, sun, diet adaptation), certain variants that are advantageous increase in frequency, whereas others decrease (2). Consequently, genome-wide association studies in these populations may identify different phenotype–genotype associations (Fig. 3). Alternatively, the two populations can be exposed to different environments, which interact with some genetic factors or epistasis.

in populations outside of Africa (Fig. 1). Subsequent demographic events, such as migration and admixture, including episodes of archaic introgression, whereby ancient humans, such as Neanderthals and Denisovans, interbred with anatomically modern humans<sup>36</sup>, further combined with evolutionary forces, such as mutation, genetic drift, gene flow and natural selection, to shape present-day human genetic diversity<sup>37</sup>. Populations that remained in Africa did not experience the severe bottleneck that reduced genetic diversity in non-African populations and hence have retained greater genetic diversity, which is key to understanding genetic architecture and disease risk<sup>6,38</sup>. Because genetic diversity and demographic history differ among populations, so too do patterns of population structure and linkage disequilibrium (that is, the non-random association of alleles with each other)<sup>39</sup>. Shorter linkage disequilibrium blocks in African populations reflect greater historical recombination and accumulated genetic variation<sup>40</sup>, producing a rich mosaic of haplotypes. By contrast, European and other populations that have experienced bottlenecks exhibit longer linkage disequilibrium blocks, as recombination has had fewer opportunities to reshuffle alleles<sup>41</sup>. Leveraging the greater genetic diversity and shorter linkage disequilibrium blocks in African populations can aid the discovery of novel loci, identify population-specific variants and improve fine-mapping resolution<sup>42,43</sup>.

## Admixture and introgression

Admixture between populations has substantially contributed to present-day human genetic diversity, as groups with different evolutionary histories came into contact and exchanged genetic material.

Although population genetic models often describe human populations as distinct groups, evidence indicates that ancestry is largely continuous, shaped by gradients of genetic variation<sup>44,45</sup>. This means that differences between populations tend to change gradually rather than aligning with discrete boundaries, and models of clearly separated groups are now recognized as oversimplified, given the extensive admixture observed across human history.

As modern humans migrated and expanded across the world, they encountered and admixed with archaic human populations, such as Neanderthals and Denisovans, leaving detectable segments of archaic DNA in non-African populations<sup>46,47</sup>. Approximately 1–3% of the genomes of non-African individuals are of Neanderthal origin, whereas Denisovan introgression is found at higher levels in populations from Oceania and parts of Asia<sup>48</sup>. These introgressed DNA segments from archaic populations can have functional effects on human physiology. For example, variants introgressed from Neanderthals have been associated with regulation of the *OAS1–OAS2–OAS3* gene cluster, which is involved in innate immunity<sup>49</sup>. Moreover, variants introgressed from Denisovans have been found in the *EPAS1* gene, which has contributed to high-altitude adaptation in Tibetans by regulating the body's response to low oxygen levels<sup>50</sup>. Archaic introgression thus represents a specific form of admixture that has shaped genetic diversity in modern humans.

## Genetic drift

One of the key evolutionary forces to consider in studies of different ancestry groups is genetic drift. Because only a subset of the alleles present in one generation is randomly passed on to the next, genetic drift leads

## Box 1 | Ancestry research in the shadow of racism

Ancestry differences in disease prevalence raise the question of whether genetics contribute to them. However, correlating disease with ancestry is often confounded because of the connection between ancestry and ethnicity or race and the myriad of differences in cultural practices as well as systemic bias, which can affect disease risk. In general, the validity and justification of group comparisons need to be considered with extreme caution. Historically, there have been repeated attempts to explain observed differences in complex traits between ‘races’ through genetics to confirm racist stereotypes<sup>21</sup>. The complexity of these group comparisons with both known and hidden biases and the pervasive influence of the exposome on complex traits lead to a high risk of incorrect conclusions, and there is also a substantial risk of intentional misappropriation of any comparative genetic research<sup>22,23</sup>.

It is important to acknowledge that current research practices often assign individuals to discrete ancestry groups — an inherently imperfect approach given the continuous nature of human genetic variation<sup>29</sup>. While these groupings serve practical purposes in assessing representation and controlling for population stratification, they are not independent of social and historical context, which influence the number and choice of reference populations and the decision of the scale at which to consider population structure. Therefore, the use of genetic ancestry groups in research risks reinforcing problematic

concepts<sup>25</sup>. The historical weaponization of ‘biological race’ concepts across colonial contexts reminds us that genetic ancestry categories, while sometimes scientifically useful, must not be conflated with socially constructed racial hierarchies<sup>159</sup>. These discredited notions continue to have a pervasive influence on people’s lives, including in the context of health-care systems<sup>160</sup>. Consequently, many observable group differences reflect the enduring legacy of racism rather than biological reality<sup>161</sup>. The historical and present-day misuse of genetic differences to justify racial hierarchies demands heightened vigilance.

Recent political developments in the USA, including legislation such as Florida’s ‘Stop WOKE Act’, which prohibits discussions of systemic racism or diversity in federally funded projects<sup>162</sup>, and executive orders that directly impede research into health disparities, exemplify the ongoing risks of conflating ancestry research with ideological agendas. Concurrently, funding for studies addressing equity and diversity has faced targeted cuts<sup>163–165</sup>, reflecting a broader backlash against equity-focused science<sup>166</sup>. These policies mirror historical efforts to suppress research that challenges racial hierarchies, underscoring the need for vigilance in defending rigorous, inclusive genomics.

to fluctuations in allele frequencies (that is, specific genetic variants become more or less common) by chance rather than selection, making it one of the main contributors to allele frequency differences and genetic diversity between populations<sup>51</sup>. Small populations, such as those formed after bottleneck or founder events, are especially susceptible to genetic drift, leading to rapid and sometimes unpredictable shifts in allele frequencies<sup>52</sup>.

Although the majority of genetic variation is shared across populations, some variants are population-specific (that is, only present or common in certain populations), and these differences can be medically relevant<sup>53–55</sup>. For example, isolated populations are communities founded by a small number of individuals who have remained relatively separate from other populations over generations, often due to geographical, cultural or other barriers<sup>56</sup>. As a result, rare variants can drift to higher frequencies, increasing the power to detect associations with complex traits (Fig. 1), or be lost entirely, with a potential impact on disease risk<sup>37,38</sup>. For example, a null mutation in *APOC3* linked to favourable lipid profiles was first discovered in the Amish population<sup>57</sup> and replicated in a Greek isolate population<sup>58</sup>.

### Natural selection

Natural selection ‘favours’ variants that increase survival or reproduction, thereby shaping the frequency of these alleles<sup>59</sup>. Adaptation to local environments has played a role in shaping genomic differences between populations (Fig. 1). For example, the sickle-cell trait, caused by inheriting one copy of a point-mutated allele of the *HBB* gene, provides malaria resistance, offering a survival advantage in malaria-endemic regions of Africa. This selective pressure has led to a higher frequency of the sickle-cell allele, and consequently sickle-cell disorder, in populations historically exposed to malaria, particularly in parts of sub-Saharan Africa, the Middle East and India<sup>60,61</sup>. The survival advantage helps maintain the gene in the population despite

the severe health risks associated with inheriting two copies, serving as an example where natural selection has driven population-specific allele frequency changes<sup>62</sup>. Of note, although natural selection acts on genetic variants, it is often in response to environmental exposures, including diet, pollution, infections and stress<sup>63,64</sup>.

Polygenic adaptation to natural selection involves subtle allele frequency changes for a large number of variants, each with a small effect on the outcome, and has shaped the genetic architecture of complex diseases<sup>65,66</sup>. Polygenic adaptation to local selection pressures could explain why some genetic associations with complex diseases are population-specific. However, it has been demonstrated that subtle population stratification can influence estimates of polygenic signatures of selection, making it challenging to quantify the impact of selection on complex traits<sup>67,68</sup>. Moreover, the interaction of the exposome (totality of exposures) with the genome through gene–environment interactions can influence health and disease, which may affect populations differently owing to variations in environmental exposures, socioeconomic conditions and geography<sup>69</sup>.

### From ancestral diversity to genomic discovery

Recognizing the importance of global representation in genetic and genomic research, recent efforts have focused on increasing ancestral diversity in large-scale genetic studies. The following sections highlight how ancestrally diverse biobanks have emerged across many regions worldwide and how this representation is reshaping genetic discovery by expanding the catalogue of genetic variants, enabling the discovery of novel trait associations and empowering fine-mapping to identify target genes, revealing new biological insights.

### Diversity in biobanks

We first describe developments in terms of global biobanks, followed by introducing some notable disease-focused cohorts. We then present

an overview of cohorts that aim to address the underrepresentation of minoritized ethnic groups in Western countries.

The field of complex trait genetics has been revolutionized by one data resource, the UK Biobank, due to its easy accessibility coupled with deep phenotyping and large sample size. About 30,000 study participants of the 500,000 cohort have some non-European ancestries. However, most of the research based on the UK Biobank has excluded the data from these individuals<sup>27</sup>.

Some of the first large biobanks outside Western countries were established in China and Japan, and these biobanks have now been used to publish extensive work and included in multi-ancestry GWAS (Table 1). Biobank Japan includes DNA samples from 270,000 individuals, focusing on 47 target diseases<sup>16</sup>. The China Kadoorie Biobank recruited 512,000 participants from 10 geographically diverse regions in China, collecting DNA, serum and electronic health-care records (EHR) to study chronic diseases and their risk factors<sup>17</sup>. These resources have used consistent phenotyping, linked EHRs and applied state-of-the-art analyses in large non-European populations, thereby enabling considerable new insights. Other global biobanks, such as the Qatar Biobank, are following suit and are starting to yield new insights into disease genetics<sup>70</sup>.

