A Roadmap to Increase Diversity in Genomic Studies

Segun Fatumo[1,2], Tinashe Chikowore[3,4], Ananyo Choudhury[3], Muhammad Ayub[5], Alicia R. Martin[6,7], Karoline Kuchenbäcker[5,8]

[1]*The African Computational Genomics* (TACG) Research Group, MRC/UVRI and LSHTM, Entebbe, Uganda.

[2]The Department of Non-communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, UK

[3]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.

[4]MRC/Wits Developmental Pathways for Health Research Unit, Department of Paediatrics, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[5]Division of Psychiatry, University College of London, London, UK.

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

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

[8]UCL Genetics Institute, University College London, London, UK

Correspondence:

Dr. Segun Fatumo
The African Computational Genomics (TACG) Research Group,
MRC/UVRI and LSHTM, Entebbe, Uganda.
segun.fatumo@lshtm.ac.uk

Competing Interests statement: The authors declare no competing interests

Abstract

Two decades ago, the sequence of the first human genome was published. Since then, advances in genome technologies have resulted in whole genome sequencing and microarray-based genotyping of millions of human genomes. However, genetic and genomic studies are predominantly based on populations of European ancestry. As a result, the potential benefits of genomic research – including better understanding of disease aetiology, early detection and diagnosis, rational drug design and improved clinical care – may elude those underrepresented populations. Here, we describe factors that have contributed to the imbalance in representation of different populations. Leveraging our experiences in setting up genomic studies in diverse global populations, we propose a roadmap to enhancing inclusion and ensuring equal health benefits of genomics advances. Our Perspective highlights the importance of sincere, concerted global efforts towards genomic equity to ensure the benefits of genomic medicine are accessible to all.

1.0 MAIN

As of June 2021, the vast majority (86.3%) of genomics studies have been conducted in individuals of European descent (Figure 1), which represents an increase from 81% in 2016. At the same time, the proportion of studies conducted in underrepresented populations have either stagnated or decreased; genetic studies including participants with multiple ancestries have increased but only very slightly, to 4.8% (figure 1)[1] . This shows that progress towards diversification has been painfully slow. The genomic research community tends to extensively use resources with relatively straightforward access models, such the UK Biobank which includes participants of mostly European descent, while other ancestry groups tend to have very few such resources and limited access models. Data from the International HundredK+ Cohorts Consortium (IHCC), a recently established consortium of international cohort studies, also shows considerable ancestral disparities (Figure 2).

Most of the data from non-European populations captured in the genome-wide association studies (GWAS) catalog and current genomic studies come from individuals in diaspora populations. For example, the 1.1% of participants of African

ancestry in the GWAS Catalog are mainly African Americans; the proportion of continental Africans in genomic studies is insignificant with respect to the prevailing genomic research. While there are five major African ethnolinguistic divisions, the African diaspora in the UK and USA predominantly consists of just one of these divisions, the Niger-Congo speakers[2]. Africans harbour a far greater amount of genetic and linguistic diversity (e.g., over 3000 indigenous languages) than populations from other continents[3,4] and this diversity is largely partitioned by geography. However, more than 90% of these ethnolinguistic groups have no representative genetic data to date. Studying a small number African diaspora populations (African American and Black participants in the UK and Europe) and grouping all participants into a broad category of African ancestry will continue to promote imbalance, widen health disparities, and fail to capture the genetic diversity in Africa. Moreover, large-scale differences in environment and lifestyle could further limit the transferability of genetic insights (such as Polygenic Risk Score models) gained from diaspora populations to continental African populations[5]. This calls for immediate measures to address the genomic studies imbalance.

Here, we discuss the factors have contributed to the current inequalities in genomic studies. We highlight successful genomic studies in Africa, Asia, Australia and Latin America and reflect on the challenges and opportunities involved in setting up studies such as these. Based on our experience, we chart a roadmap to increase diversity of populations in genomic studies which requires a concerted global effort. We emphasize that any successful roadmap must leverage established research infrastructure, capacity, expertise, and leadership within local institutions in those countries.

2.0 Lack of diversity in genomics leads to unmet scientific needs and health disparities

Eurocentric biases in genetics studies are not only inequitable, but also result in major missed scientific opportunities. Underrepresentation is driven by inequitable resource allocation, which is an ethical issue, as are potential healthcare disparities stemming from imbalanced research. Here, we focus on the major missed scientific opportunities that arise as a consequence of underrepresentation, opportunities such as identification of novel associations with population-enriched variants, pinpointing

causal variants for functional follow-up, improving genetic risk prediction accuracy for all populations (particularly underrepresented populations), and understanding shared versus unique genetic and environmental population risk factors that influence health outcomes[15,41,42,43].

Certain characteristics of underrepresented populations would undoubtedly benefit international efforts towards discovery of disease-causing variants. For example, African populations have the most genetic diversity, followed by South Asians. This helps fine-map GWAS signals and identify target genes, an essential step in gaining mechanistic insights. These populations also have the most loss-of-function variants, which can aid interpretation of genomic function and understanding mutational constraints[10]. Endogamy within subgroups and consanguinity in some South Asian populations can enhance the power for discovery of recessive inheritance.

There are already clear examples of population-enriched clinically important variants only discovered in underrepresented populations; for example the association between *APOL1* and chronic kidney disease[11], variants in *G6PD* that contribute to missed diabetes diagnosis[12], and loss-of-function variants in *PCSK9* that lower LDL cholesterol (the discovery that led to *PCSK9* inhibitor drugs)[13] , all of which were identified in populations with African ancestry.

Additionally, polygenic risk scores have become increasingly predictive as GWAS have grown and increased in power. Interest in their predictive utility, which is now comparable to other biomarkers commonly used in screening for actionable diseases such as breast cancer and cardiovascular disease[14,15], has raised their potential for clinical implementation alongside other risk factors[16]. However, their accuracy decays with increasing genetic distance from the study cohort[17,18] ; a previous study showed that Eurocentric GWAS results for several traits produce PRS that are 2-fold and 4.5-fold more accurate in individuals with European than East Asian and African ancestry, respectively[6]. Thus, increasing diversity in genomics is critical to ensure that translation of genomic screening strategies improves health outcomes for all and does not exacerbate health disparities[6].

