HHS Public Access

Author manuscript

Trends Genet. Author manuscript; available in PMC 2022 February 01.

Published in final edited form as:

Trends Genet. 2022 February ; 38(2): 152–168. doi:10.1016/j.tig.2021.09.013.

Advances in integrative African genomics

Chao Zhang[1,3], **Matthew E.B. Hansen**[1,3], **Sarah A. Tishkoff**[1,2,*]

[1]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

[2]Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA

[3]These authors contributed equally to this work

Abstract

There has been a rapid increase in human genome sequencing in the past two decades, resulting in the identification of millions of previously unknown genetic variants. However, African populations are under-represented in sequencing efforts. Additional sequencing from diverse African populations and the construction of African-specific reference genomes is needed to better characterize the full spectrum of variation in humans. However, sequencing alone is insufficient to address the molecular and cellular mechanisms underlying variable phenotypes and disease risks. Determining functional consequences of genetic variation using multi-omics approaches is a fundamental post-genomic challenge. We discuss approaches to close the knowledge gaps about African genomic diversity and review advances in African integrative genomic studies and their implications for precision medicine.

African integrative genomics is important for precision medicine globally

Genetic variants that contribute to human disease risk often vary in frequency amongst global ethnic groups. Differences in the evolutionary history of populations can lead to population-level differences in the prevalence of common and rare genetic diseases (recently reviewed in [1]). From this perspective, precision or personalized medicine is fundamentally intertwined with evolutionary history. Fossil and genetic evidence indicate that Africa is the origin of modern humans, approximately 300 000 years ago [2]. As the cradle of humanity, Africa has more cultural, linguistic, and genetic diversity than any other continent [3,4]. Africa is an extraordinarily important region in human evolutionary and biomedical genomic studies due to this genetic and cultural diversity.

African populations have a high prevalence of both communicable and noncommunicable diseases. Numerous communicable diseases that are endemic to regions within Africa include malaria [5], tuberculosis, and HIV [6], as well as tropical diseases with pandemic potential, like hemorrhagic fevers (e.g., Ebola virus disease). There has also been a rise of

*Correspondence: tishkoff@pennmedicine.upenn.edu, (S.A. Tishkoff).

Declaration of interests

No interests are declared.

Resources

noncommunicable diseases such as type 2 diabetes and cardiovascular disease, possibly due to a rise in Westernized diets and sedentary lifestyles in many regions.

African populations live in diverse environments, such as tropical rainforests, deserts, savannahs, and mountainous regions, and traditionally practice a wide variety of subsistence patterns such as agriculture, pastoralism, and hunting-gathering. This diversity allows for studies of genetic adaptations to a variety of environments and dietary factors (Box 1), which may also have pleiotropic effects on disease susceptibility. Approximately 1/7th of the world's population lives in Africa, yet proportionally little **genomics** (see Glossary) research is conducted on individuals of African ancestry compared with people of other ancestries. This has resulted in healthcare inequities and will impede implementation of **precision medicine** in people of African ancestry (e.g., [7]) because **genome-wide association study (GWAS)** results and **polygenic risk scores (PRS)** often do not transfer well between genetic ancestries, particularly from non-African to African cohorts [8,9] (Box 2).

Africans have the youngest average age of any continent, which will continue to be the case for the next 50 years at least [10], and many African countries are rapidly developing economically. Consequently, the African biomedical research community is poised to expand in terms of the number of trained scientists, the biological research infrastructure, and the scale of international scientific organization within Africa. The return-on-investment of academic research dollars spent on building capacity for biological sciences in Africa is arguably higher than many other places.

African-based genetic studies have the potential to provide new opportunities to understand disease etiology relevant to African populations but also globally. Including more African populations in integrative studies enables 'fine-mapping' in GWAS to identify causal variants for **intermediate phenotypes**, human traits, and diseases because **linkage disequilibrium (LD)** is lower in Africans than non-Africans. Additionally, higher genetic diversity and adaptation to diverse environments in Africans may lead to more genetic 'perturbations' and more phenotypic variation (e.g., gene expression and epigenetic modification), which would improve the power for association studies. Findings based on African ancestries would, therefore, benefit precision medicine in global human populations.

Currently, there is a substantial disparity in the representation of African ancestries in **integrative genomic** studies [11]. This impedes health equity as the transferability of findings based on European ancestries to other populations can be low by various metrics (Box 2). Integrative genomics research, which integrates genomic and other 'omics' data, seeks to identify the heritable molecular and latent intermediate biological causes underpinning higher-level phenotypes. In this review we first give a brief history of the advances in African genomic sequencing over the past 20 years and what gaps remain. We then highlight the current progress in African integrative genomics and some of the implications for human health and disease.

