The critical needs and challenges for genetic architecture studies in Africa
Alicia R Martin1,2,3, Solomon Teferra4,5, Marlo Möller6, Eileen G Hoal6 and Mark J Daly1,2,3,7

Human genetic studies have long been vastly Eurocentric, raising a key question about the generalizability of these study findings to other populations. Because humans originated in Africa, these populations retain more genetic diversity, and yet individuals of African descent have been tremendously underrepresented in genetic studies. The diversity in Africa affords ample opportunities to improve fine-mapping resolution for associated loci, discover novel genetic associations with phenotypes, build more generalizable genetic risk prediction models, and better understand the genetic architecture of complex traits and diseases subject to varying environmental pressures. Thus, it is both ethically and scientifically imperative that geneticists globally surmount challenges that have limited progress in African genetic studies to date. Additionally, African investigators need to be meaningfully included, as greater inclusivity and enhanced research capacity afford enormous opportunities to accelerate genomic discoveries that translate more effectively to all populations. We review the advantages, challenges, and examples of genetic architecture studies of complex traits and diseases in Africa. For example, with greater genetic diversity comes greater ancestral heterogeneity; this higher level of understudied diversity can yield novel genetic findings, but some methods that assume homogeneous population structure and work well in European populations may work less well in the presence of greater heterogeneity in African populations. Consequently, we advocate for methodological development that will accelerate studies important for all populations, especially those currently underrepresented in genetics.

Addresses
1 Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA
2 Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
3 Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
4 Department of Psychiatry, School of Medicine, College of Health Sciences, Addis Ababa University, Addis Ababa, Ethiopia
5 Department of Epidemiology, Harvard T. H. Chan School of Public Health, Harvard University, Boston, USA
6 DST-NRF Centre of Excellence for Biomedical Tuberculosis Research, South African Medical Research Council Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, Cape Town, South Africa
7 Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland

Corresponding author: Martin, Alicia R (armartin@broadinstitute.org)

Historical biases in genetic studies
Nearly a decade ago, 96% of participants in genome-wide association studies (GWAS) were of European descent [1]. While European individuals now account for 78% of GWAS participants [2], the non-European proportion has stagnated since 2014. African ancestry individuals constitute merely 2.4% of participants (although notably account for 7% of all associations) [2]. This participant bias results in interpretability gaps by ancestry with medically relevant consequences [3,4]. For example, while easily avoidable, African American patients were more likely than white Americans to be incorrectly told they have a genetic mutation that increases their risk of hypertrophic cardiomyopathy, an early-onset life-threatening heart disease, at leading genetic testing labs [5]. Additionally, drug metabolism genes such as CYP3A4 contain mutations that can alter dosage requirements, but pharmacogenetic variants are disproportionately uncatalogued among African populations [6], so genotype-based dosage guidelines are less useful. In the US, the National Human Genome Research Institute has prioritized increased diversity in genetic studies [7]. This prioritization is an important step that, if heeded, will aid interpretations in medical genomics for all ethnicities [8]. Greater inclusivity of African populations in medical genomics is important for accelerating genomic discoveries, enabling reconstruction of modern human origins, producing results that can be translated across populations more accurately, identifying genetic associations with traits for variants absent elsewhere, and building research capacity in Africa.

Genetic study biases have not happened in a vacuum, but have had widespread consequences for GWAS tools and
resources in African populations. Genotyping arrays have traditionally been biased towards alleles most frequent and imputable in European populations [9,10], com- poundeding biases in which GWAS identify variant associa- tions most common in the study population [11*,12]. In contrast, array backbones prioritizing SNPs that maxi- mally tag variants across all populations improve imputation performance, providing more even genomic coverage [13]. Perhaps more importantly, imputation panels are vastly Eurocentric, shortchanging representation of the greater haplotypic diversity present in Africans from deeper recombination history [12,14,15]. The most widely available African sequencing resources have biased representation towards African Americans and West Africans [8,12], leaving huge swaths of African diversity uncatalogued.

Existing challenges to surmount for African genetics studies

To empower African genetic studies and build capacity for research aiding biological understanding across a diverse swath of humanity, we review challenges that need to be confronted and continually addressed.