Imbalanced ancestral diversity also pervades data sets with whole genome and whole exome sequencing. This is of particular concern for resources that are available as reference panels for genotype imputation. For example, the most widely used genomic reference panel consisting of the 1000 Genomes Project dataset, has been shown to represent a minority of ancestry groups found in mainland South Asia and Africa[19]. This limits the post-imputation coverage of genomic variation for many populations.

3.0 Factors contributing to the current inequalities in genomic studies

The dominance of European and American scientists in genomic research stems from advances in genomic technologies, infrastructure, and the better funding opportunities. These are a consequence of structural advantages, some of which are related to historical and present-day exploitation. The lack of diversity among researchers is a crucial driver of bias in genetic studies[20]. Previous work shows that investigators have personal connections to their countries of origin, leading to their prioritization in research[21].

Concerns about population stratification as well as lack of capacity and analytical expertise with respect to multi-ancestry cohorts have been cited as justification for exclusion of individuals of non-European descent from genomics studies. Now, however, with advances in genetic technologies that capture the variation in diverse populations coupled with requisite analytical tools, there is ample opportunity to explore genomic studies in multi-ancestry populations.

Large-scale genetic studies are expensive and time-intensive, requiring continuity of expertise. Several countries have faced political instability that has made investments in genomic research erratic, but recent strategic funding by the US National Institutes of Health (NIH) and Wellcome Trust through the Human Heredity and Health in Africa initiative has led to the birth of genome-wide association studies on the African continent[37].

For participants to engage in research they need to trust the researchers; however, past history of research abuse and exploitation has negatively impacted on the ability of researchers to work with diverse communities[21]. The limited understanding of genetic concepts among some indigenous populations and the paucity of data on

effective models for community engagement may also contribute to poor enrollment of research participants in some population groups[22]. When community advisory boards (CAB) are sustained by community members who meet with researchers, they may facilitate positive community engagement. For example, the CAB would have the responsiblity of understanding how the researchers aim to avoid potential stigmatisation, genetic discrimination, racial stereotyping and other potential group harms in genetic research which are beyond the scope of this current review.

There are two broad groups of under-represented populations; residents of low and middle income countries (LMICs)[23] and indigenous and minority groups across the globe [24]. The factors that have caused unequal representation are overlapping in both groups. The burden of historical injustices including coercion and deception in research [25,26] and negative experience with the healthcare system[27] results in lack of trust in research. Mutual suspicion and lack of trust is a significant cause for scientists to avoid enrolling indigenous groups and for indigenous groups to avoid participating in research.

For LMICs, lack of resources such as funds, institutional capacity and a skilled work force are major barriers [28]. These countries have limited funds to invest and genomic research does not often make its way onto their list of priorities. For genomic research, scientists in these countries therefore depend on funding from high income countries, mostly through collaborative efforts. The policies and priorities of these funding agencies influence decisions about the focus of research and they set the research agenda in many LMICs [29,30]. In so-called ‘collaborative’ research, scientists from LMICs are in fact under-represented as first and last authors – and this impacts their motivation to engage in big collaborations [31].

Lack of expertise in ethical, legal, and social implications (ELSI) relevant to genomics research has hindered the conduct of research and data sharing [32,33]. Creating expertise in this area and making ELSI considerations an integral part of the study design will address this gap; local adaptation of the available guidance can help [34].

4.0 Setting up genomic studies in underrepresented populations: what has worked?

Despite the unequal representation of ancestry groups in genomic research, some studies in underrepresented populations have been very successful. In this section, we discuss flourishing genomic studies in underrepresented populations, mostly from LMICs in Africa, Asia, Latin America. As problem of genomic underrepresentation is not restricted to LMICs, we also highlight a case study from Australia. For each of these exemplar studies, we reflect on factors contributing to their successes.

4.1 AFRICA

Large-scale genomics research in Africa has so far been driven mainly by international funding, with very few examples of government funded national-level initiatives such as the Southern African human genome programme [35]. MalariaGen [36] was among the first studies to be based on a cohort that spanned multiple African countries. The focus of this study on the genetics of both the parasite and the host enabled it to capture snapshots of human genetic diversity, especially in some of the malaria endemic geographic regions of Africa. However, the H3Africa consortium was the first major pan-African study to have a comprehensive spread across the continent and across a wide variety of diseases and traits [37]. As well as investigating of communicable and non-communicable diseases, the consortium has contributed to developments in several major aspects of genetics research such as ethics and community engagement, data sharing and governance, and disease awareness, as well as technical developments including dissemination of bioinformatics skills, and design of a genotyping array and analysis tools [38]. Next, we focus on two cohorts, the Uganda genome resource (UGR) study and the AWI-Gen study (a collaborative centre of the H3Africa consortium) that are cross-sectional in terms of their populations and have been generating key insights into disease genetics.

4.1.1 Strategic collaboration and capacity building: The Uganda Genome Resource

The Uganda genome resource represents the largest published genomic study of continental Africans to date[39]. This study leveraged already existing strategic collaboration between the Uganda Virus Research Institute, and the University of

Cambridge and Sanger Institute in the United Kingdom. In 1989, the Uganda General Population Cohort was established by the Uganda Virus Research Institute and partners to examine trends in prevalence and incidence of HIV infection and their determinants[40]. A genomic study of communicable and non-communicable diseases was then launched in 2011 with this same cohort**.** The successful implementation of genomic research here can be attributed to existing local infrastructure in Uganda, long-standing collaborations with genomic centres of excellence in the UK, and strategic funding that included a research capacity-building component. For example, the author Segun Fatumo is a former H3Africa Bioinformatics Network (H3ABioNet) fellow who was funded to do postdoctoral research training in statistical genetics and bioinformatics at the Sanger Institute and University of Cambridge. During this training, he was strategically positioned to take a lead role in analyses of the Uganda genome resource. Following this training and research, Segun Fatumo has since continued to maintain the genomic resources locally, in addition to leading other genomic studies[41–44]. Furthermore, this resource has enabled significant new insights for population genetics and genetic epidemiology. For example, a genetic variant known to cause the inherited blood disorder alpha thalassemia was significantly associated with glycated haemoglobin, a biomarkercommonly used in the diagnosis of diabetes[39]. This variant is thought to have become more frequent among African populations because it can prevent severe malaria[39].

