Sankhya B:
https://doi.org/10.1007/s13571-023-00310-y

Population Diversity, Affinities and Genetic Epidemiology: A Commentary

Partha P. Majumder
National Institute of Biomedical Genomics, Kalyani, and Indian Statistical Institute, Kolkata, India

Abstract

Representation of populations of diverse ancestral histories is essential in genetic epidemiological research, failing which our understanding of the architectures of diseases will remain incomplete. Wide applications of inferences from studies with limited representation of population diversity can exacerbate health disparities. In this commentary, we identify the key reasons why inferences from studies without adequate representation of diverse populations can result in restricted applicability of inferences and the statistical challenges that need to be overcome for drawing more robust inferences. We also note that C.R. Rao, and his early mentor P.C. Mahalanobis, were pioneers to quantify population diversity and study population affinities.

AMS (2000) *subject classification.* Primary 00-01; Secondary 62-02.
Keywords and phrases Architecture of disease · Evolution · Diversity · Multi-ethnic mapping · Statistical challenges

1 Evolutionary Perspectives and Human Disease

Genetic epidemiology aims to unravel the architectures of diseases, explicitly taking genetic factors into account. For a disease, the goals are to (a) identify the factors that cause or associate with the disease, (b) estimate the impacts of the factors – singly and collectively, taking interactions among the factors into account – on the susceptibility to the disease, and (c) use the knowledge to estimate the disease-risk at the levels of the individual, the family and the population. The history of our species has influenced the architectures of our diseases. The influence of our evolutionary past has to be taken into account to understand diseases that afflict us. Fundamentally, our evolutionary history is recorded in our genomes. This evolutionary continuity justifies the use of model organisms, such as the mouse, in human disease research. However, it must also be emphasized that biological evolution is intimately linked to temporal changes in environment, that have resulted in many

Published online: 03 July 2023

differences in function of an organ or a pathway or a gene among organisms in spite of homology. This may be one reason why findings from animal models often do not translate to humans (Bart van der Worp et al., 2010) or successful clinical trials of drugs in animals often fail in humans (Mak et al., 2014). Further, many genomic changes have taken place that are human-specific; genes of the human immune system have an abundance of such changes (Quintana-Murci, 2019).

In the study of human disease, the set of environmental factors to be considered as possible contributors to the precipitation of a disease is nebulous. This uncertainty becomes a bottleneck to dissecting the architecture of a disease. In this commentary, we do not discuss the problems that can arise from differences in environmental exposures of individuals or populations. On the other hand, the genome is well-defined and the set of genomic factors that can possibly modulate susceptibility to a disease are those that are variable across individuals. Genomic variants show high inter-population differences in frequency. This variation is caused by natural selection; individuals in some populations may be exposed to an environment in which possession of the variant by an individual enhances their chance of survival. In such populations the frequency of the variant will rapidly increase compared to populations that do not experience such environmental exposure. There is also a stochastic way by which such inter-population variation arises. This is called random genetic drift. The sizes of most human populations are finite and over short periods of time the sizes remain stable. In a population comprising a finite number of diploid individuals, haploid gametes are randomly sampled from the infinite male pool of gametes (sperms) and another infinite female pool of gametes (eggs) to create a new generation of diploid individuals. This process of sampling introduces stochasticity in the relative frequencies of different variants from one generation to another. Such stochasticity results in variation in frequencies of gene variants over time between populations, even if the frequencies are equal in both populations to begin with. Similar sampling effects are also encountered when a new population group is formed from a preexisting group. If a small number of individuals carrying a restricted number of variants move away from one population to found another population, then the frequencies of only those variants carried by the founding members will alter over time in the newly-founded population, while the other variants in the ancestral population will be absent in the new population. These same phenomena also impact on associations between gene variants in stretches of the genome. In a newly-founded population, genomic positions (loci) that show high levels of associations among gene variants stretch over a large region of the genome. This association is called linkage disequilibrium (LD) and the stretch within which loci show high levels of LD is called a LD-block. The size of a LD block in a population diminishes over time because of physical exchange of genetic material between chromosomes

(genetic recombination) in each generation. Since contemporary human populations have evolved for variable lengths of time and have encountered variable pressures of selection, there are large differences both in frequencies of variants (The International Genome Sample Resource, https://www.internationalgenome.org/; archived dbSNP data are available from ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/genotype/) and in the sizes of LD blocks among contemporary populations (The International HapMap Consortium, 2007; archived data available from ftp://ftp.ncbi.nlm.nih.gov/hapmap/).

The impacts of evolutionary forces, population ancestries and admixtures among human populations impact inferences of genetic epidemiological research undertaken in multiple populations. Assessment of genomic diversities, genomic affinities and haplotype structures of populations are important, especially in respect of generalizability of inferences. These issues are discussed below in some detail.