African genomics in 2021

Most human genomic studies have focused on European populations. It is estimated that only 2% of genomic data are from individuals with African ancestries [11] and most of these were generated by genotyping array technologies [12]. African whole-genome sequencing (WGS) data is still very scarce.

The first high-coverage African WGS data (~40× coverage) was from a male individual from the Yoruba ethnic group in Nigeria generated in 2008 [13]. Two years later, complete genomes of one Khoesan individual (~10×) and one Bantu individual (~30×) from southern Africa were sequenced [14]. The Thousand Genomes Project (TGP) is a milestone population genetics dataset for global diversity [15] that now contains ~900 deep WGS (30×) from seven African ancestry populations [16]. Various lab-level efforts have examined patterns of genetic diversity from African WGS data, though most were limited in the number of samples and populations. Examples include studies of hunter-gatherers [17–22] and studies of populations from specific regions of Africa (e.g., South Africa [23]).

There have also been African genomics studies with a broader scope that aim to capture the genetic variation among major continental linguistic groups, ecosystems, and lifestyles [4,24–26]. For instance, Fan *et al.* [25] generated 92 WGS genomes and Lorente-Galdos *et al.* [26] generated 22 WGS genomes from African individuals from populations spanning a broad range of ethnic, cultural, and linguistic groups. Although these continental-scope studies included more diverse populations, they are still based on a relatively small number of samples. Cohorts covering more populations from diverse ancestries with larger sample sizes are necessary to capture the full extent of African genomic diversity. One such cohort is from the H3Africa Consortium (Box 3), which generated 425 WGS data comprising ~50 ethnolinguistic groups [27]. This cohort uncovered millions of previously undescribed variants from newly sampled populations [27]. Overall, WGS data have greatly enhanced our understanding of African populations' evolutionary history and local adaptations (Box 1).

African genomics beyond 2021

Including more WGS data for African populations will not be enough to catalogue African genomic diversity accurately. The main issue is that the majority (70%) of the sequence in the human reference genome (HRG) was obtained from a single individual, ignoring global diversity and, thus, can introduce biases when used as a reference genome. The HRG bias makes it difficult to map short reads generated by next-generation sequencing (NGS) unambiguously to the HRG. For instance, reads with non-reference alleles may not be mapped or mapped at a low rate (mapping bias), therefore inducing particular errors in the variant calling process (variant calling bias) [28].

Unmapped short reads generated by NGS technologies are usually ignored in downstream analyses. For example, the global Simons Genome Diversity Project WGS cohort uncovered 5.8 Mb of sequence contigs that do not map to the HRG [24]. Reads may not map for many reasons: the reference does not contain the sequence variant (HRG bias), the

read maps to multiple locations and/or to complex regions of the genome, or because of non-human contamination (e.g., from pathogens). We currently know little about these unmapped reads, which could contain functionally important elements of the human genome. Recently, researchers have used unmapped reads to construct **pan-genomes** for some human populations. For instance, Sherman *et al.* assembled a pan-genome from deep sequencing of 910 people of African descent [29] and Duan *et al.* assembled a pan-genome for Han Chinese with 275 samples [30]. Sherman *et al.* assembled unmapped reads into contiguous sequences (contigs) as nonreference segments [29], while Duan *et al.* performed *de novo* assembly of all sequenced reads to contigs [30]. However, it is not clear which strategy will prove most effective for constructing pan-genomes using unmapped reads.

Additionally, **structural variations (SVs)** cannot be easily captured based on short reads. It is estimated that each human genome contains >20 000 SVs, many of which are located in regions that are unmappable using short-read sequencing approaches [31]. Many emerging technologies seek to go beyond short reads to resolve the full spectrum of variation. For example, single-molecule strategies such as long-read sequencing by Pacific Biosciences and Oxford Nanopore Technologies generate contiguous reads, tens to hundreds of kilobases long, enabling direct detection of SVs and improving the alignment of unique reads in repetitive regions [31].

The current HRG is entirely linear and, as described earlier, no single sequence can accommodate the unambiguous mapping of global human genetic variation. **Graph genomes** can accommodate complex individual genetic variation by representing individual genomes as paths through a graph. The graph nodes represent genomic positions where a known variant starts or ends, while the edges represent the sequence of that particular variant (Figure 1). As new genetic variants are discovered, they can be added to the reference graph. By including all known variation, short reads in NGS data can map effectively with less bias [32]. Using a graph reference improves short-read mapping sensitivity and produces a 0.5% increase in variant calling recall, including small insertion/deletion variation [33] and larger SVs [33,34]. Since graph genomes cover more genetic diversity, they are a natural framework for constructing African-specific references that would substantially reduce the current reference biases.