Historical

Africa has long been subjected to a violent and oppressive colonial history that has bred suspicion and an anticipation of resource exploitation. This understandable mis- trust continues to strain ongoing relations, with new actors such as China in addition to European groups scrambling for African resources [16,17]. The impact on research collaborations is evident, with some authors discussing ‘neo-colonial science’ [18]. Such strained relations are more pronounced in collaborations involving genetic studies, especially when shipping samples out of Africa and the global south [19]. Some discuss ‘genomic sovereignty’ of Africans and ownership of African genetic material [20]. Proponents of international collaborations argue that working with high income countries will eventually ensure equity, justice, and benefit to Africans, with capacity building for genomic research providing imme- diate benefit for African institutions [21*], although con- cerns have been raised about the sustainability of these efforts. Ongoing tensions weigh the benefit to Africans by including more African researchers and DNA in global research against the challenges of promoting African science while integrating and importing the best science around the world into Africa (Figure 1).

Infrastructural

Conducting genetic studies in Africa is not an easy task. Infrastructural problems can include unreliable or no electricity in clinics and laboratories that process samples, impassable roads in some areas, and crime or political instability making some areas dangerous and/or inaccessible for researchers. Many African countries do not have sufficient laboratory equipment or facilities for genomics research, and most require imported reagents. Importing is not only time-consuming, but also costly—reagents are often many times more expensive in Africa than Western countries in real terms, not including shipment costs. Biobanks are less abundant, partially due to power inter- ruptions affecting storage and processing of samples. Some African institutions have experience in large-scale human genetic analyses; the H3Abiome consortium has developed core bioinformatics infrastructure in Africa [22*]. However, high-speed internet connections and powerful computers are not always available to access large data files. Human resource issues can also be a challenge, namely high staff turnover due to inadequate pay, competing demands for time from qualified staff, and/or too few qualified staff. Relatedly, brain drain is a major issue, as many skilled African scientists leave the continent in search of greener pastures [23,24]. To be sensitive to these challenges, some major international research initiatives such as H3Africa have required a relatively long embargo period on publication for African researchers [25*]. Connecting African researchers to ade- quate computing power (e.g. stable wireless connections to cloud computing) may offer more direct means to facilitate research. Compared with the relative ease of acquiring samples in the global north, the focus of data- banks on European/white populations is unsurprising, but it is nonetheless imperative that researchers rise to these challenges for the benefit of all.

Funding

Genetics research is expensive, and a lack of attention from African policy makers in resource-limited settings is primarily driven by competing priorities for more imme- diate public health concerns, including infectious
diseases over inherited conditions [26,27]. Data generation is still the most expensive part of genomics, whereas data analysis is more affordable and therefore a viable option for capacity building [28]. Furthermore, journals from the developed world often exist behind expensive pay-walls that are inaccessible to some researchers and do not always encourage publication of work from the global south, often returning manuscripts without review citing a lack of ‘sufficient general interest’. Having fewer publications has a knock-on effect on future grant funding and attracting students.

Nearly all funding for genetics research comes from outside Africa, raising questions for African scientists about the utility of investigating disease genetics with less long-term funding security and intellectual freedom to prioritize their field of study. Incentives differ from the West, heavily favoring medicine over research training—clinical demands are heavier, PhD programs are scarce, and research often does not pay. However, some external research funding in genomics, most notably by the Human Heredity and Health in Africa (H3Africa) Initiative, are being led by African scientists. H3Africa funding by the NIH (USA) and Wellcome Trust (UK) totals more than $216 million in 2015 for 185 projects in 28 African countries [29]. Its aim is to build the capacity for African scientists to conduct genomic research on heritable diseases afflicting Africans [30,31]. This international support is essential for African geneticists to continue their research [21*].