4.1.2 Building on existing resources - Africa Wits-INDEPTH partnership for Genomic Studies (AWI-GEN)

AWI-Gen is an NIH funded cross-sectional population cohort of about 12,000 older adults (40-60 years) from 6 centres spanning 4 African countries - Ghana, Burkina Faso, Kenya, and South Africa. It was set up by a strategic regional partnership between the University of the Witwatersrand, Johannesburg and the International Network for the Demographic Evaluation of Populations and Their Health (INDEPTH) study. The existing Health and Demographic Surveillance System centres and the Developmental Pathways for Health Research Unit have longitudinal cohorts which provided the research infrastructure, including long-standing community engagement, trained fieldworkers, and detailed longitudinal demographic and phenotype data. This mutually beneficial partnership enabled the project to span Africa with a wide

representation of social and genetic variability that has resulted in more than 40 publicationsacross disciplines including epidemiology, disease awareness, population genetics, candidate gene studies, and gene-environment interaction[45–49]. Several major GWASs are close to publication and have led to partnerships with large, global consortia such as the Global Lipid Genetics Consortium and Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) study. Additional funding from these partnerships has enabled the transformation of AWI-Gen into a longitudinal cohort. The achievements of the AWI-Gen study are in part attributable to the strategy of building on existing resources and forming long-term partnerships based on benefit sharing among institutions within LMIC settings.

Beyond the research itself, a major achievement of these studies lies in the sharing of bioinformatics and genomics skills across the continent. For example, the annual Introduction to Bioinformatics course run by the H3A-BioNet (Bioinformatics network of the H3A consortium) has trained over three thousand students in the last 8 years [50]. In addition, the network has hosted more than 30 workshops for basic and advanced training in areas such as GWAS, NGS, microbiome analysis and data management [38,50]. Similarly, the set up and development of several biobanks across the continent in association with these projects could have a catalytic effect for research and development initiatives in future. Finally, as these studies reach completion we anticipate that some of the outcomes will benefit the communities who participated and will also contribute to the bio-economic landscape of the respective LMICs.

4.2 ASIA

4.2.1 The importance of funding: Pakistan Alliance on genetic RisK factors for Health (PARKH)

South Asians make up one sixth of the world population, with 1.38 billion people living in India alone. Pakistan and many other countries in the region have a high rate of consanguineous marriages and have been the focus of gene mapping studies for recessive disorders for the last few decades. There is a long list of disorders for which mutations have been discovered in families from these regions including

hearing impairment [51], intellectual disability [52], microcephaly [53] and visual conditions [54]. These studies have contributed to the global efforts for study of genetic causes of recessive disorders and their underlying biology. In the process genotyping and sequencing, data has been created that can now be leveraged to address questions about population structure, population specific allele frequencies and ancestry [55]. This will require collaborative networks, data storage and access mechanisms that follow ELSI guidelines. The Greater Middle East (GME) Variome Project is one such successful example (GME (ucsd.edu).

However, South Asians are particularly underrepresented in genomic research of complex diseases. With a target recruitment of 30,000 psychiatric patients and 15,000 control participants, PARKH (Pakistan Alliance on genetic RisK factors for Health) is one of the largest international case control studies utilizing genetic data. Over a period of 20 years, the team have built extensive links with other institutions across Pakistan through small family-based studies[52,56,57], which eventually enabled a sizable pilot sample collection. Local connections, cultural understanding, knowledge of the administrative and regulatory processes, resilience, and the flexibility to navigate an ever-changing research landscape have been the key factors in the success of these projects. The collaboration between Pakistani, US- and UK-based researchers was a decisive factor in opening up access to funding resources. For example, one of the three PARKH sister studies, DIVERGE, is funded by a starting grant worth €2.5 million from the European Research Council, for which only researchers in the European Union and a select group of partner countries are eligible. The two other sister studies, the GENetics of SChizophRenia In Pakistan (GEN-SCRIP) and GENetics of BipoLar Disorder In Pakistan (GEN-BLIP) have been funded by the US National Institute of Mental Health (award numbers R01MH112904-01 and R01MH12377, respectively). PARKH demonstrates that building and maintaining infrastructure and a network for data collection as well as international collaborations can be the foundation for repeated funding success and may serve as motivation for ambitious strategies at large-scale. In the case of PARKH, none of the funders provided a dedicated capacity

building component. Rather, the investigators implemented their own strategies that included hiring local researchers for diverse roles.

Study design can also play an important role in enabling sustained research activity. For the DIVERGE study, a dedicated cross-disciplinary working group designed a protocol that captures diverse outcomes and putative risk factors for depression to enable multidisciplinary research on depression genetics, pharmacogenetics, interactions between genes and traumatic life events, and epidemiological analyses of socioeconomic factors. Importantly, local investigators took key roles in the study design to ensure that factors relevant to the studied populations were captured in the data collection.

4.3 LATIN AMERICA

Consortium-building for aggregation of large-scale genomic data - The Latin American Genomics Consortium

The term ‘Latin American’ refers to a pan-ethnicity used for the large, diverse group of people who come from Latin American countries. Additionally, people in other countries who identify with Latin American origins are often identified as Hispanic or Latinx American. Latinx populations have complex ancestry including recent admixture. Commonly used analytical approaches may not sufficiently address population stratification in these groups; for example, the use of principal components as covariates (whereby a large set of variables is condensed into a smaller, more simplistic set) only accounts for global ancestry but not for local ancestry for a given genomic region. In addition to the lack of dedicated genomic studies in these groups, individuals with admixed ancestry are systematically excluded from existing studies due to these concerns around population stratification. The recently established Latin American Genomics Consortium aims to address these issues within the field of psychiatric genetics (https://latinamericangenomicsconsortium.org). This consortium includes over 100 scientists from eight Latin American countries, Puerto Rico and the USA. The group harmonises data from existing cohorts and has a total of 100,000 samples, mostly from the USA, but there are plans to recruit new participants and establish a biobank.

The development of analytical methods for samples with admixed ancestry is an active field of research. One promising albeit computationally intensive approach is a software framework known as Tractor — it identifies haplotype segments and assigns them to ancestral origins, followed by an ancestry-specific association analysis[58].