Future efforts in genomic studies should be focused on African ancestries to examine the ‘missing diversity’ that we have never captured before (see Outstanding questions). Closing the gaps in African genomics diversity, particularly for ‘missing’ functional variation, will require, in our view, six broad research efforts: (i) including larger sample sizes of individuals with African ancestry in genomic studies; (ii) sequencing more high-depth African genomes; (iii) generation of African reference genomes from ethnically diverse populations and construction of pan-genomes that include currently unmapped reads and structural variants, using long-read sequencing and graph-based methods (Figure 1); (iv) developing online tools and computational resources for comparing continental datasets; (v) increasing training and capacity for genomics research in Africa; and (vi) detailed characterization of phenotypic diversity (including electronic health records), which can be integrated with the genomic data as described in the next sections.

Recent progress in African integrative genomics

Genomic sequences alone are insufficient to understand the functional implications of variation, since they do not directly link genotypes to phenotypes. Although GWAS have uncovered many variants associated with human traits, GWAS usually provides insufficient information about the underlying biological mechanisms influencing variable traits. Integrative genomics approaches aim to fill in this information by combining genomic data with intermediate phenotypes that link genetic variation to phenotypes (Figure 2).

Shortly after the release of the first HRG, several large-scale integrative genomic projects were initiated to annotate the functional regions of the genome. For example, the encyclopedia of DNA elements (ENCODE) project, launched in 2007 by the National Human Genome Research Institute, aims to build a comprehensive list of functional elements in the human genome, including elements that regulate protein and RNA levels [35]. The Genotype-Tissue Expression (GTEx) project, started in 2010 and funded by the National Institutes of Health (NIH), aims to build a comprehensive public resource to study tissue-specific gene expression and variants associated with gene regulation [36]. Other larger-scale integrative omics projects include the Blueprint epigenome project and the Human Protein Atlas project [37,38]. Additionally, the past 5 years have seen an increase in various regional and national medical biobanks (e.g., the UK Biobank [39], the Human Phenome Consortium of China [40], the Million Veteran Project [41], and BioBank Japan [42]) with access to both genotypes and phenotypes for large numbers of samples, including electronic health records, which enables the mapping of hundreds of genetic associations with complex traits and disease. In the following years, more countries will build their own medical biobanks to pursue the goal of improving precision medicine. However, African integrative genomics research is still in its infancy, with few large-scale projects based in African countries (Box 3). In the following subsections, we review African integrative genomics research and the implications for health and disease.

African transcriptomic studies

Transcriptomics is the study of the complete set of RNA transcripts (transcriptomes) produced by the genome under specific conditions or in a specific cell type using high-throughput methods, such as RNA-seq analysis (Box 4). Transcriptomics studies in Africa have been focused on lymphoblastoid cell lines (LCLs). The LCLs of ~450 individuals from the TGP [43] (reviewed by Kelly *et al.* [44]) is a benchmark dataset showing that more distantly related populations tend to have a greater number of differentially expressed (DE) genes [43]. A potential problem with using LCLs is that they do not reflect gene expression '*in situ*', which is influenced by both genetics and environment. A few recent (since 2016) transcriptomic studies have focused on gene expression from whole blood sampled from African populations. By investigating transcriptional responses to bacterial and viral stimuli in whole blood, studies showed individuals with European and African ancestry in the US and Europe have genetic differences correlated with differential immune responses [45,46]. We discuss two recent transcriptomic studies from whole blood in indigenous African populations [47,48].

Transcriptomics of malarial infection—Malaria is a mosquito-borne disease caused by a parasite, *Plasmodium falciparum*, endemic to tropical and subtropical regions in Africa. One particular ethnic group who live throughout western and Central Africa, the Fulani pastoralists, are relatively better protected from severe outcomes from *Plasmodium* infection than other neighboring populations [49]. Using RNA-seq analysis of monocytes from both the Fulani pastoralists and a neighboring population, the Mossi, Quin *et al.* [47] identified many DE genes between these populations. The most significant DE gene was *P2RX7*, which plays an important role in innate immunity [50]. *P2RX7* has higher expression in the Fulani than that in the Mossi. Quin *et al.* suggested that the Fulani have higher baseline levels of expression of inflammasome pathway components, resulting in stronger inflammasome activation following *P. falciparum* infection [47]. However, they did not use genome sequences in the analysis and, therefore, underlying variants that play a role in differential immune response in the Fulani are still unknown. Additionally, analyses of signatures of natural selection, particularly those that overlap regulatory variants in the Fulani, will shed light on the mechanisms that confer the lower susceptibility of Fulani to *P. falciparum* malaria.