Whilst large biobanks are not yet available for all global regions, there are noteworthy efforts focused on specific disease groups. One of the advantages of case–control designs is that they can provide much better power to study the genetics of the disease of interest compared with national biobanks or community-based recruitment. When a disease is relatively rare, only dedicated case recruitment can yield sufficiently large numbers. Moreover, severely ill participants are often underrepresented in biobanks, leading to ascertainment bias. In continental Africa, notable studies include the Africa Wits-INDEPTH Partnership for Genomic Studies (AWI-Gen) study, which has collected DNA and phenotypic data to study cardiometabolic disease in 12,000 participants from Ghana, Burkina Faso, Kenya and South Africa<sup>71</sup>. AWI-Gen, which launched in 2012, is part of the pan-African initiative Human Heredity and Health in Africa (H3Africa), which supports large-scale genomics research led by African scientists. NeuroGAP, also part of H3Africa, has enrolled almost 43,000 participants to investigate the genetic basis of neuropsychiatric disorders across countries in Africa using DNA and clinical interviews for mental health phenotyping<sup>72</sup>. Over 2 billion people live in South Asia; yet, this group is currently severely underrepresented in genetic research. An example of emerging large-scale initiatives to change this is the Pakistan Alliance on Genetic Risk Factors for Health (PARKH), which includes three sister studies: DIVERGE<sup>73</sup>, focused on major depression; Gen-SCRIP, on schizophrenia; and Gen-BLIP, on bipolar disorder<sup>74,75</sup>. PARKH has recruited approximately 28,000 patients with severe mental illness and 14,000 controls.

Another development has been the improved representation of ancestrally diverse participants in biobanks in high-income countries (Table 1). To empower meaningful research, it is usually necessary to recruit minority groups at a higher proportion compared with their representation in the general population. The All of Us research programme was a trailblazer for representation and engagement of communities that are historically underrepresented in biomedical research<sup>15</sup>. The initiative aims to enrol over 1 million participants across the USA, collecting DNA, serum, EHRs, surveys and data from wearables to study a wide range of diseases<sup>15</sup>. Genes & Health is a large community-based, long-term study recruiting 100,000 British Bangladeshi and British Pakistani people<sup>18</sup>. Genes & Health has recognized the increased burden of cardiometabolic diseases in British South

Asians (Box 2) and is releasing advanced omics datasets on subsets of individuals, enabling deep multi-omic analyses of health and disease.

Finally, the Global Biobank Meta-analysis Initiative (GBMI) is a global network of 29 biobanks and initiatives representing diverse origins and ancestries across four continents. GBMI comprises data from over 2.2 million individuals with genetic and EHR data, aiming to harmonize and integrate data from diverse biobanks<sup>76,77</sup>.

## Diversity-driven discovery

By encompassing greater ancestral diversity, these new data resources are expected to capture a broader spectrum of genomic variation compared with studies limited to participants of European ancestries, thereby expanding the set of genetic variants that can be tested in association studies. Here, we examine whether this increased diversity has, in practice, enabled the discovery of novel variants and trait associations. We then discuss the identification of causal variants and target genes by fine-mapping, one of the main challenges after locus discovery.

**Novel variants.** Studies with ancestrally diverse participants have markedly expanded the genomic variation that genetic association studies are able to assess. For example, the increased genomic diversity of the non-European samples in GBMI enabled researchers to analyse an additional 21.8 million genetic variants that were not present in the 1.4 million biobank participants with European ancestries<sup>76</sup>. Of these variants, 3.4 million were common in at least one non-European-ancestry group. Using whole-genome sequencing, the All of Us study found 275 million previously unreported genetic variants, including 3.9 million coding variants<sup>15</sup>. In the Million Veteran Program (MVP), over half of the variants analysed in a multi-ancestry meta-analysis were not present in a GWAS restricted to European ancestries<sup>78</sup>. Many of these variants are unique to certain populations. The immense genomic diversity present in African populations makes this group particularly promising to identify population-specific risk alleles<sup>79</sup>. However, a limitation is that most large cohorts used DNA microarrays for genotyping, which do not capture all of the genomic variation present in diverse populations either because they were designed primarily for populations with European ancestries or because they do not have sufficient numbers of variants on them to capture the variation in highly diverse populations, for example, from Africa<sup>80</sup>. Whole-genome sequencing avoids these biases, and its falling costs may make its use increasingly feasible across diverse populations<sup>81</sup>. To improve coverage of genetic variation, studies based on DNA microarrays use genotype imputation where variants that are missing on the array are stochastically inferred based on complete genomic information from a reference data set with whole-genome sequencing. Studies have shown that including ancestry-matched samples in the imputation reference panel improves both coverage and imputation quality, especially for low-frequency and rare-frequency variants<sup>82</sup>. When public reference panels do not adequately represent the ancestry or genetic diversity of a target group, a useful strategy has been to sequence a small subset of individuals from the study population to build a custom reference panel, which can then be used for improved imputation across the rest of the sample<sup>79</sup>.

**Novel trait associations.** Variants unique to certain populations and alleles with higher frequencies in a given ancestry group can increase statistical power to discover novel trait associations. A comprehensive assessment of the additional discoveries across a wide array of phenotypes has been implemented by the Pan-UKB project<sup>83</sup>. Overall,

# Review article

**Table 1 | Notable global or ancestrally diverse genetics data resources**

Study	Country	Acestry	Current enrolment <sup>a</sup>	Data access <sup>b</sup>	Cohort design	Official website	Refs.
<b>Global biobanks and general population cohorts</b>							
Biobank Japan	Japan	East Asian	270k	Yes	PDC	<a href="https://biobankjp.org/en">https://biobankjp.org/en</a>	115
Tohoku Medical Megabank	Japan	East Asian	150k	Yes	CBC	<a href="https://www.megabank.tohoku.ac.jp/english/">https://www.megabank.tohoku.ac.jp/english/</a>	116
China Kadoorie Biobank	China	East Asian	512k	Yes	NBR	<a href="https://www.ckbiobank.org">https://www.ckbiobank.org</a>	17
Chinese Millionome Database	China	East Asian	141k	NA	RFR	<a href="https://cmdb.bgi.com/">https://cmdb.bgi.com/</a>	117
Taiwan Biobank	Taiwan	East Asian	200k	Yes	NBR	<a href="https://www.twbiobank.org.tw/">https://www.twbiobank.org.tw/</a>	118
Taiwan Precision Medicine Initiative	Taiwan	East Asian	500k	Yes	NBR	<a href="https://tpmi.ibms.sinica.edu.tw">https://tpmi.ibms.sinica.edu.tw</a>	119
Qatar Biobank	Qatar	Middle Eastern	40k	Yes	NBR	<a href="https://www.qphi.org.qa">https://www.qphi.org.qa</a>	70
Genome India	India	South Asian	20k	Yes	NBR	<a href="https://genomeindia.in">https://genomeindia.in</a>	120
BELIEVE	Bangladesh	South Asian	75k	Yes	CBC	<a href="https://www.believestudy-bangladesh.org">https://www.believestudy-bangladesh.org</a>	121
The Egypt Genome Project	Egypt	North African	NA	NA	NBR	<a href="https://egp.sci.eg">https://egp.sci.eg</a>	122
Uganda Genome Resource	Uganda	African	22k	Yes	CBC	NA	123,124
H3Africa	Multiple African countries	African	Multiple studies	Yes	PDC	<a href="https://h3africa.org">https://h3africa.org</a>	125,126
Mexico City Prospective Study	Mexico	Latin American	150k	Yes	CBC	<a href="https://datashare.ndph.ox.ac.uk/mexico/">https://datashare.ndph.ox.ac.uk/mexico/</a>	127
Estonian Biobank	Estonia	European	212k	Yes	NBR	<a href="https://genomics.ut.ee/en/content/estonian-biobank">https://genomics.ut.ee/en/content/estonian-biobank</a>	128,129
FinnGen	Finland	European	500k	Yes	NBR	<a href="https://www.finnngen.fi/en">https://www.finnngen.fi/en</a>	130
UK Biobank	UK	Diverse	500k	Yes	NBR	<a href="https://www.ukbiobank.ac.uk">https://www.ukbiobank.ac.uk</a>	131
Genomics England	UK	Diverse	100k	Yes	PDC	<a href="https://www.genomicsengland.co.uk/">https://www.genomicsengland.co.uk/</a>	132
<b>Disease-specific cohorts</b>							
AWI-Gen	Multiple African countries	African	12k	Yes (part of H3Africa)	PDC	<a href="https://h3africa.org/index.php/awi-gen/">https://h3africa.org/index.php/awi-gen/</a>	71,104
NeuroGAP	Multiple African countries	African	43k	No	PDC	<a href="https://www.broadinstitute.org/stanley-center-psychiatric-research/stanley-global-neuropsychiatric-genetics-african-populations-neurogap">https://www.broadinstitute.org/stanley-center-psychiatric-research/stanley-global-neuropsychiatric-genetics-african-populations-neurogap</a>	72,133
PARKH	Pakistan	South Asian	42k	No	PDC	<a href="https://www.genes-and-mental-illness.com/">https://www.genes-and-mental-illness.com/</a> <a href="https://lird.org/projects/the-gen-scrip-study/">https://lird.org/projects/the-gen-scrip-study/</a>	73
PROMIS	Pakistan	South Asian	40k	No	PDC	<a href="https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000917.v1.p1">https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000917.v1.p1</a>	134,135
<b>Resources addressing underrepresentation in Western countries</b>							
Our Future Health	UK	Diverse	>2.5 million	Yes	NBR	<a href="https://ourfuturehealth.org.uk">https://ourfuturehealth.org.uk</a>	136
Genes & Health	UK	South Asian	100k	Yes	CBC	<a href="https://www.genesandhealth.org">https://www.genesandhealth.org</a>	18,101
All of Us	USA	Diverse	>867k	Yes	NBR	<a href="https://allofus.nih.gov">https://allofus.nih.gov</a> <a href="https://www.researchallofus.org">https://www.researchallofus.org</a>	15,137
Million Veteran Program	USA	Diverse	1 million	Yes	CBC	<a href="https://www.mvp.va.gov/pwa/">https://www.mvp.va.gov/pwa/</a>	78
PAGE	USA	Diverse	NA	Yes	CBC	<a href="https://www.pagestudy.org">https://www.pagestudy.org</a>	138