4.4 AUSTRALIA

4.4 The importance of the community in setting research priorities - The Tiwi Island Aboriginal Population.

Aboriginal and Torres Strait Islander people in Australia are one of the largest indigenous populations in the world. They comprise hundreds of groups, each with their own distinct language, history, and cultural traditions. The Tiwi Land Council signed an historic research agreement to formalize Tiwi control of research priorities, research information, and samples including biobanking in genomic studies[60]. The Tiwi people have therefore proactively participated and engaged with research into kidney disease and other chronic conditions in their community for more than 30 years or more, with stakeholders providing ethical guidance for researchers and support for communities themselves[59]. At one point, the Tiwi community raised local financial support and external funds, specifically the Stanley Tipiloura Fund, to support research into kidney disease[61].

Crucially, members of the Tiwi community have worked as staff in all research projects conducted within their community[61], and have contributed to the application of genetics research to study its origins, migrations, customs, relationships, and health issues[61]. The Tiwi Island Aboriginal Population is therefore an example of best practice for indigenous-led initiatives with a substantial proportion of indigenous researchers and leaders. This is further illustrated by the recently launched National Centre for Indigenous Genomics (NCIG) which not only demonstrates genuine partnerships with community but is also governed by an indigenous-majority board.

4.5 Collective lessons or Key learnings

The success of the cohorts and studies described above illustrate that with sufficient funding, it is possible for indigenous groups and those at LMIC institutions to scale up in resources and skills to enable high-quality genomics research in less than a decade.

These examples should motivate funders to support both ongoing and new ventures that are led by LMIC researchers. Moreover, publications in top tier journals and presentations at major conferences have provided them the opportunity to participate inlarge-scale, global studies. We hope that in future they would not only be able to extend their research to larger cohorts but would also be able to move closer to leading some of these large-scale global studies. As an example of this, two key contributors to the AWI-Gen study (including one of the authors of the current manuscript, Tinashe Chikoware) were recently provided the opportunity to co-lead one of the CHARGE consortium Phase 2 studies.

5.0 A Roadmap for establishing sustainable diverse genomics research worldwide.

Based on our experiences in setting up genomic studies in diverse populations, we recommend key priority steps (Figure 3) which we discuss in detail below.

5.1 Stakeholder will

The importance of diversity in research studies has been known for a long time and is evident in legislation and guidelines, such as those enacted in the USA in 1993 to increase participation of women and minority groups in clinical studies [NIH Revitalization Act of 1993 Public Law 103-43. Federal Register, 59FR14508]. However, participation of the minority groups such as Hispanics and African-Americans has remained limited in America[20]. The lack of diversity in genomics requires boldness and willingness of the varied stakeholders, including research institutions, researchers, participants, funders and governments, to collaboratively work together to address this imbalance. To help correct the lack of diversity in genomic research, several key ingredients are needed.

First, research institutions must be willing to ensure they have a diverse workforce. This has been shown to improve trust among minority groups, leading to improved recruitment. Diverse researchers have been reported to be more interested in studying about their population groups, thereby increasing diversity in genomics[55] . Notably, programs such as the NIH UNITE have been set up to address structural racism in the workplace and ensure diverse researchers have equitable access to opportunities in

biomedical research. In view of the global nature of research, there is a need for institutions that support open access to research outputs which will help other researchers to carry forward similar work globally, and also to replicate findings in diverse settings.

Next, researchers must to be willing to form genuine partnerships with communities that result in ethical conduct of genetic research which benefits all[22]. In order to address the historical perceptions and distrust of clinical research by minority groups, researchers should take time to engage in dialogue about the goals of genomic studies and clarify concerns of potential harm—ultimately leading to integration of participants values and expectations in the implementation of genomic studies[27].

The willingness of research participants from minority groups to participate in genomic studies is key to the success of these studies. When participants trust the researchers and their governments, they are not only more willing to participate, but may even offer broad consent in BioBank studies[62] – thereby indicating that if researchers and government work together to ensure ethical and trustworthy research is conducted, more minority groups will likely participate. However, there is a need for more research to inform policy with regards to who benefits from commercialisation of the research outputs and how genomic sovereignty can be maintained in the context of broad or tiered consent.

Studies that are focused on cohorts from previously marginalized populations have the additional burden of managing the damage that has been caused by earlier studies, in which an extractive attitude coupled with a lack of engagement with the community and under-appreciation of their beliefs and sentiments has led to a general distrust in researchers. In addition to an extensive and prolonged engagement with such communities, it is crucial that research be focused on areas that are health priorities for the respective communities and that have a potential to bring about tangible benefits to them. Only through such an approach can these communities come to view researchers as allies and partners.

Funders must be willing to set up strategic schemes which promote research of underrepresented population groups. Genomic research in underrepresented

population groups has been noted to require more time and resources and funders need to be able to commit to this. Most scientists from these population groups have a lower competitive edge compared to those of European ancestry and they will need earmarked funding to ensure they can grow their capacity to compete for grants in the future. The H3Africa and the Data Science for Health Discovery and Innovation in Africa are examples of strategic funding commitments by the NIH to bolster genetic research in Africa.

Finally, governments must be willing to institute policies that create environments conducive to sustainable, diverse genomics studies. A number of governments are realising the potential and value of genomic studies even among underrepresented populations; examples such as the China Kadoorie Biobank and the South African Human Genome Project offer hope that more governments might take such steps and sustainable diverse genomics may become a reality.

5.2 Funding

Genetic research is expensive, making it a secondary priority for funding in LMIC. One route towards greater inclusion of underrepresented populations is by leveraging funding mechanisms from international institutions and those in research-intensive nations. Funders have an opportunity to help address imbalances in global genetics research through their research priorities; dedicated funding calls, such as the 'Genetic Architecture of Mental Disorders in Ancestrally Diverse Populations' by the National Institute of Mental Health in the US, can be a strategic tool to empower fast progress.

Barriers to access

Many funding calls are exclusively targeted to researchers at institutions in the funder's country. Given the immense and wide-reaching benefits of increasing diversity in genetic research, funders should reconsider such restrictions. In addition to eligibility restrictions, fewer researchers in LMICs have track records competitive for large funding calls due to the limited research capacity, infrastructure, and funding at their local institutions. This catch-22 makes it very difficult for those researchers to build up large genomic studies without collaborators from research intensive nations.