**Ethical**

Ethics review boards may lack familiarity with genomics research, which creates challenges for advising long-term, large-scale collaborative genetic studies that can in turn delay funded projects [28]. These challenges are partially driven by restrictive ethical guidelines and uncertainty about the benefits of such studies to African populations [28]. Unlike in the US, genetics projects are subject to ethics review both at the provincial and national levels as a legacy of colonialism, which can lead to years-long delays. Ethics approval by regulatory bodies in Africa is mostly restricted to project-specific research questions, often raising questions around ‘broad consent’ and ‘indefinite storage’ of samples that are not easy to answer. A primary concern about loss of control and ownership over the DNA samples arises when they are shipped abroad [32]. Burgeoning interest in building large-scale genomics collaborations in Africa has resulted in a recent best practices ethical framework for genomics research and biobanking in Africa [25*].

Some communities have set up local councils to oversee research projects and publication of results allowed from the research [33]. While these are excellent in theory, in practice there can be long delays, misunderstanding due to unfamiliarity of lay people with jargon, and a lack of continuity in leadership. Consequently, even when extensive consultation on planned or existing research projects has taken place, this often needs to be repeated at each subsequent visit. New council leaders sometimes try to enforce sample destruction before allowing further sampling, even when consent forms specify long sample storage. A middle ground of continuous community leadership from members more familiar with research methodology and terminology that is acceptable to the council would be ideal, but is often infeasible. Furthermore, while returning scientific discoveries to communities or participants should be the norm, re-contacting study participants in communities can be challenging as people lose cell phones or move for employment opportunities.

**Respect and consent**

To ensure mutual respect in collaborative African genomics studies, it is important to avoid generalizing ‘Africanness’ in such a vast continent, comprising not only more genetic diversity than the rest of the world, but also many cultures, language groups, and world views, some of which are marginalized or discriminated against. Thus, it is important to obtain perspectives from diverse continental Africans when communicating science broadly. Furthermore, meaningful engagement with African colleagues is vital to healthy collaborations and to avoid tokenism. Additionally, obtaining informed consent for genomics research can be complex in any setting, but poses more challenges where there are lower income and literacy levels or language barriers. Furthermore, some diseases such as mental illness are subject to greater stigma in some African communities, requiring cultural awareness and sensitivity to differences. Participants may misunderstand the study purpose or expect benefits that are not included, such as better disease treatment [34*] or individual-level ancestry results useful for land claims. Additionally, in some African societies, decisions to participate in research studies are made collectively as well as at the individual level [35], necessitating consultation with community leaders.

Communicating science respectfully can be challenging when nomenclature is subject to sociopolitical debate, as with the descendants of the original hunter-gatherers of Southern Africa. In an attempt to be politically correct, many population geneticists use ‘KhoeSan’ to refer to the Khoe and San groups collectively. However, the San Council of Southern Africa prefers to keep these terms separate (i.e., San and Khoe or Nama) to denote different cultures. Many ‘San’ individuals prefer being called ‘Bushmen’, while others consider the word to be pejorative. Labels are only useful insofar as they are universally informative, and respect is imperative. Wide pre-publication consultation is obviously necessary [36], but complete consensus is unlikely.
GWAS design challenges in Africa

Unlike most of the GWAS and complex trait studies that have been conducted in Europe, assumptions of homogeneous population structure are more likely to be violated in Africa, as few populations have remained isolated and unchanged over the past 4000 years [37]. This higher level of diversity across African populations relative to others [38,39] creates greater challenges when attempting to balance case/control distributions at the outset of many studies due to greater complexities in population structure, including variable LD patterns between study sites. Consequently, false positives are more likely to arise from confounding due to unaccounted population stratification, especially for rare variants, which are challenging to analyze [40,41]. Higher rates of genetic diversity also result in a larger number of effective tests, meaning that the standard multiple testing threshold of $p < 5 \times 10^{-8}$ needs to be roughly twice as stringent in African GWAS ($p < \sim 2.5 \times 10^{-8}$) [42]. Additional challenges arise from a dearth of large, easily accessible reference panels in Africa. While the African Genome Variation Project and related projects have worked to ameliorate this gap, data access is somewhat more challenging and slower than the publicly available 1000 Genomes Project [15].