**Table 1 (continued) | Notable global or ancestrally diverse genetics data resources**

Study	Country	Acestry	Current enrolment <sup>a</sup>	Data access <sup>b</sup>	Cohort design	Official website	Refs.
<b>Resources addressing underrepresentation in Western countries (continued)</b>							
Michigan Genomics Initiative	USA	Diverse	>90k	NA	PDC	<a href="https://aidhi.umich.edu/mgi-community">https://aidhi.umich.edu/mgi-community</a>	139
HCHS/SOL	USA	Hispanic	16k	Yes	CBC	<a href="https://sites.csc.unc.edu/hchs">https://sites.csc.unc.edu/hchs</a>	140,141
TopMed	USA	Diverse	206k	Yes	CBC PDC	<a href="https://topmed.nhlbi.nih.gov">https://topmed.nhlbi.nih.gov</a>	142
23 and Me	USA	Diverse	15 million	Yes	Commercial	<a href="https://blog.23andme.com/tag/gwas">https://blog.23andme.com/tag/gwas</a> <a href="https://research.23andme.com/research-innovation-collaborations/">https://research.23andme.com/research-innovation-collaborations/</a>	NA
BioVU	USA	Diverse	300k	Yes	PDC	<a href="https://vict.vumc.org/biovu-description/">https://vict.vumc.org/biovu-description/</a>	143
BioME	USA	Diverse	>57k	Yes	PDC	<a href="https://icahn.mssm.edu/research/ipm/programs/biome-biobank">https://icahn.mssm.edu/research/ipm/programs/biome-biobank</a>	NA
CanPath	Canada	Diverse	330k	Yes	NBR	<a href="https://canpath.ca/cohort/ontario-health-study/">https://canpath.ca/cohort/ontario-health-study/</a>	144

AWI-Gen, Africa Wits-INDEPTH Partnership for Genomic Studies; CBC, community-based cohort; H3Africa, Human Heredity and Health in Africa; NA, not available; NBR, national biobank or population registry linkage; PARKH, Pakistan Alliance on Genetic Risk Factors for Health; PDC, patient-based/disease-oriented cohort; RFR, reference genome or frequency resource.

<sup>a</sup>Enrolment as of September 2025. <sup>b</sup>Standardized procedures in place (Yes/No).

this study found 237,360 independent associations across 431 phenotypes. Out of these, 14,676 loci (6.2%) were uncovered by including non-European ancestries from the UK Biobank (4.5% of the samples)<sup>83</sup> (Fig. 2). A total of 1,270 traits were assessed in the MVP cohort, which identified 26,049 significant variant–trait associations, out of which 3,477 were identified when individuals of non-European ancestries were included<sup>78</sup>. On a cross-study level, the GBMI identified 163 associations across 14 diseases in addition to the 337 loci identified in the analysis restricted to European-ancestry participants<sup>76</sup>.

Many of the novel associations found in diverse populations have lead variants, also known as sentinel variants, with lower allele frequencies in European ancestries (Fig. 2), suggesting that ancestral diversity played an important role in these discoveries.

**Fine-mapping.** Although genome-wide studies in diverse populations have uncovered many novel associations, identifying the specific causal variants in a trait-associated locus is difficult. Due to linkage disequilibrium across the genome, many variants in a GWAS-nominated region can appear statistically associated with a trait, even if only one or a few of them are causally linked to the trait or disease. Without knowledge of the causal variants, the target genes of these associations remain elusive, given that most associations in GWAS are found outside of protein-coding genome regions<sup>84</sup>. Linkage disequilibrium differences between ancestry groups are predicted to improve our ability to pinpoint the causal variants at a GWAS locus<sup>85,86</sup>, a process known as fine-mapping<sup>87</sup>.

If a causal variant is the same in both groups – a common assumption – then it should be present in the intersection of the two lists of potential causal variants, known as the credible sets, from both groups. By contrast, non-causal variants with small *P* values can show heterogeneity in associations across ancestry groups. For example, a locus near *CHRM3* associated with schizophrenia risk contained seven plausible variants when fine-mapped using European-ancestry data alone. When East Asian data were included, a single variant (*rs11587347*) showed a strong association in both populations, whereas the other six variants remained equally associated in individuals of European ancestries but not in those of East Asian ancestries<sup>88</sup>.

The smaller average linkage disequilibrium blocks in individuals of African ancestries are particularly beneficial for fine-mapping because fewer variants that correlate with the causal variant imply smaller credible sets. For example, in a sample-size-matched comparison, fine-mapping using data from individuals of African ancestries yielded significantly smaller credible sets compared with individuals of European ancestries in the MVP (Wilcoxon signed-rank  $P = 3.8 \times 10^{-52}$ )<sup>78</sup>.

Several fine-mapping methods have been developed to carry out multi-ancestry fine-mapping, for example, MESuSiE and MGFlashFM<sup>89,90</sup> (Table 2). When comparing fine-mapping using European-ancestry data and multi-ancestry data for major depression and bipolar disorder<sup>43,91</sup>, the inclusion of multi-ancestry data demonstrates improved fine-mapping resolution with smaller credible sets for both disorders. For example, the median size of the 99% credible sets was reduced from 66 to 30 variants for major depression. Similarly, a multi-ancestry effort for fine-mapping breast cancer loci reduced the size of credible sets by half, from 12 to 6 variants, compared with a previous study using only European-ancestry data<sup>92</sup>. At one locus on 17q21.31, the credible set was reduced from 2,277 credible causal variants to two variants. Overall, out of 332 association signals, 50 of the credible sets contained a single putatively causal variant. However, we acknowledge that none of these comparisons accounts for the fact that the multi-ancestry fine-mapping also featured a larger sample size compared to the European-ancestry-only GWAS, which can contribute to the reduced size of the credible sets.

**Novel biological insights.** Novel trait associations identified in diverse ancestries lead to new biological insights and have real-world implications for health and clinical management. For example, research using data from the China Kadoorie Biobank identified an association of variants in *MYLK4* with cholecystitis. The gene has been linked to serum alkaline phosphatase, which is an established marker of bile duct stones and acute cholecystitis<sup>17</sup>. Biobank Japan uncovered three disease-linked missense variants that were specific to East Asians<sup>93</sup>. One of these variants was associated with lung cancer and was located in *POT1*, which encodes a protein that regulates telomere length, further strengthening the evidence that telomere dysregulation is pathogenic for lung cancer and other cancers. Participants with African ancestries

## Box 2 | Type 2 diabetes in British South Asians

Individuals belonging to South Asian communities in the UK have two to four times higher rates of type 2 diabetes mellitus compared with white British individuals<sup>167</sup>. The Genes & Health study found a trans-ancestry genetic correlation of 0.68 ( $r_g$ ) and a power-adjusted transferability ratio of 0.75 (ref. 168) when assessing type 2 diabetes associations from previously published genome-wide association studies in European ancestries with British Pakistani and Bangladeshi people, suggesting that some of the genetic effects are population-specific. To investigate the causes of these differences, Farmaki et al. compared diabetes risk between first-generation and second-generation migrants as well as individuals with South Asian, European and mixed ancestry<sup>169</sup>. Prevalence was 20% lower in

second-generation versus first-generation South Asians, a difference likely driven by environmental factors (see the figure; reference groups have been assigned 100 for comparison). The South Asian–European mixed ethnicity group had a 70% lower risk compared with the South Asian group, which could only partially be explained by differences in socioeconomic deprivation and obesity (see the figure; reference groups have been assigned 100 for comparison). The association between genetic admixture and diabetes risk, once environmental risk factors were accounted for, approached linearity. Whilst this finding points to genetic factors playing a role in the increased risk of diabetes, it remains possible that there are non-genetic confounders that correlate with these ancestry proportions.

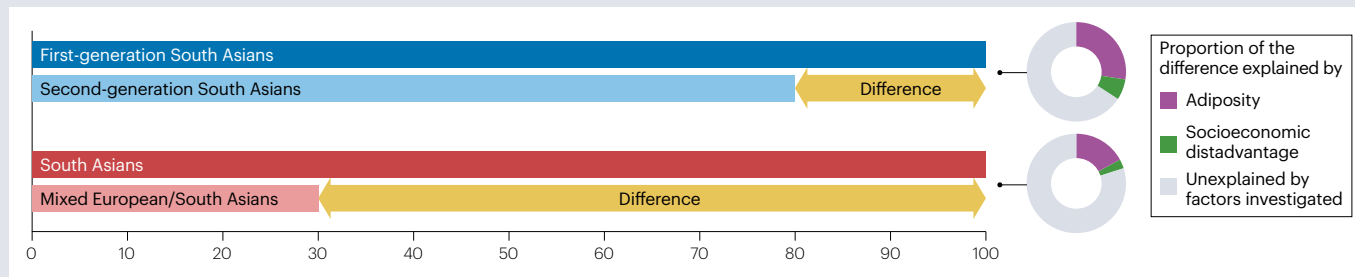


Figure adapted from ref. 169, CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>).

in the MVP enabled the identification of variants near *PCAT2* that were associated with a twofold increased risk of prostate cancer<sup>78</sup>.

### Comparisons of genetic associations across populations

Most genotype–phenotype associations identified to date have been inferred from European-ancestry data, which raises the question of whether these associations are transferable across different ancestries and global settings. Differences in allele frequencies due to genetic drift or different selection pressures can affect the statistical power to observe an association in a given ancestry group. It is also possible for distinct causal variants in one gene or pathway to have shared functional consequences across populations. For example, in East Asian populations, a common missense variant in *ALDH2* markedly reduces enzyme activity, leading to acetaldehyde accumulation, alcohol flushing, and strong protection against heavy drinking and alcohol dependence<sup>94</sup>. By contrast, in European-ancestry groups, a missense variant in the gene *ADH1B* was associated with alcohol use and alcohol use disorders, whereas a distinct missense variant in the same gene was identified in samples with African ancestries<sup>95</sup>. Although these alleles differ by ancestry and gene, they converge on the same metabolic pathway of ethanol oxidation and acetaldehyde clearance, illustrating how different populations can harbour distinct causal variants that exert shared functional effects. Similar patterns were observed at immune loci in a GWAS of rheumatoid arthritis, where ancestry-specific variants converged on the same gene function or pathway<sup>96</sup>.