2 C.R. Rao: A Pioneer in Understanding Population Affinities Using Physical and Biological Markers

C.R. Rao, together with P.C. Mahalanobis, laid the foundations of studies on population diversity and affinities in India. India has a vast array of populations with diverse ancestries and lifestyles. These populations have remained largely, but not completely, unadmixed for many thousands of years. They also follow some sets of social beliefs and practices, and have distinct types of social organization that have evolved over time. (Thapar, 1978). Contemporary populations of India comprise about 400 tribal populations with simple social organization, about 4000 caste populations with hierarchical social organization and about 150 populations who do not belong to these two main rungs of Indian society (Singh, 1993). While viewing the inter-population socio-cultural and physical diversities, one eminent Indian anthropologist (D.N. Majumdar) wondered whether population groups irrespective of the social rung to which they belong would be physically and biologically more similar if they lived in geographical proximity (perhaps because of environmental similarity and higher possibility of marital exchange across populations) than if the same population was divided into two subgroups separated by a large geographical distance. P.C. Mahalanobis, a founder of statistical science and the Indian Statistical Institute, and C.R. Rao, a student of Mahalanobis and later a colleague in the Indian Statistical Institute, got interested in these questions and undertook two landmark studies in northern India (United Provinces) and eastern India (Bengal) during the 1940s and 1950s. In each of these carefully-designed studies (Mahalanobis et al., 1949; Majumdar and Rao, 1958), a large number of individuals were sampled from about two dozen populations. A large set of body measurements were taken on each sampled individual.

These data were statistically analyzed to answer the anthropological questions. During the analyses, a number of novel statistical methods were devised by Mahalanobis and Rao. The studies did not produce unequivocal answers to the anthropological questions that they set out to answer. C.R. Rao recognized, towards the end of the Bengal anthropometric survey, that body dimensions were possibly highly influenced by environmental factors. He stated that attempts to understand biological affinities should rely on such attributes that are not influenced by environmental factors, such as blood groups. This was an incisive understanding. In the Bengal anthropometric survey, the ABO blood group type was determined on a limited number of individuals. C.R. Rao analyzed these data, but still the answers to the original anthropological questions remained equivocal. He stated that data on one blood group was insufficient; more biological markers were necessary. The Indian Statistical Institute in later years undertook many population surveys to estimate genetic diversity and affinities among population groups resident in different regions of India (Majumder and Mukherjee, 1993). The inferences from the data generated in these surveys have not completely consistent across geographical regions in India in respect of genetic affinities among population groups at different levels of cultural hierarchy (caste, tribe, etc.).

C.R. Rao revisited these problems of understanding genetic diversity and affinities during the late 1970s and early 1980s. He actually devised statistical methodologies for dissecting genetic diversity by a set of qualitative factors. He called the general theory as Analysis of Diversity (ANODIV). In particular, he considered the problem of apportioning of the total genetic diversity in a set of populations to within and between populations and devised a method that he termed as APDIV, apportionment of diversity. He published a series of papers and an expository paper that was published in Sankhya (Rao, 1982, and references therein).

Even though understanding the genomic and environmental underpinnings of human diseases was not the explicit reason for conducting the surveys undertaken by Mahalanobis, Majumder and Rao, these population surveys sought to answer questions that are precisely those necessary for the design of genetic epidemiological studies on diseases and also for generalizing the inferences of a genetic epidemiological study across many populations. Thus, conceptually these studies were forerunners of genetic epidemiological studies not just in India but also globally.

3 Representing Human Diversity in Genetic Epidemiological Research is Essential

The two major imperatives of genetic epidemiology are (i) to robustly excavate the architecture of a complex disease, and (ii) to use the results of an excavation to estimate risks of the disease in individuals belonging to various populations with a high degree of precision. A complex disease is influenced by variants

in multiple genes. The impacts of these gene variants are usually different in individuals drawn from the same population. The impact of the same variant may also differ among individuals drawn from different populations. The extent of difference in impact depends on the ancestral histories of the populations, differences in environmental conditions prevailing in the populations and differences in the nature and extent of exposures among individuals in these populations. Thus, both excavation of genomic and environmental underpinnings and estimation of applicable risks are difficult problems.