Collaboration

For most of the case studies we have presented here, collaborations between local investigators and those from research-intensive nations were critical for funding success. Collaborations can provide diverse expertise that includes competitive research track records, experience in grant writing, administrative support, and the necessary local expertise and knowledge about the target population. Therefore, networking and building long-lasting productive collaborations remains a key route for investigators to access funding for large-scale genetics research. However, the potential for power imbalance needs to be considered when establishing collaborations with instutites from research intensive nations, as well as the potential for negative reactions by some members of local communities to initiatives led by foreigners. When capacity-building is incorporated, the collaborative approach may eventually support local expertise to enable more genomic research led by investigators in LMIC's. Moreover, data-sharing agreements are important to ensure the interests of the local researcher are respected..

Sustainability

Sustainability should be a primary consideration for awarded funds to most effectively improve the diversity of genomic studies long-term. Many funding calls do not provide a dedicated capacity-building component. In these cases, researchers can still invest funds to enhance local capacity for long-term benefits, such as by hiring local students or researchers for training or research positions (see also Capacity Building).

5.3 Infrastructure and administrative components

To conduct cohort-based genomic research, it is not only critical to access some key infrastructure components but to align the study with the legal, administrative and ethical frameworks applicable at the institutional and national level. **(Table 1)**. A comprehensive understanding of ethical concerns, regulations and policies could enable researchers to avoid major delays in cross-border shipping of biological samples and also to ensure the ability to re-use/share these valuable datasets in future. Most of the studies described above report pre-study consultation with legal experts (often available via their institutions) and implementation of necessary material and data transfer agreements to ensure efficient movement of samples and data.

Infrastructure for steps such as sample processing, biobanking, genotyping or sequencing and computational analysis are often outsourced or accessed via local and international collaborations (**Table2**). However, developing the ability and infrastructure to be able to do one or more of these at the institutional level could be a major capital for securing continued funding for the study and future research.

5.4 Capacity Building

To narrow gaps in genomic studies for underrepresented populations, education models that retain trained individuals are critical; these provide knock-on opportunities to transfer technology and knowledge locally, thereby creating a critical mass of appropriately trained individuals.

For example, capacity development has been one of the key aims of the AWI-Gen study. In addition to training over 20 postgraduate students and postdoctoral fellows in statistical genetics, the consortium has been a key contributor to several major GWAS training initiatives on the continent. This includes hosting or organizing courses and workshops independently as well as in partnership with bodies such as H3Africa Bioinformatics Network, Wellcome Trust Overseas course, and Sweden South Africa University Forum. AWI-Gen has also been a key contributor to the development of the H3Africa GWAS pipeline and imputation facility that is anticipated to help future genomics research on the continent. However, like most other studies in LMIC settings, retaining trained students and scientists continues to be a challenge that AWI-Gen must deal with.

In Pakistan, the PARKH team utilizes their international links to support researchers to visit labs in the US and Canada for training. These visits were organized in collaboration with the Higher Education Commission of Pakistan that guarantees that scholars return to work in their institutions. PARKH have formed virtual analysis teams bringing together experts in the United States and Canada and trainees and junior faculty from Pakistan. Senior researchers from PARKH collaborative network have co-supervised graduate students from Pakistani universities.

5.5 Partnership with Global Consortia

Increasing diversity in genomic studies contributes to more robust findings from replicated results as well as novel discoveries, particularly when combined with existing large-scale studies. Developing local research capacity enables contributions to global genomics consortia, as demonstrated in several consortia already such as the Global Lipids Genetics Consortium[63], GIANT Consortium, Psychiatric Genomics Consortium and other major initiatives. These have dual and mutual benefits by enabling the discovery of ancestry-specific findings, raising the profile of these findings to a broader audience, and enhancing the careers of local contributing investigators. Participation in global consortia by diverse groups requires trust, which can only be built when all contributors benefit.

Conclusion

Despite some notable efforts, representation of non-European ancestry groups in genetic research remains low, and this affects diverse global populations. The benefits of greater diversity extend beyond the studied population. We present a vision with a concrete roadmap on how to address this imbalance; leveraging established local infrastructure and offering strategic funding that is tied to capacity building could empower sustainable global research. To be successful in achieving equitable inclusion of underrepresented groups in genomic studies, the stakeholders must stimulate local participation, build trust, and ensure mutual respect.

Figure Legends:

Figure 1: The proportion of samples from individuals cumulatively reported by GWAS Catalog[1] as of July 8, 2021.

Figure 2. Disparities in representation of continents in genomic studies will grow wider in the next few years without immediate measures to increase diversity**.** Upcoming large-scale (>100 K participant) cohort-based studies included within the IHCC were employed as an indicator of the representation of various continents in genomics research over the next few years. (a) Number of enrolled participants from each geographic region (b) Number of cohorts from each geographic region. The estimates are based on cohorts that are collecting, or aim to collect, genomic data (https://ihccglobal.org/membercohorts/).

Figure 3: Roadmap showing the key pillars for setting up and sustaining diverse global genomic studies.

Reference

1. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
2. Patin, E. *et al.* Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science* **356**, 543–546 (2017).
3. Auton, A. & Salcedo, T. The 1000 genomes project. in *Assessing Rare Variation in Complex Traits: Design and Analysis of Genetic Studies* 71–85 (Springer New York, 2015). doi:10.1007/978-1-4939-2824-8_6.
4. Fan, S., Hansen, M. E. B., Lo, Y. & Tishkoff, S. A. Going global by adapting local: A review of recent human adaptation. *Science* **354**, 54–59 (2016).
5. Majara, L. *et al. Low generalizability of polygenic scores in African populations due to genetic and environmental diversity*. 2021.01.12.426453 https://www.biorxiv.org/content/10.1101/2021.01.12.426453v1 (2021) doi:10.1101/2021.01.12.426453.
6. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
7. Huang, Q. Q. *et al. Transferability of genetic loci and polygenic scores for cardiometabolic traits in British Pakistanis and Bangladeshis*. 2021.06.22.21259323 https://www.medrxiv.org/content/10.1101/2021.06.22.21259323v1 (2021) doi:10.1101/2021.06.22.21259323.

8. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).
9. Asimit, J. L., Hatzikotoulas, K., McCarthy, M., Morris, A. P. & Zeggini, E. Trans-ethnic study design approaches for fine-mapping. *Eur. J. Hum. Genet. EJHG* **24**, 1330–1336 (2016).
10. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
11. Genovese, G. *et al.* A risk allele for focal segmental glomerulosclerosis in African Americans is located within a region containing APOL1 and MYH9. *Kidney Int.* **78**, 698–704 (2010).
12. Rotimi, C. N. *et al.* The genomic landscape of African populations in health and disease. *Hum. Mol. Genet.* **26**, R225–R236 (2017).
13. Cohen, J. *et al.* Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat. Genet.* **37**, 161–165 (2005).
14. Gao, C. *et al.* Risk of Breast Cancer Among Carriers of Pathogenic Variants in Breast Cancer Predisposition Genes Varies by Polygenic Risk Score. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **39**, 2564–2573 (2021).
15. Inouye, M. *et al.* Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. *J. Am. Coll. Cardiol.* **72**, 1883–1893 (2018).
16. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
17. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).

18. Scutari, M., Mackay, I. & Balding, D. Using Genetic Distance to Infer the Accuracy of Genomic Prediction. *PLoS Genet.* **12**, e1006288 (2016).
19. Sengupta, D., Choudhury, A., Basu, A. & Ramsay, M. Population Stratification and Underrepresentation of Indian Subcontinent Genetic Diversity in the 1000 Genomes Project Dataset. *Genome Biol. Evol.* **8**, 3460–3470 (2016).
20. Oh, S. S. *et al.* Diversity in Clinical and Biomedical Research: A Promise Yet to Be Fulfilled. *PLoS Med.* **12**, e1001918 (2015).
21. Bentley, A. R., Callier, S. & Rotimi, C. N. Diversity and inclusion in genomic research: why the uneven progress? *J. Community Genet.* **8**, 255–266 (2017).
22. Tindana, P. *et al.* Community engagement strategies for genomic studies in Africa: a review of the literature. *BMC Med. Ethics* **16**, 24 (2015).
23. Hindorff, L. A. *et al.* Prioritizing diversity in human genomics research. *Nat. Rev. Genet.* **19**, 175–185 (2018).
24. Tan, S.-H., Petrovics, G. & Srivastava, S. Prostate Cancer Genomics: Recent Advances and the Prevailing Underrepresentation from Racial and Ethnic Minorities. *Int. J. Mol. Sci.* **19**, E1255 (2018).
25. Reverby, S. M. Ethical Failures and History Lessons: The U.S. Public Health Service Research Studies in Tuskegee and Guatemala. *Public Health Rev.* **34**, 1–18 (2012).
26. Löwy, I. The best possible intentions testing prophylactic approaches on humans in developing countries. *Am. J. Public Health* **103**, 226–237 (2013).
27. Kraft, S. A. *et al.* Beyond Consent: Building Trusting Relationships With Diverse Populations in Precision Medicine Research. *Am. J. Bioeth.* **18**, 3–20 (2018).
28. *The 10/90 report on health research 2000*. (2000).

29. McGregor, S., Henderson, K. J. & Kaldor, J. M. How are health research priorities set in low and middle income countries? A systematic review of published reports. *PloS One* **9**, e108787 (2014).
30. Sridhar, D. Who sets the global health research agenda? The challenge of multi-bi financing. *PLoS Med.* **9**, e1001312 (2012).
31. Mbaye, R. *et al.* Who is telling the story? A systematic review of authorship for infectious disease research conducted in Africa, 1980-2016. *BMJ Glob. Health* **4**, e001855 (2019).
32. Stein, C. M. Challenges of Genetic Data Sharing in African Studies. *Trends Genet. TIG* **36**, 895–896 (2020).
33. Wright, G. E. B., Koornhof, P. G. J., Adeyemo, A. A. & Tiffin, N. Ethical and legal implications of whole genome and whole exome sequencing in African populations. *BMC Med. Ethics* **14**, 21 (2013).
34. Ascencio-Carbajal, T., Saruwatari-Zavala, G., Navarro-Garcia, F. & Frixione, E. Genetic/genomic testing: defining the parameters for ethical, legal and social implications (ELSI). *BMC Med. Ethics* **22**, 156 (2021).
35. Choudhury, A. *et al.* Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nat. Commun.* **8**, 2062 (2017).
36. Malaria Genomic Epidemiology Network. A global network for investigating the genomic epidemiology of malaria. *Nature* **456**, 732–737 (2008).
37. H3Africa Consortium *et al.* Research capacity. Enabling the genomic revolution in Africa. *Science* **344**, 1346–1348 (2014).

38. Choudhury, A., Sengupta, D., Aron, S. & Ramsay, M. *The H3Africa Consortium: Publication Outputs of a Pan-African Genomics Collaboration (2013 to 2020)*. 257–304 (Brill, 2021). doi:10.1163/9789004500228_011.
39. Gurdasani, D. *et al.* Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa. *Cell* **179**, 984-1002.e36 (2019).
40. Asiki, G. *et al.* The general population cohort in rural south-western Uganda: a platform for communicable and non-communicable disease studies. *Int. J. Epidemiol.* **42**, 129–141 (2013).
41. Chikowore, T. *et al. Polygenic prediction of type 2 diabetes in continental Africa*. 2021.02.11.430719 https://www.biorxiv.org/content/10.1101/2021.02.11.430719v1 (2021) doi:10.1101/2021.02.11.430719.
42. Fatumo, S. The opportunity in African genome resource for precision medicine. *EBioMedicine* **54**, 102721 (2020).
43. Fatumo, S. *et al.* Discovery and fine-mapping of kidney function loci in first genome-wide association study in Africans. *Hum. Mol. Genet.* **30**, 1559–1568 (2021).
44. Fatumo, S. *et al.* Metabolic Traits and Stroke Risk in Individuals of African Ancestry: Mendelian Randomization Analysis. *Stroke* **52**, 2680–2684 (2021).
45. Boua, P. R. *et al.* Novel and Known Gene-Smoking Interactions With cIMT Identified as Potential Drivers for Atherosclerosis Risk in West-African Populations of the AWI-Gen Study. *Front. Genet.* **10**, 1354 (2019).
46. Dlamini, S. N. *et al.* Associations Between CYP17A1 and SERPINA6/A1 Polymorphisms, and Cardiometabolic Risk Factors in Black South Africans. *Front. Genet.* **12**, 687335 (2021).