Genotype–phenotype associations can display heterogeneity across ancestries; that is, a variant can have a stronger or weaker effect size depending on the ancestry group. For example, a multi-ancestry

GWAS of type 2 diabetes found that 40% of genome-wide significant loci showed evidence of heterogeneity across ancestries<sup>97</sup>. One possible explanation for this result could be ancestry differences in linkage disequilibrium (see ‘Transferability of disease loci across populations’). In this scenario, the true causal effect may actually be consistent across ancestries. Alternatively, causal effects can be population-specific, for example, if they are subject to the presence of specific environmental factors (such as an obesogenic environment). Understanding the transferability of genetic associations is vital to ensuring that downstream applications yield equitable benefits. We first discuss research that has aimed to quantify similarity of genetic associations between two ancestry groups across the genome using trans-ancestry genetic correlations. Subsequently, we review ancestry-specific effects at specific loci.

### Genome-wide comparisons of genetic effect sizes

Several methods have been developed to compare effects of genetic variants on complex traits between ancestries (Table 2). Genetic correlation estimates are a widely used tool to assess the similarity of genetic effect sizes between two ancestry groups across the genome. These estimates reflect shared genetic effects with respect to the genetic variation that is captured by the respective genotyping technology of the study, that is, common single-nucleotide polymorphisms for most GWAS. A correlation coefficient of 1 indicates perfect alignment of the genetic effects, whereas estimates closer to 0 highlight heterogeneity in effect sizes between two groups. Popcorn can estimate the genetic correlation between two ancestry groups by modelling the linkage disequilibrium structure<sup>98</sup>. A modified version that accounts for bias where linkage disequilibrium similarity affects the estimates

was developed to better estimate sharing of true causal effect sizes<sup>99</sup>. Another method, S-LDXR, can further estimate enrichment of stratified squared trans-ancestry genetic correlation across functional categories of single-nucleotide polymorphisms<sup>100</sup>.

One of the most extensive efforts to assess the similarity of genetic effects for people of different ancestries in the USA was conducted in the MVP cohort<sup>78</sup> and included data from individuals assigned to African ( $n = 121,177$ ), admixed American ( $n = 59,048$ ), East Asian ( $n = 6,702$ ) and European ( $n = 449,042$ ) ancestry groups for 131 diseases and disease-related binary outcomes. Comparisons of individuals with European versus African ancestries had the greatest statistical power. The median genetic correlation was 0.64 (inter-quartile range (IQR) 0.58–0.70), and most estimates differed significantly from 1. Other studies reported genetic correlations between East Asian-ancestry and European-ancestry GWAS, which were closer to 1. For example, for Biobank Japan, the average genetic correlation with European-ancestry studies across 31 complex diseases and traits was 0.85 (standard error (SE) 0.01)<sup>100</sup>. Few estimates have been reported for South Asians. Comparing British Bangladeshi or Pakistani people from the Genes & Health study with those with European ancestries, the genetic correlation was 0.98 for coronary artery disease and 0.68 (SE 0.15) for type 2 diabetes<sup>101</sup>.

Of note, besides ancestry differences, these genetic correlation estimates can be influenced by differences in study design, for example, when a study in one ancestry group ascertained disease status using EHR data within a specific health-care system and another study used a case–control design<sup>102</sup>. Using meta-analysed data to combine studies with different designs can be a challenge in this context.

In summary, recent large studies report relatively high trans-ancestry genetic correlations for several diseases when comparing GWAS from East Asia with those of European ancestries. However, trans-ancestry correlations were lower for some traits, and for individuals of African and European ancestries in the USA the genetic correlations for more than 100 disease outcomes differed significantly from 1. Overall, there is evidence for both widespread sharing of genetic effects and some heterogeneity.

## Cross-population heterogeneity in genetic effects

Given the evidence from genome-wide comparisons of genetic effects that there can be heterogeneity, it is important to assess whether the association of a given locus is transferable to other ancestry groups. For example, investigators carried out ancestry-specific genome-wide association analyses of 843 traits in the UK Biobank, followed by meta-analysis, yielding 14,438,869 genome-wide significant associations; they identified 20,287 associations in 82 independent loci that demonstrated heterogeneity across ancestry groups<sup>103</sup>. Similarly, the All of Us study identified several associations that were specific to some ancestries<sup>15</sup>. Whilst they replicated associations of the *HLA-DQB1* locus with increased risk of type 1 diabetes across ancestries, the association of this locus with increased risk of coeliac disease was only observed in individuals with European ancestries. Moreover, an association at the *TCF7L2* locus with type 2 diabetes was not observed in participants with East Asian ancestries. These cross-ancestry differences in associations were not driven by low minor allele frequency. Finally, the MVP study confirmed heterogeneity in the effect of *APOE-e4*, with a 30% lower risk of dementia for carriers in the African-ancestry group compared with the European-ancestry group<sup>78</sup>.

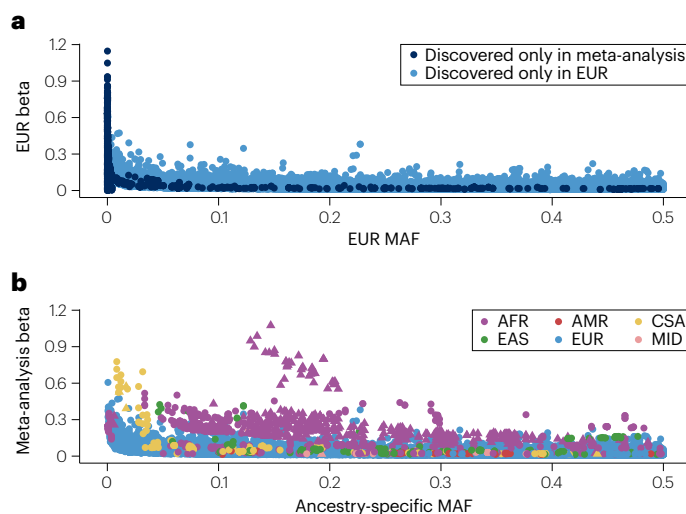
These studies investigated individuals of diverse ancestries from the same country. However, it is vital to consider global cohorts because

different environments can have an impact on genetic effects due to gene–environment interactions. The AWI-Gen study investigated the genetics of hypertension using data from over 10,000 individuals from three countries in sub-Saharan Africa<sup>104</sup>. Two genome-wide significant blood pressure associations were identified that had not been previously found in much larger cohorts comprised mostly of participants of European ancestries, despite one of the lead variants being common across all ancestries. Neither locus demonstrated evidence of association in other cohorts with African ancestries, which suggests that environmental factors, such as diet or other exposures, interact with these variants to influence blood pressure. Similarly, the Qatar Biobank reported a significant association of a locus near *APOBEC3H-CBX7* with type 2 diabetes that did not demonstrate evidence of association in European-ancestry samples despite similar allele frequencies and substantially larger sample sizes<sup>105</sup>.

Variants classified by the Human Gene Mutation Database as disease-causing were assessed in the Uganda Genome Resource. Several variants were unexpectedly common in the Ugandan samples and did not display evidence of association with relevant traits<sup>79</sup>, including variants in or near *LPA* ( $P = 0.40$  for association with total cholesterol), *ADAMTS13* ( $P = 0.90$  for association with platelet count) and *HNF1A*, which has been linked to maturity-onset diabetes of the young type 3 ( $P = 0.20$  for HbA1c level). These findings raise questions about the pathogenicity of these variants in the Ugandan population.

## Transferability of disease loci across populations

Assessing whether a specific disease-linked locus is ‘transferable’, that is, relevant to other ancestries, is challenging. The examples



**Fig. 2 | Differences across ancestries yield novel genetic discoveries.**

**a**, Significantly associated variants for a set of 140 quantitative traits identified in the European genetic ancestry group (blue) versus those discovered only in the multi-ancestry meta-analysis (black) of the Pan-UKB project, with allele frequencies and effect sizes in European ancestries shown. **b**, The same significantly associated variants as shown in part **a** but with ancestry-specific frequencies and effect sizes estimated from the multi-ancestry meta-analysis. Associations on the X chromosome are denoted with triangles. Contrasting parts **a** and **b** highlights the importance of higher allele frequencies in underrepresented ancestry groups for empowering associations. Reproduced from ref. 83, Springer Nature. AFR, African; AMR, Admixed American; CSA, Central or South Asian; EAS, East Asian; EUR, European; MAF, minor allele frequency; MID, Middle Eastern.