It is obvious that unless genetic epidemiological studies encompass a large set of populations representative of global diversity – both genomic and environmental diversity – the inferences drawn from such studies can hardly be translated widely for public health management. Many large genetic studies that have aimed to excavate determinants of diseases have been carried out in the past. How have we done in representing human global diversity in these studies? Terribly! In 2009, Need and Goldstein (2009) estimated that about 96% of participants in genome-wide association studies (GWAS) were of European descent, even though the global population only comprises about 16% of individuals of this group. The most adverse impact of non-representation of diversity in a genetic epidemiological study is that the estimated risks from the study may be thought to be applicable across all populations, when in fact these are not. The need to include larger ensembles of diverse populations has been emphasized (e.g., Bustamante, Burchard, de la Vega, 2011). As a result of some course corrections, the estimate of 96% reduced, seven years later, to about 80% (Popejoy and Fullerton, 2016). The current DNA sequence databases are somewhat less distorted. The gnomAD database (https://macarthurlab.org/2017/02/27/the-genome-aggregation-database-gnomad/) includes ~60% European sequences. But sampling bias is still evident. For example, less than 10% sequences are from individuals of African ancestry.

Does inclusion of diversity improve the ability to fully excavate the determinants of a disease? The answer is yes; for both Mendelian and complex disorders, as discussed later. A complex disease is often defined by the level of an associated biomarker. For example, an individual is diagnosed to be suffering from rheumatoid arthritis if the person's level of C-Reactive Protein (CRP) is above a certain threshold. However, some gene variants found in populations of African ancestry lower the CRP level (Kocarnik et al., 2018). An individual from such a population carrying one of these genetic variants will be declared to be normal when in fact the individual is actually suffering from rheumatoid arthritis. Medication required for ameliorating the disease condition will not be provided to this person. Also, importantly, inferences of the architecture of a complex disease inferred from case–control studies can be severely compromised if many 'cases' are systematically misdiagnosed as 'controls' in

some populations; misdiagnosis resulting from the phenomenon just described. Thus, unless populations with high levels of African ancestry are included in efforts to excavate determinants of rheumatoid arthritis, the efforts may remain incomplete or may even result in incorrect inferences.

Another similar example is in respect of beta-globin, a component of human hemoglobin. A variant (rs334) in the gene that encodes the beta-globin is known to determine whether an individual will have sickle-cell disease, a Mendelian disease, which is highly prevalent in some regions of the world, notably Africa. This variant is also known to lower HbA1c, a biomarker used to determine whether an individual is suffering from type-2 diabetes (Lacy et al., 2017). The variant also occurs at a fairly high frequency among Hispanics/Latinos (Moon et al., 2019). If this fact is not taken into account, many Hispanics/Latinos will be declared as non-diabetic, when in fact they are diabetic. A study to excavate the determinants and estimate risk of type-2 diabetes will seriously suffer if Hispanics/Latinos are not included. Further, individuals carrying the rs334 variant will appear to have their blood glucose under control and hence will not be provided medication, even though in reality they may require medication.

As discussed earlier, various evolutionary forces, notably natural selection and random genetic drift, have resulted in large inter-population variations in frequencies of alleles at various loci. In GWA studies, association between a disease and a predisposing allele cannot be uncovered unless the study is conducted in a population in which the allele has a reasonably high frequency. Since this cannot be known a priori, inclusion of a diversity of populations is essential. A compelling example is that a nonsense variant that causes premature termination of the *PCSK9* gene was found to be associated with a dramatic reduction (28–40%) in the concentration of low-density lipoprotein (LDL) cholesterol (Cohen et al., 2006). Elevated level of LDL cholesterol enhances the risk to coronary heart disease. The nonsense variant in *PCSK9* is rare in populations of European ancestry, but is quite common in populations of African ancestry, perhaps due to genetic drift (Horton et al., 2007). Association between the *PCSK9* variant and LDL cholesterol level was uncovered by studying individuals with African ancestry; it would have been hardly possible to detect this association in other populations. Even though the discovery was made among individuals of a specific population, the benefits of the discovery has been applicable to all. The finding that this naturally occurring variant adversely impacts on the level of PCSK9 protein, which in turn reduces LDL-cholesterol level, prompted the initiation of a search for ways to inhibit the synthesis of PCSK9 protein. The search has led to the identification of PCSK9

inhibitor drugs (Gouni-Berthold et al., 2016; Roth et al., 2016) that are now widely used to lower LDL cholesterol level to reduce the risk of heart disease.

Although the emphasis on analysis of DNA from diverse populations for disease gene mapping is correct, there is an interest twist to this. African populations being older than populations of other geographical regions show low levels of linkage disequilibrium (LD). Commercial DNA microarrays that are popularly used in GWA studies interrogate few SNPs from regions of strong LD. Resultantly, because of low LD in African populations, use of these commercial microarrays reduces the statistical power to detect genome-wide associations with common diseases in Africa. However, the low levels of LD in Africa it has been emphasized (Teo et al., 2010) will make it easier to localize the causal variants responsible for GWA signals of association, which is one of the major roadblocks for GWA studies of European populations.