Because some complex trait genetics methods assume homogeneity that is more often violated in African populations with higher diversity, methodological advancements that explicitly account for structure over a range of time periods will be especially useful [43]. For example, heritability estimates in the presence of admixture can be biased and inflated [44]. Alternatively, higher heritability estimates may be driven by higher relatedness among geographically proximal individuals. The presence of structure can create challenges disentangling the heritable component due to genetics versus similar environments [44]. Other methods for inferring heritability (e.g. LD score regression) are suboptimal in the presence of admixture, as LD from these populations are often not reflective of the study cohort and vary locally [45,46]. Other methods for inferring genetic architecture, including Bayesian linear mixed models (LMMs) such as the Bayesian sparse LMM (BBSLM), Bayes R, and BOLT-LMM, have been shown to be effective at controlling population stratification, cryptic relatedness, and also increase power in structured populations [47*,48–50]. These studies demonstrate that more advanced GWAS methods may be more fruitful generally, but especially in Africa where higher rates of substructure are typical.

Successful GWAS strategies in African and African descent populations

Despite these challenges, many successful examples illuminate paths forward. Because of high prioritization of infectious disease studies, most positive examples exist for genetic susceptibility studies, including of tuberculosis [51–53], malaria [54], sickle cell disease modifiers [55], HIV [56–59], nontyphoidal Salmonella [60*], and trypanosomes [61]. Significant findings were aided by simpler genetic architectures and higher genetic risk divergence between endemic cases versus high-risk controls due to natural selection. Some challenges of studying these evolutionary important traits, however, are high levels of genetic diversity in the parasite and variable LD patterns among populations, sometimes necessitating specialized association approaches that allow for multiple independent origins of resistance loci and/or allelic heterogeneity [54]. Some anthropometric studies have faced similar challenges and advantages due to high divergence, natural selection, and genetic architectures, such as in skin pigmentation [62*,63*]. In smaller cohorts that are underpowered for discovering individual loci, gene-based associations can sometimes be useful in conjunction with functional follow up [64]. Studies of traits with elevated prevalence in African Americans, such as BMI, prostate cancer, and low birth weight [65–67] have analyzed genome-wide significant loci by local ancestry and/or more easily fine-mapped variants with narrower LD. Additionally, multiethnic studies including African Americans have demonstrated the utility of integrative genomics approaches for fine-mapping, e.g. with pulmonary function variants [68*]. Several recent GWAS reviewed here have used linear mixed models, with a random effect to account for genetic relatedness. These models are useful but can produce inflated heritability estimates, which can be corrected using a second random effect to measure spatial distance as a proxy for environmental effects [69*].

Advantages and opportunities for genetic architecture studies in Africa

The opportunities for large-scale genetic studies in Africa are ample. Growing inclusion of African Americans in medical genomics studies is crucial, but still leaves behind many populations and large swaths of sub-Saharan African genetic diversity, and these populations may greatly increase our understanding of complex trait genetic architecture [70]. There is more genetic and often phenotypic diversity in Africa that has been understudied, meaning there is considerable low-hanging fruit for novel findings and insights into the genetic architectures and etiologies of complex traits. More rapid LD decay in Africa also means there is greater fine-mapping resolution to pinpoint causal variants influencing traits than will be discovered in any other global population [71], as reviewed recently [72]. For example, several variants in TCF7L2 were associated with type 2 diabetes in European and East Asian populations in the early GWAS era, but candidate loci were narrowed considerably via comparison with more diverse West African cohorts, even with smaller cohort sizes [73].

Major opportunities also present themselves in precision medicine. For example, polygenic risk scores have been of growing interest as large-scale GWAS now offer low-
cost tests that can outpace the clinical status quo [74,75]. However, these scores generalize poorly across diverse populations [11]. European GWAS results consistently predict genetic risk several-fold less accurately in non-Europeans, performing the worst in African Americans (and by extension, likely even worse in eastern, central, and southern African populations) [76–79]. A typical but somewhat misguided argument in favor of immediate translational implementation of polygenic risk scores is that standard clinical lab tests from blood panels are often differentially informative across ethnicities and more reliable in European descent populations. However, interpretability gaps for current clinical tests are less acutely and consistently worse in non-European populations than genetic risk prediction; the underlying biology remains the same, such that for all diseases, drugs do not routinely work many-fold better in European than African-descent populations. Further, new population-specific interpretation of common clinical lab tests enables better prognostic value than existing reference intervals [80*]. In contrast, the most significant and highest frequency genetic variants from GWAS used to predict genetic risk are not likely to be the same across populations, even when the underlying causal variants are the same. This is due to GWAS discovery biases, as variants used to predict risk tend to explain more phenotypic variation in the study population. While improved analytical methods hold promise, the only way genetic prediction power of inherited diseases in non-Europeans can truly be made equal is with massive investments to produce similar-sized GWAS of these phenotypes in non-European populations. Additionally, discoveries based on African genetics contribute to global knowledge, but many African population groups are sufficiently different [37] that insights made from trans-ethnic studies can similarly be gained by analyzing multiple GWAS of different African populations.