**Table 2 | Methods and tools for genetic studies with diverse ancestries**

Method	Type	Description	Ref.
S-LDXR	Genetic correlation	Estimates stratified LD scores to assess genetic correlation	100
Popcorn	Genetic correlation	Estimates transethnic genetic correlation from summary statistics, accounting for LD and allele frequency differences	98
Causal effect size correlation estimator	Genetic correlation	Adjusts observed genetic correlation to estimate the correlation of true causal effect sizes across populations; corrects for differences in LD patterns	99
Unbiased cross-ancestry genetic correlation estimator	Genetic correlation	Uses individual-level data to build ancestry-specific genomic relationship matrices that account for allele frequency and trait architecture differences; produces unbiased estimates	145
LAVA (Local Analysis of Variance Association)	Local genetic correlation	Estimates local genetic correlations using variance association models; works for one ancestry at a time and requires ancestry-specific GWAS summary statistics and corresponding LD panels	146
MR-MEGA	Association testing	Assesses heterogeneity in genetic associations across diverse populations using meta-regression	147
Environment-adjusted meta-regression model (env-MR-MEGA)	Association testing	Accounts for both environmental exposures and genetic ancestry in meta-analyses	148
METASOFT	Association testing	Performs meta-analysis of GWAS using a novel random-effects model (RE2) that avoids overly conservative <i>P</i> values Enhances power to detect associations in the presence of cross-study heterogeneity	149
TRACTOR	Association testing	Local ancestry-aware GWAS tool improving discovery in admixed populations	150
Power-adjusted transferability ratios	Transferability assessment	Aggregates information across loci and accounts for all three factors, sample size, MAF and differences in LD	101
PESCA	Transferability assessment	Identifies population-specific and shared causal variants using GWAS summary statistics and LD	106
msCAVIAR	Fine-mapping	Performs fine-mapping across multiple studies using a random-effects model to account for cross-study heterogeneity	151
MESuSiE	Fine-mapping	Multiple-ancestry extension of the SuSiE model, accounting for shared and ancestry-specific genetic effects (up to two ancestries)	89
MultiSuSiE	Fine-mapping	Multiple-ancestry extension of the SuSiE model	152
SuSiEx	Fine-mapping	Cross-ancestry extension of the SuSiE model	86
MGflashfm	Fine-mapping	Jointly fine-maps signals from multiple traits and population groups	90
TESLA	TWAS	Aggregates GWAS summary statistics, whole-genome sequences and eQTL data from diverse ancestries	153
METRO	TWAS	Incorporates expression prediction models constructed in different genetic ancestries through a likelihood-based inference framework, producing calibrated <i>P</i> values with substantially improved TWAS power	154
MA-FOCUS	TWAS fine-mapping	A multi-ancestry framework leveraging differences in ancestry-specific patterns of LD and eQTL signals Consistently outperforms single-ancestry TWAS fine-mapping approaches	155
PRS-CSx	Polygenic prediction	Improves cross-population polygenic prediction by integrating GWAS summary statistics from multiple populations	156
CT-SLEB	Polygenic prediction	Calculates PRS using summary statistics from multi-ancestry training samples, integrating clumping and thresholding, empirical Bayes, and superlearning	157
PolyPred	Polygenic prediction	Improves cross-population PRS by combining two predictors: a new predictor that leverages functionally informed fine-mapping to estimate causal effects (instead of tagging effects), addressing LD differences; and BOLT-LMM, a published predictor	158

eQTL, expression quantitative trait locus; GWAS, genome-wide association study; LD, linkage disequilibrium; LMM, linear mixed model; MAF, minor allele frequency; PRS, polygenic risk scores; TWAS, transcriptome-wide association study.

discussed so far have tested whether sentinel variants discovered in European-ancestry GWAS show similar effects in other ancestry groups. However, as discussed earlier, sentinel variants are not necessarily

causal and instead may be in linkage disequilibrium with the (possibly unmeasured) true causal variant in the discovery population. If the measured variant is not in linkage disequilibrium with the causal variant

# Review article

in other ancestries, its effect may be reduced or absent, leading to a lack of replication (Fig. 3). Moreover, some studies may not have sufficient power to replicate an association. Most non-European-ancestry studies have a much smaller sample size compared with their respective European-ancestry discovery GWAS, or the allele frequency of a variant may be lower. Methods have been developed to assess locus sharing whilst accounting for these factors (Table 2).

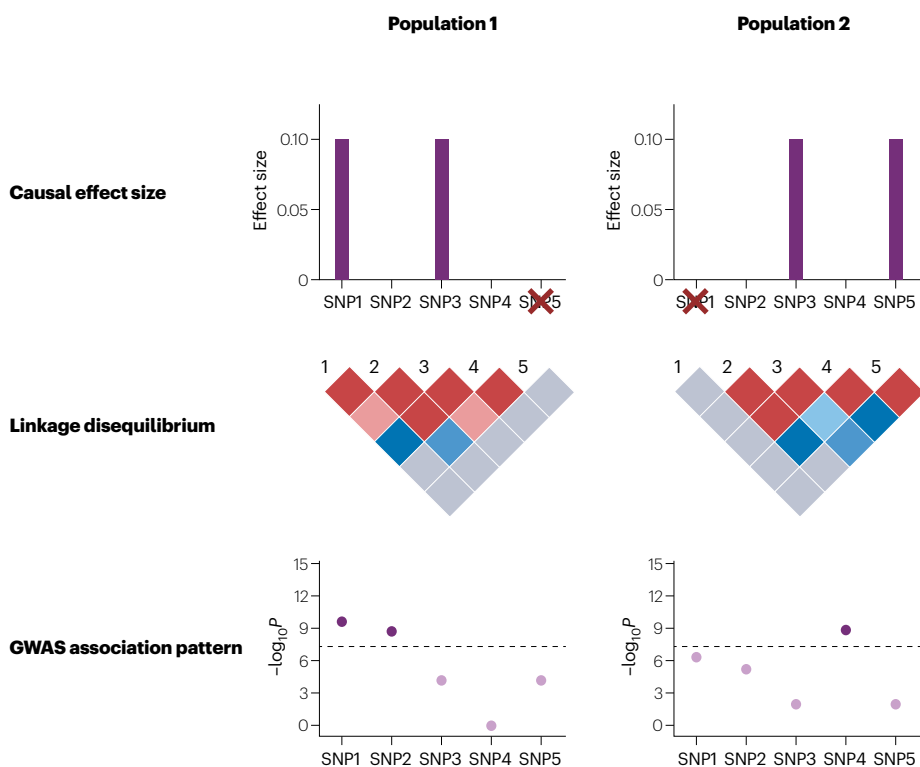
Power-adjusted transferability (PAT) ratios first assess the evidence of transferability for each locus, taking into account linkage disequilibrium<sup>43,101,106</sup>. A ratio of the number of observed transferable loci over the expected number given the statistical power is then computed. For example, a study assessing 71 previously identified loci for coronary artery disease replicated only nine of them in British South Asian individuals from the Genes & Health study<sup>101</sup>. This small number was not explained by differences in allele frequencies, which were similar in both ancestry groups for most loci. However, sample size differences had a great role, with only four loci having power >60% to show significant effects in the Genes & Health study; indeed, three of these loci were amongst the nine that showed evidence of replication. The PAT ratio accounts for replication power by considering the expected number of transferable loci given the power. In this study<sup>101</sup>, 13% of loci showed evidence of replication; however, the study was powered to replicate 21% of loci, yielding a PAT ratio of 0.62. This discrepancy suggests that, even when accounting for power and differences in linkage disequilibrium, one would expect to see more significant associations of coronary artery disease loci in the Genes & Health study. By contrast, for continuous cardiometabolic traits, the study found PAT ratios of 1.

Another study assessed PAT ratios for major depression, which were about 0.30 for African, East and South Asian ancestries<sup>43</sup>. The

genetic architecture of major depression is exceptionally polygenic, that is, involving many variants each exerting a small effect<sup>107</sup>. A complex interplay of the genetic risk factors with geographically or culturally stratified environmental factors, as well as ascertainment differences and global differences in disease diagnosis, is likely to contribute to the low PAT ratios.

Fundamental biology is shared across all humans. Possible reasons why effect sizes can differ nonetheless include heterogeneity in study design or outcome, as well as epistasis, that is, the interaction between different genetic variants. Gene–environment interactions and local polygenic adaptation could also lead to heterogeneity in effect estimates. For example, the All of Us study assessed ancestry-matched replication rates for published variants from the phenotype–genotype reference map in ancestrally diverse participants from the USA<sup>15</sup>. This study found high replication rates ranging from 72% to 100% for all ancestry groups, except for individuals of East Asian ancestries, for whom the replication rate was 46.6%. This group was the only group for which results from USA-based East Asians were compared to GWAS discovery data mostly from East Asian countries. Comparisons for African-ancestry participants of All of Us were based on GWAS findings mostly from other diaspora communities of people with African ancestries in the USA or the UK<sup>108</sup>. Therefore, the low replication rate for East Asians may result from a location mismatch for this group relative to previous GWAS with East Asian samples, and could implicate environmental factors underlying these effects.

Lack of transferability can have important implications, for example, in the context of drug development. The cholesteryl-ester transfer protein, encoded by *CETP*, is essential for reverse cholesterol transport, facilitating the movement of cholesterol from peripheral tissues back to the liver. It mediates the exchange of triglycerides



**Fig. 3 | The impact of ancestral diversity on GWAS associations.** Representation of two ancestral populations, each with a population-specific causal variant. Population 1 has experienced a bottleneck event, leading to reduced genomic diversity (single-nucleotide polymorphism 5 (SNP5) is lost), whilst alleles at other variants have risen in frequency (SNP1). Population 2 has lost the alternative allele of SNP1 as a consequence of adaptation to the local environment. Additionally, due to different linkage disequilibrium patterns in the two populations, a shared causal SNP3 is tagged by different variants, leading to different observed associations: SNP2 is significant in the genome-wide association study (GWAS) in population 1 and SNP4 in population 2. Adapted with permission from ref. 106, Elsevier.

and cholesteryl esters between high-density lipoprotein cholesterol (HDL-C) and apolipoprotein-B-containing lipoproteins, such as low-density lipoprotein cholesterol (LDL-C). Given its influence on HDL-C and LDL-C levels, researchers have extensively explored CETP inhibitors as potential therapeutics to lower the risk of coronary heart disease<sup>109</sup>. GWAS then uncovered that the LDL-C association observed in European ancestries was not transferable to people of South Asian and East Asian ancestries<sup>101,110</sup>. This finding raised concerns about the cross-population efficacy of CETP inhibitors, especially when a study using data from the China Kadoorie Biobank was unable to replicate the protective effect of lower CETP levels on coronary heart disease that had been reported for a European-ancestry group<sup>110</sup>. However, recent work has confirmed a protective effect for a wider East Asian population group at the global ancestry level<sup>109</sup>. This example highlights the importance of identifying and following up on observed population differences in genetic associations for loci with potential therapeutic implications. The precise mechanisms underlying the different association patterns of *CETP* variants with lipids in European and Asian ancestries are not yet fully understood.

## Ancestry and disease prevalence

Prevalence rates for some complex diseases differ both between countries and among ethnic groups within the same country<sup>5</sup>. Similarly, many disease-linked variants exhibit frequency differences across ancestries, which raises the question of whether prevalence differences are at least partially attributable to genetic factors.

Founder effects have driven up the frequencies of some mutations with large effects in isolated populations. For example, founder mutations in the *BRCA1* and *BRCA2* genes in Ashkenazi Jewish people and several other groups have been well studied. However, the presence of such large-effect variants does not necessarily increase overall disease prevalence in these populations<sup>111</sup>. Examples where genetic differences may contribute to prevalence differences exist but clear causal links are difficult to establish. For example, the MVP study identified a variant in *SLC22A18–SLC22A18AS* associated with keloid scarring<sup>78</sup>. This variant is common in African-ancestry groups and monomorphic in individuals with European ancestries. The authors hypothesize that this variant may contribute to differences in the prevalence rate of keloid scarring, which is 300 times greater in individuals of African ancestries than in those of European ancestries.