4 Challenges

4.1 Identifying Focal Points

The inclusion of a diverse set of populations in a genetic epidemiological study introduces many challenges in study design, data collection and statistical analysis. We shall indicate some of these challenges in the context of a genome-wide association study with a complex disease. For ease of exposition, let us consider the key steps in a GWAS that are relevant to this discussion. First, based on sample size calculations, an adequate number of individuals – patients (cases) and normal individuals (controls) – are sampled, usually in equal numbers, from a population under study to provide the desired level of statistical power. Second, on each individual recruited into the study, data on a large number of clinical and exposure variables are collected, and a blood sample taken. Third, DNA of each individual is analyzed using a DNA microarray that generates genotypes of the individual at a large number of loci (of the order of several hundred thousand) at fixed points on the genome. Fourth, genotype imputation – that seeks to generate genotype information on the sampled individuals at a set of loci not directly genotyped using the microarray – is carried out to increase statistical power of the study. Fifth, the statistical significance of the difference in genotype proportions – both directly observed and imputed – between the set of cases and controls is assessed by a contingency chi-squared test at individual loci and/or for haplotypes (explained in some detail later) are done, after adjusting for differences in environmental exposures and other concomitant variables. Finally, inferences on genomic association are drawn after correcting the test results for multiple testing.

4.2 Designing Content of DNA Microarrays

The nature of genome-wide genotype data on a DNA sample hinges critically on the content of the DNA microarray used. Normally, these microarrays are designed on the basis of data that have been collected in various population genetic studies. Initially, because of the high cost involved in the conduct of a GWA study, most studies were carried out on relatively homogeneous populations, such as those of Finland, Iceland, etc. In general, the focus was on European populations. The HapMap project (The International HapMap Consortium, 2007) generated information on haplotype structures of populations. Because loci (single nucleotide polymorphic loci; SNP loci or simply SNPs) within a haplotype block are associated, a few SNPs, called tagSNPs, from a haplotype block usually suffice to capture ("tag") the information on genomic variation within the block. These tagSNPs are usually placed on a DNA microarray by design, so that genotype data are generated on these tagSNPs. As the HapMap data have shown, the haplotype structures of ancestrally close populations are similar, tagSNPs from one population can be ported to an ancestrally close population. In other words, in populations recently derived from a common ancestral population, the tagSNPs of one population will remain informative in another population, in the sense of retaining their polymorphic status and also capturing information on haplotype blocks. Genotype information captured by tagSNPs can then be used to impute genotypes at the other loci within the haplotype block with reasonable accuracy in these populations (discussed later). However, problems arise when a DNA microarray enriched with tagSNPs of a population is used on an ancestrally distant population. First, many loci on the microarray often turn out to be monomorphic, thereby resulting in wastage of resources. Second, because of differences in haplotype block structure, many of the SNPs may not be as informative for imputing genotype information of unobserved SNPs, i.e., SNPs not placed on the microarray. These issues prompted genome scientists to consider working with the industry to develop microarrays that have greater applicability in multiethnic GWA studies. A collaborative effort has resulted in the development of the Multi-Ethnic Global Array (MEGA). This collaboration was among the genomics company Illumina, the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA), and PAGE II (Population Architecture using Genomics and Epidemiology) consortium (http://www.pagestudy.org). MEGA comprises a "backbone" set of loci derived from results of a few large multiethnic studies and a custom-content set; for details see Bien et al. (2016). The backbone set comprises SNPs included in previously designed arrays by Illumina (e.g., Infinium Human Core Bead Chip; African Diaspora Consortium Power Chip) and SNPs derived from DNA sequencing data from > 36,000 individuals in diverse ethnic groups, especially loss of function and splice variants. The

custom-content set comprises SNPs taken from published and unpublished studies, regulatory SNPs, etc. Subsequently, an improved microarray, the Expanded Multi-Ethnic Genotyping Array (MEGA^{EX}), was developed to provide extensive genotyping coverage of European, East Asian, and South Asian populations. In sum, technologies for data generation are improving, and are expected to continue to improve, as data are collected using these technologies in larger numbers of multi-ethnic studies in diverse regions of the world. However, rapid statistical analysis of the data generated by GWA studies using the newer microarrays in different populations and feedback to the industry are required in order that the SNPs placed on the microarrays are polymorphic and imputation-informative in a larger set of ancestrally-diverse populations. By simultaneously using data from whole-genome or whole-exome sequencing studies, these statistical analyses can be done more profitably. One such example is a methodological improvement for selecting SNPs for microarray design applicable to multi-ethnic GWA studies developed by Wojcik et al. (2018).