47. Gómez-Olivé, F. X. *et al.* Regional and Sex Differences in the Prevalence and Awareness of Hypertension: An H3Africa AWI-Gen Study Across 6 Sites in Sub-Saharan Africa. *Glob. Heart* **12**, 81–90 (2017).
48. Nonterah, E. A. *et al.* Classical Cardiovascular Risk Factors and HIV are Associated With Carotid Intima-Media Thickness in Adults From Sub-Saharan Africa: Findings From H3Africa AWI-Gen Study. *J. Am. Heart Assoc.* **8**, e011506 (2019).
49. Sengupta, D. *et al.* Genetic substructure and complex demographic history of South African Bantu speakers. *Nat. Commun.* **12**, 2080 (2021).
50. Aron, S. *et al.* The Development of a Sustainable Bioinformatics Training Environment Within the H3Africa Bioinformatics Network (H3ABioNet). *Front. Educ.* **6**, 356 (2021).
51. Acharya, A., Schrauwen, I. & Leal, S. M. Identification of autosomal recessive nonsyndromic hearing impairment genes through the study of consanguineous and non-consanguineous families: past, present, and future. *Hum. Genet.* (2021) doi:10.1007/s00439-021-02309-9.
52. Harripaul, R. *et al.* Mapping autosomal recessive intellectual disability: combined microarray and exome sequencing identifies 26 novel candidate genes in 192 consanguineous families. *Mol. Psychiatry* **23**, 973–984 (2018).
53. Khan, N. M. *et al.* Updates on Clinical and Genetic Heterogeneity of ASPM in 12 Autosomal Recessive Primary Microcephaly Families in Pakistani Population. *Front. Pediatr.* **9**, 695133 (2021).
54. Khan, A. A. *et al.* P.arg102ser is a common Pde6a mutation causing autosomal recessive retinitis pigmentosa in Pakistani families. *JPMA J. Pak. Med. Assoc.* **71**, 816–821 (2021).
55. Manolio, T. A. Using the Data We Have: Improving Diversity in Genomic Research. *Am. J. Hum. Genet.* **105**, 233–236 (2019).

56. Knight, H. M. *et al.* Homozygosity mapping in a family presenting with schizophrenia, epilepsy and hearing impairment. *Eur J Hum Genet* **16**, 750–758 (2008).
57. Hampshire, D. J. *et al.* MORM syndrome (mental retardation, truncal obesity, retinal dystrophy and micropenis), a new autosomal recessive disorder, links to 9q34. *Eur J Hum Genet* **14**, 543–548 (2006).
58. 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).
59. Kowal, E. & Anderson, I. *Genetic research in Aboriginal and Torres Strait Islander communities : continuing the conversation*. (Lowitja Institute, 2012).
60. Elsum, I. *et al.* Inclusion of Indigenous Australians in biobanks: a step to reducing inequity in health care. *Med. J. Aust.* **211**, (2019).
61. Thomson, R. J. *et al.* New Genetic Loci Associated With Chronic Kidney Disease in an Indigenous Australian Population. *Front. Genet.* **10**, 330 (2019).
62. Gaskell, G. *et al.* Publics and biobanks: Pan-European diversity and the challenge of responsible innovation. *Eur. J. Hum. Genet. EJHG* **21**, 14–20 (2013).
63. Klarin, D. *et al.* Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat. Genet.* **50**, 1514–1523 (2018).

Table 1. Details of pre-study administration the LMIC cohorts described in this article

Supplementary Table 1. Details of pre-study administration the LMIC cohorts described in this article	
Area	Comment
Ethics and regulations	Interactions with IRB, identifying relevant government and other statutory bodies and requesting necessary permissions is critical. Lack of ethical framework for aspects such as biobanking might be a problem in LMIC settings.
Community engagement	Finding the right people who represent the community, understanding and addressing their aspirations and concerns. Finding the correct language to communicate the nuances. Aids such as video, comics might be helpful. Consent documents might need translation to local languages. Moreover, vocabularies might not contain the exact terms, so a conceptual translation instead of literal translation might be required.
legal	Material transfer agreements, country specific modalities of fund transfer, tax implications, customs regulations needs to be identified and formalized.
Funding infrasturcutre	Dedicated personnel/bodies with expertise and experience for grant administration and management is helpful.
Other	The field staff are often benefited by a focused training for sample collection. A careful development of the questionnaire is necessary. Existing questionnaires need to be modified to encompass the variables specific to local settings. For example, chewing tobacco or smoking tobacco in forms other than cigarettes may be uncommon in some regions but quite common in others.

Table 2. Origin of the infrastructure employed by some of the LMIC cohorts described in this article. These have been broadly categorized into Study (owned by/genrated for the study), Local (Shared with local and national insitutions and service providers) and External (International collaborators and suppliers)

Study step	AWI-Gen	UGR	PARKH	Tiwi Islander	Comments
Sample collection	Study	Study	Study	Study	The sample collection infrastructure includes- basic devices for physical measurements and training of field staff for accurate and reproducible measurements; set up and SOP fo interviewing participants and recording their inputs digitally (either on the field or as a follow up); a module for labeling the tubes/aliquots for blood, body fluid and other biological sample collections.
Sample processing	Study	Study	External	External	If the institutional settings are limited, instead of doing this locally processing might be done via collaborators or service providers. However, this aspect needs to be considered and planned for.
Sample storage	Study	Study	External	External	Even if it is not necessary to have a full facility a the project site, having a partnership with a secure and fully authenticated biobank could be valuable in the long run. Moreover, devoting resources to be able to store at least a part of the samples on site, even if for short time scales, could be handy from a logistics point of view.
Genotyping	External	External	External	External	This might be especially challenging for LMIC settings. Collaboration with service providers- government, private or academic institutions might be required. Also, some level of resource (such as packaging and dry ice) and training might be required depending on the type of the biological samples.
Computational facility and resources	Local	Local	Local	Local	Although processing genotype data from a sma cohort is often possible with minimal resources, partnering with high-performance computing facilities at the national level or at local universities could facilitate the process significantly. Also, a proper policy and mechanism for determining who can access the data and how it needs to be used needs to be arrived upon. Some level of training for data QC and management before the arrival of the is recommended.