As complex diseases are multifactorial and polygenic, assessing the overall impact of genetics requires consideration of multiple genetic variants simultaneously, for example, through PRS, where the number of disease-associated alleles is summed up across genetic variants to form a score. However, relating PRS to prevalence rates is highly problematic owing to ascertainment biases in variant discovery. Even large differences in average PRS across ancestries can be purely artefactual<sup>112</sup>. For example, the average schizophrenia PRS is much lower in people of African ancestries compared with people of European and other ancestries, although the disease prevalence is similar across populations<sup>113</sup>. The variants were discovered and effect sizes estimated in data predominantly from one ancestry group, usually European. When carried forward to another ancestry group, the allele frequencies of these variants are expected to differ in that group due to genetic drift and other population genetic forces, which yields a different mean for the PRS in that group. These variants only explain a small proportion of disease risk (in the discovery sample), and most of them are not causal. Differences in linkage disequilibrium across ancestries mean that these non-causal variants are less predictive of the outcome in

another ancestry group, and therefore differences in average PRS do not translate to differences in disease prevalence. These observations highlight the risk of misinterpreting differences in PRS as evidence of differences in disease susceptibility between populations, which has important implications for both research and clinical translation of PRS.

## Conclusions and future perspectives

The expansion of diverse genetic cohorts has fundamentally altered the landscape of complex disease genetics, helping to address critical gaps in our understanding. These new resources have enabled the study of millions of additional genetic variants, empowering the identification of novel disease associations with important biological implications. The field's next major challenge lies in systematically elucidating the biological mechanisms of these associations, particularly for variants in poorly characterized genomic regions. Here, genetic diversity has proven particularly valuable, as demonstrated by the improved resolution of fine-mapping in diverse cohorts. Emerging tools, such as SuSiEx, MESuSiE and MGFlashFM, are advancing multi-ancestry fine-mapping, creating new opportunities to pinpoint causal variants and their target genes. As most loci identified in GWAS lie outside of coding regions, this effort to identify causal associations needs to go hand in hand with continued work to functionally interrogate gene regulatory effects.

Genetics can contribute to differences in disease prevalence across populations, with implications for risk prediction and stratification. However, the polygenic architecture of complex diseases requires consideration of thousands of variants to estimate the genetic impact on disease risk. Whilst causal variants remain unknown for most loci, differences in linkage disequilibrium patterns, allele frequencies and heterogeneity in causal genetic effects continue to challenge the transferability of PRS<sup>12</sup>. Methodological innovations can address some technical limitations. However, dedicated genetic studies in diverse populations remain essential to fully capture population-specific effects.

Although genetics contributes to disease prevalence, socio-environmental factors often dominate, which underscores the need for researchers to carefully contextualize findings and avoid deterministic narratives through active community engagement embedded in study design. Considerable barriers remain, particularly a lack of large biobank resources in some global regions, including Africa and South Asia. Addressing these gaps is crucial for a comprehensive understanding of complex disease genetics and their environmental interactions.

To fully harness the growing global database, more research with a dedicated focus on the role of ancestry and different environments, as well as their implications for equity in genomic medicine, is urgently needed<sup>114</sup>. This includes studies on the transferability of genetic findings, performance of genomics in risk stratification and the effect of population genetic forces, such as local adaptation, on genetic architectures. As the field advances, four imperatives emerge: translating ancestry-aware discoveries into mechanisms through experimental validation; expanding resources to underrepresented regions; dedicated equity-focused research to ensure precision medicine benefits all; and confronting the historical misuse of genetic differences. The promise of this research lies not merely in cataloguing diversity but in leveraging it to dismantle disparities and deliver on the unmet potential of genomics for global health.

Published online: 06 January 2026

## References

- NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128.9 million children, adolescents, and adults. *Lancet* **390**, 2627–2642 (2017).
- World Health Organization. *World Health Statistics 2024: Monitoring Health for the SDGs, Sustainable Development Goals* (WHO, 2024).
- de Bont, J. et al. Ambient air pollution and cardiovascular diseases: an umbrella review of systematic reviews and meta-analyses. *J. Intern. Med.* **291**, 779–800 (2022).
- Vaduganathan, M., Mensah, G. A., Turco, J. V., Fuster, V. & Roth, G. A. The global burden of cardiovascular diseases and risk: a compass for future health. *J. Am. Coll. Cardiol.* **80**, 2361–2371 (2022).
- Delon, C. et al. Differences in cancer incidence by broad ethnic group in England, 2013–2017. *Br. J. Cancer* **126**, 1765–1773 (2022).
- Fatumo, S. et al. A roadmap to increase diversity in genomic studies. *Nat. Med.* **28**, 243–250 (2022).
- Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
- Lewis, C. M. & Vassos, E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med.* **12**, 44 (2020).
- Wang, Y. et al. Polygenic prediction across populations is influenced by ancestry, genetic architecture, and methodology. *Cell Genom.* **3**, 100408 (2023).
- Kullo, I. J. Clinical use of polygenic risk scores: current status, barriers and future directions. *Nat. Rev. Genet.* <https://doi.org/10.1038/s41576-025-00900-8> (2025).
- Hingorani, A. D. et al. Performance of polygenic risk scores in screening, prediction, and risk stratification: secondary analysis of data in the polygenic score catalog. *BMJ Med.* **2**, e000554 (2023).
- Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
- Koyama, S. et al. Population-specific putative causal variants shape quantitative traits. *Nat. Genet.* **56**, 2027–2035 (2024).
- Kintu, C. et al. Meta-analysis of African ancestry genome-wide association studies identified novel locus and validates multiple loci associated with kidney function. *BMC Genomics* **24**, 496 (2023).
- All of Us Research Program Genomics Investigators. Genomic data in the All of Us research program. *Nature* **627**, 340–346 (2024).
- Nagai, A. et al. Overview of the Biobank Japan project: study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
- Walters, R. G. et al. Genotyping and population characteristics of the China Kadoorie Biobank. *Cell Genom.* **3**, 100361 (2023).
- Finer, S. et al. Cohort profile: East London Genes & Health (ELGH), a community-based population genomics and health study in British Bangladeshi and British Pakistani people. *Int. J. Epidemiol.* **49**, 20–21 (2020).
- Executive Office of the President. Executive Order 14151 — Ending radical and wasteful government DEI programs and preferencing. *Federal Register* <https://public-inspection.federalregister.gov/2025-01953.pdf> (2025).
- Executive Office of the President. Executive Order 14173 — Ending illegal discrimination and restoring merit-based opportunity. *Federal Register* <https://public-inspection.federalregister.gov/2025-02097.pdf> (2025).
- Martschenko, D., Trejo, S. & Domingue, B. W. Genetics and education: recent developments in the context of an ugly history and an uncertain future. *AERA Open* **5**, 233285841881051 (2019).
- Saini, A. *Superior: The Return of Race Science* (HarperCollins, 2021).
- Rutherford, A. *How to Argue with a Racist: History, Science, Race and Reality* (Orion Publishing Group Limited, 2021).
- Wakeley, J. The limits of theoretical population genetics. *Genetics* **169**, 1–7 (2005).
- Lewis, A. C. F. et al. Getting genetic ancestry right for science and society. *Science* **376**, 250–252 (2022).
- Krieger, N. Who and what is a ‘population’? Historical debates, current controversies, and implications for understanding ‘population health’ and rectifying health inequities. *Milbank Q.* **90**, 634–681 (2012).
- Peterson, R. E. et al. Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell* **179**, 589–603 (2019).
- Coop, G. Genetic similarity versus genetic ancestry groups as sample descriptors in human genetics. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2207.11595> (2022).
- National Academies of Sciences, Engineering, and Medicine; Division of Behavioral and Social Sciences and Education; Health and Medicine Division; Committee on Population; Board on Health Sciences Policy; Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research. *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field* (National Academies Press, 2023).
- Tattersall, I. Out of Africa: modern human origins special feature: human origins: out of Africa. *Proc. Natl Acad. Sci. USA* **106**, 16018–16021 (2009).
- Stringer, C. B. The emergence of modern humans. *Sci. Am.* **263**, 98–104 (1990).
- Collins, F. S. & Mansoura, M. K. The human genome project: revealing the shared inheritance of all humankind. *Cancer* **91**, 221–225 (2001).
- Ashraf, Q. & Galor, O. The ‘out of Africa’ hypothesis, human genetic diversity, and comparative economic development. *Am. Econ. Rev.* **103**, 1–46 (2013).
- Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019).
- Amos, W. & Hoffman, J. I. Evidence that two main bottleneck events shaped modern human genetic diversity. *Proc. Biol. Sci.* **277**, 131–137 (2010).
- Nielsen, R. et al. Tracing the peopling of the world through genomics. *Nature* **541**, 302–310 (2017).
- Saeb, A. T. M. & Al-Naqeb, D. The impact of evolutionary driving forces on human complex diseases: a population genetics approach. *Scientifica* **2016**, 2079704 (2016).
- Ashraf, B. & Lawson, D. J. Genetic drift from the out-of-Africa bottleneck leads to biased estimation of genetic architecture and selection. *Eur. J. Hum. Genet.* **29**, 1549–1556 (2021).
- Tiret, L. et al. Heterogeneity of linkage disequilibrium in human genes has implications for association studies of common diseases. *Hum. Mol. Genet.* **11**, 419–429 (2002).
- Campbell, M. C. & Tishkoff, S. A. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu. Rev. Genomics Hum. Genet.* **9**, 403–433 (2008).
- Lucena-Perez, M. et al. Bottleneck-associated changes in the genomic landscape of genetic diversity in wild lynx populations. *Evol. Appl.* **14**, 2664–2679 (2021).
- Kanai, M. et al. Meta-analysis fine-mapping is often miscalibrated at single-variant resolution. *Cell Genom.* **2**, 100210 (2022).
- Meng, X. et al. Multi-ancestry genome-wide association study of major depression aids locus discovery, fine mapping, gene prioritization and causal inference. *Nat. Genet.* **56**, 222–233 (2024).
- Serre, D. & Pääbo, S. Evidence for gradients of human genetic diversity within and among continents. *Genome Res.* **14**, 1679–1685 (2004).
- Patterson, N. et al. Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
- Racimo, F., Marnetto, D. & Huerta-Sánchez, E. Signatures of archaic adaptive introgression in present-day human populations. *Mol. Biol. Evol.* **34**, 296–317 (2017).
- Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S. & Akey, J. M. Analysis of human sequence data reveals two pulses of archaic Denisovan admixture. *Cell* **173**, 53–61.e9 (2018).
- Húsz, M. J., Williams, A. L. & Siepel, A. Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. *PLoS Genet.* **16**, e1008895 (2020).
- Dannemann, M., Prüfer, K. & Kelso, J. Functional implications of Neandertal introgression in modern humans. *Genome Biol.* **18**, 61 (2017).
- Zhang, X. et al. The history and evolution of the Denisovan-EPAS1 haplotype in Tibetans. *Proc. Natl Acad. Sci. USA* **118**, e2020803118 (2021).
- Johnson, O. L., Tobler, R., Schmidt, J. M. & Huber, C. D. Fluctuating selection and the determinants of genetic variation. *Trends Genet.* **39**, 491–504 (2023).
- Panoutsopoulou, K. et al. Genetic characterization of Greek population isolates reveals strong genetic drift at missense and trait-associated variants. *Nat. Commun.* **5**, 5345 (2014).
- Gravel, S. When is selection effective? *Genetics* **203**, 451–462 (2016).
- Maher, M. C., Uricchio, L. H., Torgerson, D. G. & Hernandez, R. D. Population genetics of rare variants and complex diseases. *Hum. Hered.* **74**, 118–128 (2012).
- Hernandez, L. M., Blazer, D. G. & Institute of Medicine (US) Committee on Assessing Interactions Among Social, Behavioral, and Genetic Factors. *Genetics and Health* (National Academies Press, 2006).
- Hatzikotoulas, K., Gilly, A. & Zeggini, E. Using population isolates in genetic association studies. *Brief. Funct. Genomics* **13**, 371–377 (2014).
- 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).
- Tachmazidou, I. et al. A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. *Nat. Commun.* **4**, 2872 (2013).
- Vasseur, E. & Quintana-Murci, L. The impact of natural selection on health and disease: uses of the population genetics approach in humans. *Evol. Appl.* **6**, 596–607 (2013).
- Esoh, K. & Wonkam, A. Evolutionary history of sickle-cell mutation: implications for global genetic medicine. *Hum. Mol. Genet.* **30**, R119–R128 (2021).
- 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).
- Bersaglieri, T. et al. Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74**, 1111–1120 (2004).
- Vermeulen, R., Schymanski, E. L., Barabási, A.-L. & Miller, G. W. The exposome and health: where chemistry meets biology. *Science* **367**, 392–396 (2020).
- Miller, G. W. & Jones, D. P. The nature of nurture: refining the definition of the exposome. *Toxicol. Sci.* **137**, 1–2 (2014).
- Zeng, J. et al. Widespread signatures of natural selection across human complex traits and functional genomic categories. *Nat. Commun.* **12**, 1164 (2021).
- Berg, J. J. & Coop, G. A population genetic signal of polygenic adaptation. *PLoS Genet.* **10**, e1004412 (2014).
- Sohail, M. et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife* **8**, e39702 (2019).
- Berg, J. J. et al. Reduced signal for polygenic adaptation of height in UK Biobank. *eLife* **8**, e39725 (2019).
- Stingone, J. A. et al. Toward greater implementation of the exposome research paradigm within environmental epidemiology. *Annu. Rev. Public Health* **38**, 315–327 (2017).
- Al Thani, A. et al. Qatar Biobank cohort study: study design and first results. *Am. J. Epidemiol.* **188**, 1420–1433 (2019).