Increasingly, there is a shift from using DNA microarrays to DNA sequencing in order to identify disease-associated variants. DNA sequencing will alleviate many of the problems mentioned in this section. However, sequencing still remains expensive in most global regions and currently access to sequencing platforms is constraining.

4.3 Challenges in Imputation

Genotype imputation relies on the observation that because of shared ancestry of individuals within a population, or even between populations, there are stretches of DNA that are shared among individuals. This sharing results in the formation of haplotype blocks, stretches of loci at which individuals share genotypes. The association of genotypes at loci within a haplotype block allows for inferring of genotypes at an unobserved locus based on information of genotypes at directly observed loci. A good description of the principles and methods of genotype imputation is available in Li et al. (2009). Depending on the ancestral histories of populations, the lengths of haplotype blocks can be quite variable across populations, as we have alluded to earlier. Haplotyping can be done using heuristic methods (Clark, 1990) or by the maximum-likelihood method using the EM algorithm (Excoffier and Slatkin, 1995). Imputation methods (Shi et al., 2019), with some methodological differences, have been encoded in packages of which the popular ones are Beagle4.1(https://faculty.washington.edu/browning/beagle/beagle.html), IMPUTE2 (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html), MACH (http://csg.sph.umich.edu/abecasis/mach/), Minimac3 (http://genome.sph.umich.edu/wiki/), and SHAPEIT2 (https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/). The overall performance of these packages are similar; IMPUTE2 is

more time-consuming than the others but its imputation-accuracy is also slightly higher. Successful and accurate genotype imputation, critically hinges on the choice of a reference population with well-characterized haplotypes. The haplotype data from the reference population are used to impute genotypes of individuals of the GWAS study. Well-characterized and dense haplotype data are available only for a limited number of global populations; 1000 Genomes phase 3 data (Sudmant et al., 2015) and Haplotype Reference Consortium (HRC) (McCarthy et al., 2016) data, are currently widely used. If the cases and controls of the GWAS study are drawn from a study population which is different from one of the reference populations, then the reference population that is used for imputation is the one that is genetically the closest to the study population. However, in a multi-ethnic study, if different reference populations are used to impute genotypes, then different sets of genotypes can be imputed possibly with differing levels of accuracy. This can pose a great problem in combined association analysis. Innovative statistical methods have been conceptualized and implemented, including the use of mixtures of reference panels (Huang et al., 2009; Schurz et al., 2019); however, further statistical improvements are necessary.

4.4 Challenges in Meta-analysis

The case–control association methodology in GWA studies is fairly standard. Having obtained reliable association results, various laboratory experiments are performed to identify a set of causal SNPs. However, the allelic effects of even these causal SNPs may vary across populations, primarily, but not exclusively, for two reasons: (a) interactions with environment and differing levels of environmental exposures among populations, and (b) variation in allele frequencies among populations. Meta-analysis is often performed to estimate the 'true' allelic effects, by pooling results of multiple populations. Two basic models are popular in meta-analysis; fixed effects and random effects. The fixed effects model, that essentially assumes that all the studies have been performed in the same population because allelic effects are considered to be equal in all populations, is clearly untenable in a multi-ethnic context. The random effects model is also of limited applicability because it is expected that study populations derived from the same ancestral population will show less heterogeneity than those that are derived from distinct ancestral populations. In a multi-ethnic study, it is more likely that subsets of study populations will be derived from different ancestral populations. To overcome the limitations arising from assumptions that underlie the random effects model, Doi et al. (2015) had proposed an inverse variance quasi-likelihood-based alternative to the random effects model. To our knowledge, this model has not been applied to meta-analysis of multi-ethnic GWA studies. The expected nature

of heterogeneity among populations vis-à-vis their ancestral derivation, as mentioned earlier, has been addressed by Morris (2011). He has developed a meta-analysis methodology that takes into account the observation that populations derived from the same ancestral population, and hence closely clustered, are expected to show similar allelic effects, while those that are derived from different ancestral populations will belong to distinct clusters. His method relies on a Bayesian partition model (Knorr-Held and Rasser, 2000; Denison and Holmes, 2001) and takes into account differences in local LD structures among populations. The allelic effects of a variant are assumed to be the same in populations belonging to a cluster, but the effects may be variable across clusters of populations. In some sense, this model is a hybrid of fixed- and random-effects models. A software package MANTRA incorporating this method has been developed. Although there have been some applications of this method (Li and Keating, 2014), the number of applications has been limited. With more empirical experience, the usefulness of this method will become more apparent. Formulation of alternative statistical methods is also required.