As a major genetics mission is to understand the biological basis and evolutionary origins of diseases and traits and use this knowledge to perform biologically-informed drug discovery, human evolution tells us that Africa has a huge role to play. Progress so far has been slower due to a need for increased capacity and collaborative engagement with African investigators. Several outstanding examples of this potential already exist, such as the Southern African Human Genome Programme (SAHGP), one of the first genetic architecture studies of African participants fully funded and analyzed by Africans [81*]. International collaborations have also blazed the trail for meaningful collaborations with deep investments in building research capacity in human genomics, such as MalariaGEN, partnerships by the African Center of Excellence for Genomics of Infectious Diseases (ACEGID), as well as the Global Initiative for Neuropsychiatric Genetics Education in Research (GINGER) program. Calls from African researchers for funding and building research capacity in genetics [70,82] should be thoughtfully heeded to ensure that those with the greatest public health needs are not the last to benefit.

Conflict of interest statement
Nothing declared.

Acknowledgements
We thank Brenna Henn and Luis Quintana-Murci for inviting this review. We also thank Karestan Koenen, Lori Chibnik, and Anne Stevenson for their thoughtful comments. Research reported in this publication was supported by NIH K99MH117229. This research was partially funded by the South African government through the South African Medical Research Council. The content is solely the responsibility of the authors and does not necessarily reflect the official view of the South African Medical Research Council. This work was also supported by the National Research Foundation of South Africa.

Appendix A. Supplementary data
Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.gde.2018.08.005.

References and recommended reading
Papers of particular interest, published within the period of review, have been highlighted as

- of special interest


Provides a framework for describing ancestry in genetic studies, and describes the disproportionately high contribution of associations identified in African and Hispanic/Latin American populations.


Reviews NIH NHGRI efforts to prioritize diversity in human genomics research.

Genetics of human origins


In a GWAS of nontyphoidal Salmonella, a common and frequently fatal cause of bacteremia in children and HIV-infected adults in Africa, this study identified a significant susceptibility locus in STAT4 in Kenyan children with controls with replication in Malawians. This eQTL modulates interferon-γ production in stimulated natural killer cells, highlighting the shared genetic architecture of infectious and autoimmune disease.


Along with HBS1L, one of the two first studies of skin pigmentation in continental Africa, which showed that skin pigmentation is more polymorphic in Khoesan populations than described elsewhere previously.


Along with HBS1L, one of the two first studies of skin pigmentation in continental Africa, which showed functional evidence of a role for FMSD12 in pigmentation.


This GWAS of pulmonary function identified 60 novel genetic loci, a 50% increase in the number of known loci, with 8 identified in African ancestry and 34 from multi-ethnic meta-analyses. Notably, fine-mapping of loci was improved by incorporating LD, functional data, and/or multi-ethnic analysis.


This study showed that using standard linear mixed models with kinship as a random effect to estimate heritability can produce inflated estimates, but that including a second random effect, a covariance matrix approximating geographical distance, produces less inflated heritability estimates.


This preprint demonstrates the clinical value of defining reference intervals for blood panels not by the standard small (N=128) patient population with biased ancestry, but by very large patient samples from diverse ancestries (N>10,000). Serum creatinine and hemoglobin A1c in particular produce actionably different reference intervals across self-reported ethnicities.


A pilot study of deep whole genome sequencing of 24 southern African individuals exhibiting complex, region-specific population structure. This study was led and funded entirely by African investigators and an African government, respectively.