71. Tluway, F. et al. Cohort profile: Africa Wits-INDEPTH partnership for genomic studies (AWI-Gen) in four sub-Saharan African countries. *Int. J. Epidemiol.* **54**, dyae173 (2024).
72. Stevenson, A. et al. Neuropsychiatric genetics of African populations-psychosis (NeuroGAP-Psychosis): a case-control study protocol and GWAS in Ethiopia, Kenya, South Africa and Uganda. *BMJ Open* **9**, e025469 (2019).
73. Valkovskaya, M. et al. Study protocol of DIVERGE, the first genetic epidemiological study of major depressive disorder in Pakistan. *Psychiatr. Genet.* **33**, 69–78 (2023).
74. The GEN-BLIP Study. <https://lird.org/projects/the-gen-blip-study/>.
75. The GEN-SCRIP Study. <https://lird.org/projects/the-gen-scrip-study/>.
76. Zhou, W. et al. Global biobank meta-analysis initiative: powering genetic discovery across human disease. *Cell Genom.* **2**, 100192 (2022).
77. GBMI Home. *Global Biobank Meta* <https://www.globalbiobankmeta.org/>.
78. Verma, A. et al. Diversity and scale: genetic architecture of 2068 traits in the VA Million Veteran Program. *Science* **385**, eadj1182 (2024).
79. Gurdasani, D. et al. Uganda genome resource enables insights into population history and genomic discovery in Africa. *Cell* **179**, 984–1002.e36 (2019).
80. Verlouw, J. A. M. et al. A comparison of genotyping arrays. *Eur. J. Hum. Genet.* **29**, 1611–1624 (2021).
81. Hanks, S. C. et al. Extent to which array genotyping and imputation with large reference panels approximate deep whole-genome sequencing. *Am. J. Hum. Genet.* **109**, 1653–1666 (2022).
82. Zhou, W. et al. Improving power of association tests using multiple sets of imputed genotypes from distributed reference panels. *Genet. Epidemiol.* **41**, 744–755 (2017).
83. Karczewski, K. J. et al. Pan-UK Biobank genome-wide association analyses enhance discovery and resolution of ancestry-enriched effects. *Nat. Genet.* **57**, 2408–2417 (2025).
84. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
85. 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).
86. Yuan, K. et al. Fine-mapping across diverse ancestries drives the discovery of putative causal variants underlying human complex traits and diseases. *Nat. Genet.* **56**, 1841–1850 (2024).
87. Li, Z. & Zhou, X. Towards improved fine-mapping of candidate causal variants. *Nat. Rev. Genet.* <https://doi.org/10.1038/s41576-025-00869-4> (2025).
88. Lam, M. et al. Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nat. Genet.* **51**, 1670–1678 (2019).
89. Gao, B. & Zhou, X. MESuSIE enables scalable and powerful multi-ancestry fine-mapping of causal variants in genome-wide association studies. *Nat. Genet.* **56**, 170–179 (2024).
90. Zhou, F. et al. Leveraging information between multiple population groups and traits improves fine-mapping resolution. *Nat. Commun.* **14**, 7279 (2023).
91. O'Connell, K. S. et al. Genomics yields biological and phenotypic insights into bipolar disorder. *Nature* **639**, 968–975 (2025).
92. Jia, G. et al. Refining breast cancer genetic risk and biology through multi-ancestry fine-mapping analyses of 192 risk regions. *Nat. Genet.* **57**, 80–87 (2025).
93. Ishigaki, K. et al. Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. *Nat. Genet.* **52**, 669–679 (2020).
94. Li, D., Zhao, H. & Gelernter, J. Strong protective effect of the aldehyde dehydrogenase gene (ALDH2) 504lys (\*2) allele against alcoholism and alcohol-induced medical diseases in Asians. *Hum. Genet.* **131**, 725–737 (2012).
95. Polimanti, R. & Gelernter, J. ADH1B: From alcoholism, natural selection, and cancer to the human phenome. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **177**, 113–125 (2018).
96. Ishigaki, K. et al. Multi-ancestry genome-wide association analyses identify novel genetic mechanisms in rheumatoid arthritis. *Nat. Genet.* **54**, 1640–1651 (2022).
97. Mahajan, A. et al. Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *Nat. Genet.* **54**, 560–572 (2022).
98. 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).
99. Galinsky, K. J. et al. Estimating cross-population genetic correlations of causal effect sizes: GALINSKY et al. *Genet. Epidemiol.* **43**, 180–188 (2019).
100. Shi, H. et al. Population-specific causal disease effect sizes in functionally important regions impacted by selection. *Nat. Commun.* **12**, 1098 (2021).
101. Huang, Q. Q. et al. Transferability of genetic loci and polygenic scores for cardiometabolic traits in British Pakistani and Bangladeshi individuals. *Nat. Commun.* **13**, 4664 (2022).
102. Cai, N. et al. Minimal phenotyping yields genome-wide association signals of low specificity for major depression. *Nat. Genet.* **52**, 437–447 (2020).
103. De Lillo, A. et al. Cross-ancestry genome-wide association studies identified heterogeneous loci associated with differences of allele frequency and regulome tagging between participants of European descent and other ancestry groups from the UK Biobank. *Hum. Mol. Genet.* **30**, 1457–1467 (2021).
104. Singh, S. et al. Genome-wide association study meta-analysis of blood pressure traits and hypertension in sub-Saharan African populations: an AWI-gen study. *Nat. Commun.* **14**, 8376 (2023).
105. Elashi, A. A. et al. Genome-wide association study and trans-ethnic meta-analysis identify novel susceptibility loci for type 2 diabetes mellitus. *BMC Med. Genomics* **17**, 115 (2024).
106. Shi, H. et al. Localizing components of shared transethnic genetic architecture of complex traits from GWAS summary data. *Am. J. Hum. Genet.* **106**, 805–817 (2020).
107. Zhang, Y. D. et al. Assessment of polygenic architecture and risk prediction based on common variants across fourteen cancers. *Nat. Commun.* **11**, 3353 (2020).
108. Mills, M. C. & Rahal, C. The GWAS diversity monitor tracks diversity by disease in real time. *Nat. Genet.* **52**, 242–243 (2020).
109. Dunca, D. et al. Comparing the effects of CETP in east asian and European ancestries: a Mendelian randomization study. *Nat. Commun.* **15**, 5302 (2024).
110. Millwood, I. Y. et al. Association of CETP gene variants with risk for vascular and nonvascular diseases among Chinese adults. *JAMA Cardiol.* **3**, 34–43 (2018).
111. Neuhausen, S. L. Ethnic differences in cancer risk resulting from genetic variation. *Cancer* **86**, 2575–2582 (1999).
112. Martin, A. R. et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
113. Whiteford, H. A. et al. Global burden of disease attributable to mental and substance use disorders: findings from the global burden of disease study 2010. *Lancet* **382**, 1575–1586 (2013).
114. Lehmann, B. et al. Methodological opportunities in genomic data analysis to advance health equity. *Nat. Rev. Genet.* **26**, 635–649 (2025).
115. Kanai, M. et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).
116. Minegishi, N. et al. Biobank establishment and sample management in the Tohoku medical megabank project. *Tohoku J. Exp. Med.* **248**, 45–55 (2019).
117. Li, Z. et al. CMDB: the comprehensive population genome variation database of China. *Nucleic Acids Res.* **51**, D890–D895 (2023).
118. Feng, Y.-C. A. et al. Taiwan Biobank: a rich biomedical research database of the Taiwanese population. *Cell Genom.* **2**, 100197 (2022).
119. Yang, H.-C. et al. The Taiwan precision medicine initiative: a cohort for large-scale studies. *Genetics* <https://doi.org/10.1038/s41586-025-09680-x> (2024).
120. Bhattacharyya, C. et al. Mapping genetic diversity with the Genomelndia project. *Nat. Genet.* **57**, 767–773 (2025).
121. Chowdhury, R. et al. Cohort profile: the BangladEsh longitudinal investigation of emerging vascular and nonvascular events (BELIEVE) cohort study. *BMJ Open* **15**, e088338 (2025).
122. Elmonem, M. A. et al. The Egypt genome project. *Nat. Genet.* **56**, 1035–1037 (2024).
123. Gurdasani, D. et al. The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**, 327–332 (2015).
124. Fatumo, S. et al. Uganda genome resource: a rich research database for genomic studies of communicable and non-communicable diseases in Africa. *Cell Genom.* **2**, None (2022).
125. Mulder, N. et al. H3Africa: current perspectives. *Pharmgenomics. Pers. Med.* **11**, 59–66 (2018).
126. Choudhury, A. et al. High-depth African genomes inform human migration and health. *Nature* **586**, 741–748 (2020).
127. Ziyatdinov, A. et al. Genotyping, sequencing and analysis of 140,000 adults from Mexico City. *Nature* **622**, 784–793 (2023).
128. Leitsalu, L. et al. Cohort profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int. J. Epidemiol.* **44**, 1137–1147 (2015).
129. Mägi, R. Diverse landscape of genomic research within the Estonian Biobank. *Hum. Mol. Genet.* <https://doi.org/10.1093/hmg/ddaf026> (2025).
130. Kurki, M. I. et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
131. Sudlow, C. et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
132. Samuel, G. N. & Farsides, B. The UK's 100,000 genomes project: manifesting policymakers' expectations. *New Genet. Soc.* **36**, 336–353 (2017).
133. Majara, L. et al. 56. Genome wide association study of schizophrenia in the neuropsychiatric genetics in African population-psychosis (neurogap-psychosis) study. *Eur. Neuropsychopharmacol.* **87**, 79–80 (2024).
134. Ahmad, S. et al. Physical activity, smoking, and genetic predisposition to obesity in people from Pakistan: the PROMIS study. *BMC Med. Genet.* **16**, 114 (2015).
135. Saleheen, D. et al. The Pakistan risk of myocardial infarction study: a resource for the study of genetic, lifestyle and other determinants of myocardial infarction in South Asia. *Eur. J. Epidemiol.* **24**, 329–338 (2009).
136. Cook, M. B. et al. Our future health: a unique global resource for discovery and translational research. *Nat. Med.* **31**, 728–730 (2025).
137. Venner, E. et al. The frequency of pathogenic variation in the All of Us cohort reveals ancestry-driven disparities. *Commun. Biol.* **7**, 174 (2024).
138. Wojcik, G. L. et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
139. Zawistowski, M. et al. The Michigan Genomics Initiative: a biobank linking genotypes and electronic clinical records in Michigan Medicine patients. *Cell Genom.* **3**, 100257 (2023).
140. Gallo, L. C. et al. The hispanic community health study/study of latinos sociocultural ancillary study: sample, design, and procedures. *Ethn. Dis.* **24**, 77–83 (2014).
141. LaVange, L. M. et al. Sample design and cohort selection in the Hispanic community health study/study of Latinos. *Ann. Epidemiol.* **20**, 642–649 (2010).
142. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TopMed program. *Nature* **590**, 290–299 (2021).
143. Roden, D. M. et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin. Pharmacol. Ther.* **84**, 362–369 (2008).