4.5 Risk Estimation

Fisher (1918) had introduced the concept of estimating the probability that an individual will possess a polygenic trait. The concept has been resurrected some years ago (e.g., Wray et al.. 2007) and a method has been suggested and refined for calculating the risk of a complex disease for an individual. A score, called the Polygenic Risk Score (PRS), is calculated (Choi et al., 2020). Briefly, a the PRS is constructed using information derived from a GWAS. A GWAS provides estimates of effect sizes of risk alleles at significantly associated loci. Using the effect sizes or a function of these effect sizes as weights, a person's PRS is calculated as the weighted sum of the risk allele counts (0, 1 or 2) at the significant loci detected in GWAS. Various regression-based methods have been described to estimate effect sizes. Further, since loci within a haplotype block are associated, SNPs are "clumped" (or, "pruned") so that the retained SNPs are not significantly associated. Then, "thresholding" is done to remove variants with a *p*-value larger than a chosen level of significance ($p > p_T$). The PRS has been widely applied (see Torkamani et al., 2018; Lewis and Vassos, 2020). However, it has been emphasized that unrestricted use of the PRS may exacerbate health disparities (Martin et al., 2019). This is because the effect sizes of risk variants are estimated from GWA studies that have primarily been conducted on populations of European ancestry. Effect sizes of these variants be very different in non-European populations, that can lead to hugely different risk estimates. Since PRSs are increasing being used for clinical

decision-making (Martin et al., 2019), their use without exercising adequate caution may actually be harmful. The key problems that plague the use of PRSs in multi-ethnic settings are exactly the same as those discussed in earlier sections. Similar solutions have also been suggested. For example, Grinde et al. (2019) have suggested that allele-weighting may use results of trans-ancestry meta-analysis to improve prediction accuracy. Marquez-Luna et al. (2017) has suggested a method (MultiPRS) that combines estimates of effect sizes based on large European training data with estimates based on a small data set from the non-European target population. The authors have shown that better predictive value is obtained in the target population by this combined approach. In spite of these methodological advances, problems persist; a detailed discussion is available in Kaplan and Fullerton (2022).

5 Conclusion

Undoubtedly, the study of diverse populations is of great value in deciphering the architectures of diseases. The understanding of the etiologies of complex diseases will remain incomplete unless more inclusive studies on populations of diverse ancestries living with a diversity of environmental exposures are undertaken. However, because of differences in evolutionary histories of extant populations, the profiles of the genomes of individuals drawn from diverse populations impact on the efficiency of current methodologies for understanding diseases. The only way to resolve these problems and to improve the robustness of conclusions and predictions from multi-ethnic studies is to design the studies carefully and use efficient and innovative statistical methods of data analysis.

Acknowledgements Thanks are due to Dr. Samsiddhi Bhattacharjee for his careful reading of the manuscript and for offering valuable comments.

References

BART van der WORP, H., HOWELLS, D.W., SENA, E.S., PORRITT, M.J., REWELL, S., O'COLLINS, V. and MACLEOD, M.R. (2010) Can animal models of disease reliably inform human studies? *PLoS Medicine* **7**, 1-8.

BIEN, S.A,, WOJCIK, G.L., ZUBAIR, N., GIGNOUX, C.R., MARTIN, A.R., KOCARNIK, J.M., et al. (2016) Strategies for Enriching Variant Coverage in Candidate Disease Loci on a Multiethnic Genotyping Array. *PLoS ONE* **11**(12), e0167758.

BUSTAMANTE, C.D., BURCHARD, E.G. and DE LA VEGA, F.M. (2011) Genomics for the world. *Nature* **475**,163-165.

CHOI, S.W., MAK, T.S. and O'REILLY, P.F. (2020) Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols* **15**, 2759-2772.

CLARK, A.G. (1990) Inference of haplotypes from PCR-amplified samples of diploid Popul.tions. *Molecular Biology and Evolution* **7**,111–22.

COHEN, J.C., BOERWINKLE, E., MOSLEY, T.H. and HOBBS, H.H. (2006) Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *New England Journal of Medicine* **354**,1264–1272.

DENISON, D.G.T. and HOLMES, C.C. (2001) Bayesian partitioning for estimating disease risk. *Biometrics* **57**,143–149.

DOI, S.A., BARENDREGT, J.J., KHAN, S., THALIB, L. and WILLIAMS, G.M. (2015) Advances in the Meta- analysis of heterogeneous clinical trials I: The inverse variance heterogeneity model. *Contemporary Clinical Trials* **45**,130-138.

EXCOFFIER, L. and SLATKIN, M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*,**12**, 921–927.

FISHER R.A. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **53**, 399–433.

GOUNI-BERTHOLD, I., DESCAMPS, O.S., FRAASS, U., HARTFIELD, E., ALLCOTT, K., DENT, R. and MÄRZ, W. (2016) Systematic review of published Phase 3 data on anti-PCSK9 monoclonal antibodies in patients with hypercholesterolaemia. *British Journal of Clinical Pharmacology* **82**, 1412–1443.