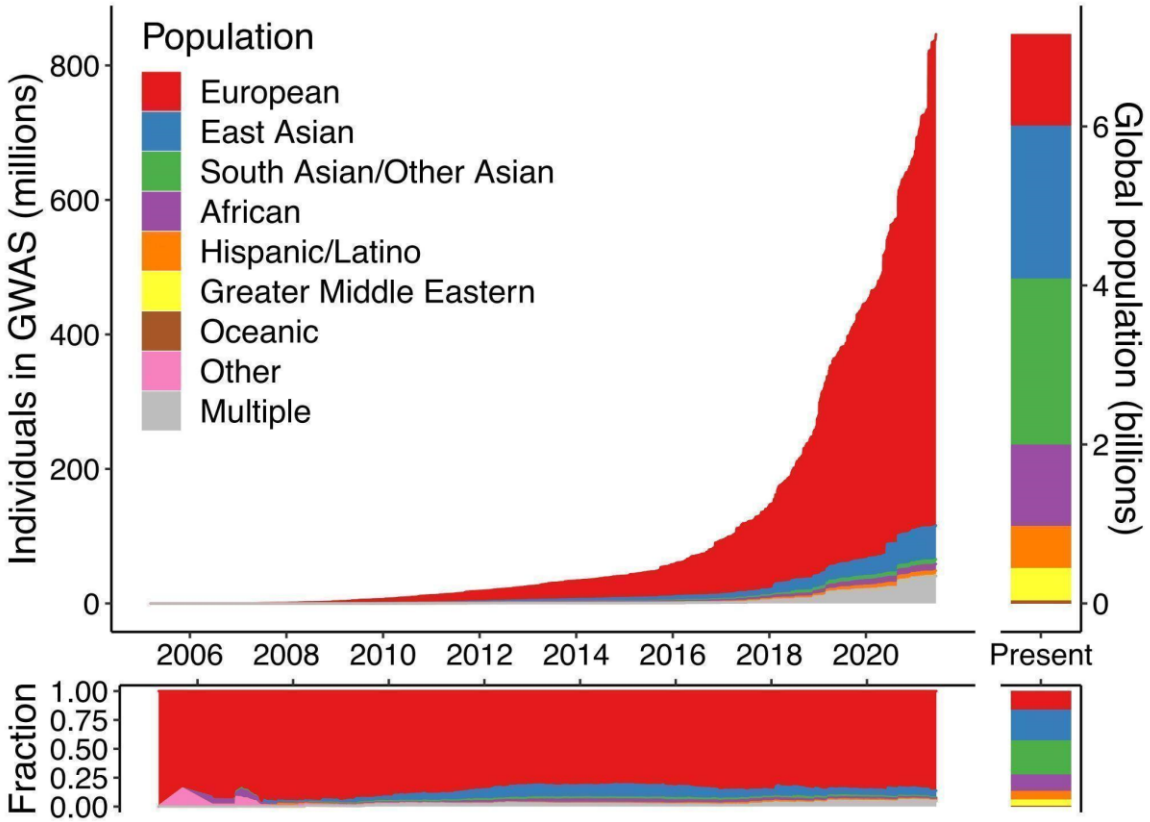
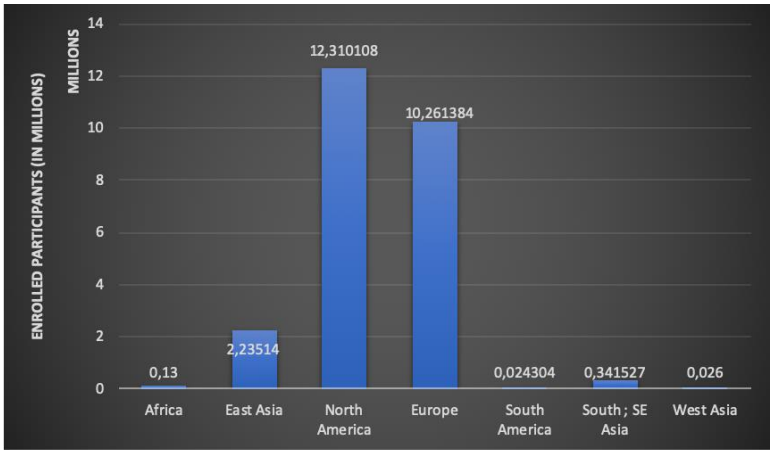


Figure 1: The proportion of samples from individuals cumulatively reported by GWAS Catalog [1] as of July 8, 2021



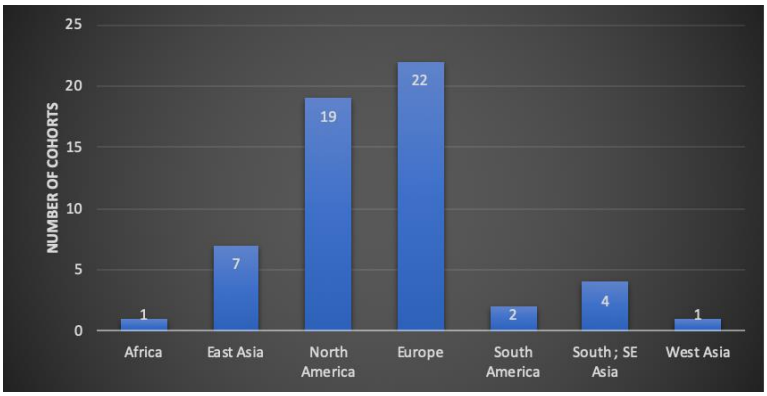


Figure 2. Disparity in representations of continents will increase in the next few years without immediate measures to increase diversity in genomic studies**.** Upcoming large-scale (>100 K participant) cohort-based studies included within the IHCC was employed as an indicator of the representation of various continents in genomics research over the next few years. (a) Number of enrolled participants from each geographic region (b) Number of cohorts from each geographic region. The estimates are based on cohorts that are collecting or aim to collect genomic data (https://ihccglobal.org/membercohorts/).

Figure 3: Roadmap showing the key pillars for setting up and sustaining diverse global genomic studies.