144. Kirsh, V. A. et al. Cohort profile: the Ontario Health Study (OHS). *Int. J. Epidemiol.* **52**, e137–e151 (2023).
145. Momin, M. M. et al. A method for an unbiased estimate of cross-ancestry genetic correlation using individual-level data. *Nat. Commun.* **14**, 722 (2023).
146. Werme, J., van der Sluis, S., Posthuma, D. & de Leeuw, C. A. An integrated framework for local genetic correlation analysis. *Nat. Genet.* **54**, 274–282 (2022).
147. Mägi, 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).
148. Wang, S. et al. Accounting for heterogeneity due to environmental sources in meta-analysis of genome-wide association studies. *Commun. Biol.* **7**, 1512 (2024).
149. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* **88**, 586–598 (2011).
150. Atkinson, E. G. et al. Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nat. Genet.* **53**, 195–204 (2021).
151. LaPierre, N. et al. Identifying causal variants by fine mapping across multiple studies. *PLoS Genet.* **17**, e1009733 (2021).
152. Rossen, J. et al. MultiSuSiE improves multi-ancestry fine-mapping in All of Us whole-genome sequencing data. Preprint at *medRxiv* <https://doi.org/10.1101/2024.05.13.24307291> (2024).
153. Chen, F. et al. Multi-ancestry transcriptome-wide association analyses yield insights into tobacco use biology and drug repurposing. *Nat. Genet.* **55**, 291–300 (2023).
154. Li, Z. et al. METRO: Multi-ancestry transcriptome-wide association studies for powerful gene-trait association detection. *Am. J. Hum. Genet.* **109**, 783–801 (2022).
155. Lu, Z. et al. Multi-ancestry fine-mapping improves precision to identify causal genes in transcriptome-wide association studies. *Am. J. Hum. Genet.* **109**, 1388–1404 (2022).
156. Ruan, Y. et al. Improving polygenic prediction in ancestrally diverse populations. *Nat. Genet.* **54**, 573–580 (2022).
157. Zhang, H. et al. A new method for multi-ancestry polygenic prediction improves performance across diverse populations. *Nat. Genet.* **55**, 1757–1768 (2023).
158. Weissbrod, O. et al. Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nat. Genet.* **54**, 450–458 (2022).
159. Johnson, R. & Pasaniuc, B. Implications of self-identified race, ethnicity, and genetic ancestry on genetic association studies in biobanks within health systems. Preprint at *arXiv* <https://doi.org/10.48550/ARXIV.2402.15696> (2024).
160. Yearby, R., Clark, B. & Figueroa, J. F. Structural racism in historical and modern US health care policy. *Health Aff.* **41**, 187–194 (2022).
161. Kittles, R. A. & Weiss, K. M. Race, ancestry, and genes: implications for defining disease risk. *Annu. Rev. Genomics Hum. Genet.* **4**, 33–67 (2003).
162. The Florida Senate. CS/CS/HB 999: Postsecondary Educational Institutions. <https://www.flsenate.gov/Session/Bill/2023/999>.
163. The White House. Ending Radical and Wasteful Government DEI Programs and Preferencing <https://www.whitehouse.gov/presidential-actions/2025/01/ending-radical-and-wasteful-government-dei-programs-and-preferencing/> (2025).
164. Whitehurst, L. Supreme Court lets trump administration cut \$783 million of research funding in anti-DEI push. *AP News* <https://apnews.com/article/supreme-court-trump-nih-dei-320a6b3749bf56703b50739362d1238c> (2025).
165. Schwabish, J. & Axelrod, J. NSF has canceled more than 1,500 grants. Nearly 90 percent were related to DEI. *Urban Institute* <https://www.urban.org/urban-wire/nsf-has-canceled-more-1500-grants-nearly-90-percent-were-related-dei> (2025).
166. Nunes, F. DEI initiatives removed from federal agencies that fund science, but scientific research continues. *The Conversation* <https://doi.org/10.64628/AI.w4fgjjcve> (2025).
167. Tillin, T. et al. Insulin resistance and truncal obesity as important determinants of the greater incidence of diabetes in Indian Asians and African Caribbeans compared with Europeans: the Southall and Brent REvisited (SABRE) cohort. *Diabetes Care* **36**, 383–393 (2013).
168. Hodgson, S. et al. Integrating polygenic risk scores in the prediction of type 2 diabetes risk and subtypes in British Pakistanis and Bangladeshis: a population-based cohort study. *PLoS Med.* **19**, e1003981 (2022).
169. Farmaki, A.-E. et al. Type 2 diabetes risks and determinants in second-generation migrants and mixed ethnicity people of South Asian and African Caribbean descent in the UK. *Diabetologia* **65**, 113–127 (2022).

## Acknowledgements

The authors thank A. Choudhury for helpful discussions and assistance in the preparation of this work, and M. Koromina for input on fine-mapping. The authors' work was funded by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant agreement No. 948561).

## Author contributions

The authors contributed equally to all aspects of the article.

## Competing interests

The authors declare no competing interests.

## Additional information

**Peer review information** *Nature Reviews Genetics* thanks Yen-Feng Lin, who co-reviewed with Tzu-Ting Chen; Chikashi Terao; and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature Limited 2026