GRINDE, K.E., QI, Q., THRONTON, T.A., LIU, S., SHADYAB, A.H., CHAN, K.H.K., REINER, A.P. and SOFER, T. (2019) Generalizing polygenic risk scores from Europeans to Hispanics/ Latinos. *Genetic Epidemiology* **43**, 50-62.

HAIKO, SCHURZ, H., MÜLLER, S.J., VAN, HELDEN, P.D., TROMP, G., HOAL, E.G., KINNEAR, C.J. and MÖLLER, M. (2019) Evaluating the Accuracy of Imputation Methods in a Five-Way Admixed Population. *Frontiers in Genetics*, 05 February 2019 | https://doi.org/10.3389/fgene.2019.00034

HORTON, J.D., COHEN, J.C. and HOBBS, H.H. (2007) Molecular biology of PCSK9: its role in LDL metabolism. Trends Biochemical Sciences **32**,71–77.

HUANG, L., L.I., Y., SINGLETON, A.B., HARDY, J.A., ABECASIS, G., ROSENBERG, N.A. and SCHEET, P. (2009) Genotype-Imputation Accuracy across Worldwide Human Populations. *American Journal of Human Genetics* **84**: 235-250.

KAPLAN, J.M., FULLERTON, S.M. (2022) Polygenic risk, population structure and ongoing difficulties with race in human genetics. *Philosophical Transactions of The Royal Society London Series B Biological Sciences*. **377**(1852): 20200427. https://doi.org/10.1098/rstb.2020.0427.

KNORR-HELD, L. and RASSER, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics* **56**:13–21.

KOCARNIK, J.M., RICHARD, M., GRAFF, M., HAESSLER, J., BIEN, S., CARLSON, C., CARTY, C.L., REINER, A.P., AVERY, C.L., BALLANTYNE, C.M., LACROIX, A.Z., ASSIMES, T.L., BARBALIC, M., PANKRATZ, N., TllANG, W., TAO, R., CHEN., D., TALAVERA, G.A., DAVIGLUS, M.L., CHIRINOS-MEDINA, D.A., PEREIRA, R., NISHIMURA, K., BUŽKOVÁ, P., BEST, L.G., AMBITE, J.L., CHENG, I., CRAWFORD, D.C., HINDORFF, L.A., FORNAGE, M., HEISS, G., NORTH, K.E., HAIMAN, C.A., PETERS, U., L.E., MARCHAND, L. and KOOPERBERG, C. (2018) Discovery, fine-mapping, and conditional analyses of genetic variants associated with C-reactive protein in multiethnic populations using the Metabochip in the Population Architecture using Genomics and Epidemiology (PAGE) study. *Human Molecular Genetics* **27**, 2940-2953.

LACY, M.E., WELLENIUS, G.A., SUMNER, A.E., CORREA, A., CARNETHON, M.R., LIEM, R.I., WILSON, J.G., SACKS, D.B., JACOBS, D.R., J.R., CARSON, A.P., LUO, X., GJELSVIK, A., REINER, A.P., NAIK, R.P., LIU, S., MUSANI, S.K., EATON, C B. and WU, W.C. (2017) Association of Sickle Cell Trait With Hemoglobin A1c in African Americans. *Journal of the American Medical Association* **317**, 507-515.

LEWIS, C.M. and VASSOS, E. (2020) Polygenic risk scores: from research tools to clinical instruments. Genome Medicine **12**, 44.

L.I., Y., WILLER, C., SANNA, S. and ABECASIS, G. (2009) Genotype Imputation. *Annual Review of Genomics and Human Genetics* **10**, 387-406

L.I., Y.R. and KEATING, B.J. (2014) Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. Genome Medicine **6**, 91.

MAHALANOBIS, P.C., MAJUMDAR, D.N. and RAO, C.R. (1949). Anthropometric survey of the United Provinces, 1941: A statistical study. *Sankhya* **9**, 90–324.

MAJUMDAR, D.N. and RAO, C.R. (1958). Bengal anthropometric survey, 1945: A statistical study. *Sankhya* **19**, 203–411.

MAJUMDER, P.P. and MUKHERJEE, B.N. (1993) Genetic diversity, affinities among human populations: An overview. In *Human Population Genetics: A Centennial Tribute to J.B.S. Haldane* (P.P. Majumder, ed.), New York: Plenum Press, pp. 255-275.

MÁRQUEZ-LUNA, C. and LOH, P.R., SOUTH ASIAN TYPE 2 DIABETES (SAT2D) CONSORTIUM; SIGMA TYPE 2 DIABETES CONSORTIUM and PRICE, A. L. (2017). Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genetic Epidemiology* **41**, 811–823.

MARTIN, A.R., KANAI, M. and KAMATANI, Y. et al. (2019) Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics* **51**, 584–591

MAK, I. W. Y., EVANIEW, N. and GHERT, M. (2014) Lost in translation: animal models and clinical trials in cancer treatment. *American Journal of Translational Research* **6**, 114–118.

MCCARTHY, S., DAS, S., KRETZSCHMAR, W., DELANEAU, O., WOOD, A. R. and TEUMER, A., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics* **48**, 1279–1283.

MOON, J.Y., LOUIE, T.L., JAIN, D., SOFER, T., SCHURMANN, C., BELOW, J.E., LAI, C.Q., AVILES-SANTA, M.L., TALAVERA, G.A., SMITH, C.E., PETTY, L.E., BOTTINGER, E.P., CHEN, Y.I., TAYLOR, DAVIGLUS, M.L., CAI, J., WANG, T., TUCKER, K.L., ORDOVÁS, J.M., HANIS, C.L., LOOS, R.J.F, SCHNEIDERMAN, N., ROTTER, J.I., KAPLAN, R.C. and Q.I., Q. (2019) A Genome-Wide Association Study Identifies Blood Disorder-Related Variants Influencing Hemoglobin A1c With Implications for Glycemic Status in U.S. Hispanics/Latinos. *Diabetes Care* **42**,1784-1791.

MORRIS, A.P. (2011) Transethnic meta-analysis of genomewide association studies. *Genetic Epidemiology* **35**, 809-822.

NEED, A.C. and GOLDSTEIN, D.B. (2009) Next generation disparities in human genomics: concerns and remedies. *Trends in Genetics*, **25**, 489-494.

POPEJOY, A. and FULLERTON, S. (2016) Genomics is failing on diversity. *Nature* **538**, 161–164. https://doi.org/10.1038/538161a

QUINTANA- MURCI, L. (2019) Human immunology through the lens of evolutionary genetics. *Cell* **177**, 184–199.

RAO, C.R. (1982) Diversity: Its Measurement, Decomposition, Apportionment and Analysis. *Sankhyā*: The Indian Journal of Statistics, Series A (1961-2002) **44**, 1-22.

ROTH, E.M., MORIARTY, P.M., BERGERON, J., LANGSLET, G., MANVELIAN, G., ZHAO, J., BACCARA-DINET, M.T., RADER,D.J. and ODYSSEY CHOICE I investigators (2016) A phase III randomized trial evaluating alirocumab 300 mg every 4 weeks as monotherapy or add-on to statin: ODYSSEY CHOICE I. *Atherosclerosis* **254**, 254-262.

SHI, S., YUAN, N., YANG, M., DU, Z., WANG, J., SHENG, X., WU, J. and XIAO, J. (2019) Comprehensive Assessment of Genotype Imputation Performance. *Human Heredity* **83**, 107–116.

SINGH, K.S. (1993) *People of India, Volume I*. Oxford University Press, New Delhi.

SUDMANT, P. H., RAUSCH, T., GARDNER, E. J., HANDSAKER, R. E., ABYZOV, A. HUDDLESTON, J. et al (2015). An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81.

TEO, YY., SMAKK, K. & KWAITKOWSKI, D. (2010) Methodological challenges of genome-wide association analysis in Africa. *Nature Reviews Genetics* **11**, 149–160.

THAPAR, R., (1978) *Ancient Indian Social History: Some Interpretations.* Orient BlackSwan, Hyderabad, India

The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861.

TORKAMANI, A., WINEINGER, N.E. and TOPOL, E.J. (2018) The personal and clinical utility of polygenic risk scores. *Naturer Reviews Genetics* **19,** 581–590.

WOJCIK, G.L., FUCHSBERGER, C., TALIUN, D., WELCH, R., MARTIN, A.R., SHRINGARPURE, S., CARLSON, C.S., ABECASIS, G., KANG, H.M., BOEHNKE, M., BUSTAMANTE, C.D., GIGNOUX, C.R., and KENNY, E.E. (2018) Imputation-aware tag SNP selection to improve power for large-scale, multi-ethnic association studies. *G3 (Bethesda)* **8**, 3255-3267.

WRAY, N.R., GODDARD, M.E. and VISSCHER, P.M. (2007) Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research* **17**, 1520-1528.

Partha P. Majumder
National Institute of Biomedical Genomics,
Kalyani, Indian Statistical Institute,
Kolkata, India
E-mail: ppm@isical.ac.in

Paper received: 16 September 2021; accepted: 19 May 2023