Transcriptomic differentiation between hunter-gatherers and agriculturalists—The shift from a hunter-gatherer to an agricultural mode of subsistence has been associated with changes in infectious disease burdens [51]. To investigate the evolution of genetic variation influencing the immune system in hunter-gather and agriculturalist populations, Harrison *et al.* [48] analyzed the difference in transcriptomic changes induced by immune responses to bacterial and viral stimuli between the Batwa, a rainforest hunter-gatherer population from southwest Uganda and their Bantu-speaking agriculturalist neighbors, the Bakiga. They identified significant immune response differences in transcriptional profiles between the Batwa and Bakiga [48]. Genes with higher expression levels in Batwa individuals were enriched in pathways involved in immune responses to viruses, while genes with higher expression levels in Bakiga individuals were enriched for inflammatory response genes. The increased divergence between hunter-gatherers and agriculturalists in the early transcriptional response to viruses compared with that for bacterial stimuli, suggests that differences in viral exposure may have been an important factor contributing to the immune response divergence between the Batwa and the Bakiga. This study demonstrated that positive natural selection has contributed to present-day differences in innate immune responses between the Batwa and the Bakiga.

African epigenomic studies

Epigenomics is the study of the complete set of chemical modifications of genetic material that impact gene expression. Methylation of CpG dinucleotides, histone modification, and chromatin accessibility are important contributors to epigenetic regulation of gene expression in numerous cellular processes. Genetic variation, environmental factors, and other factors (e.g., age [52,53]) impact these epigenetic processes, which, in turn, impacts phenotypic variation [54]. For example, maternal nutritional status during early pregnancy causes persistent and systemic epigenetic changes [55], and seasonal fluctuations in nutritional status affects DNA methylation at human metastable epialleles [56]. Recent

research has shown that both prenatal environmental factors and genotype, together, contribute to DNA methylation [57].

Only a few studies have investigated epigenetic variation among ethnically diverse populations and showed that DNA methylation differs between populations with diverse ancestries [58–62]. For example, Fraser *et al.* examined DNA methylation in cell lines derived from one African and one European population and found population-specific DNA methylation patterns at over a third of all genes [58]. By mapping quantitative trait loci for chromatin accessibility (**caQTLs**) from ten diverse populations, Tehranchi *et al.* [63] observed a clear trend for increased sharing within continents: the mean fraction of shared caQTLs was 59.9% within Africans and 59.8% within Europeans, compared with 48.4% between these two groups.

Since 2015, there have been several population-based epigenomic analyses focused on indigenous African populations [53,60], not including studies of international reference panels containing African ancestry individuals. Fagny *et al.* generated genotype and DNA methylation profiles for 362 rainforest hunter-gatherers and farmers [60]. They found that the current habitat and ancient lifestyle of a population both have critical impacts on the methylome. Methylation variation associated with recent changes in habitat (determined by comparing forest farmers with farmers in an urban area) mostly involves immune and cellular functions, whereas that associated with ancient lifestyle (determined by comparing forest dwelling hunter-gatherers with forest dwelling farmers) affects developmental processes. Gopalan *et al.* investigated genome-wide methylation patterns using saliva- and whole-blood-derived DNA from two hunter-gather African populations: the Baka of the western Central African rainforest and the ‡Khomani San of the South African Kalahari Desert [53]. Hundreds of CpG sites with methylation levels significantly associated with age were identified, including the age-associated site in the promoter of the gene *ELOVL2*, which has been well studied [64], and 277 age-associated sites that have not been reported in previous studies. Studies like this exemplify the need for more African population-based studies to identify additional ancestry-specific epigenetic variation that may contribute to phenotypic differences.

African proteomic studies

Proteomics is the study of the complete set of proteins (proteomes) produced by the genome under specific conditions or in a specific cell using high-throughput methods such as mass spectrometry (MS) and antibody-based methods (Box 4). Current studies have examined proteomic data from different cell types [65], tissues [38,66], or disease contexts [67–69] that have provided an atlas of the human proteome in different biological contexts (see review [70]). Only a few studies investigated proteomic diversity amongst ethnically diverse populations, including Africans. The most comprehensive study comparing proteomic differences among populations was from Wu *et al.* [71]. They determined relative protein levels of 5953 genes in LCLs from 95 diverse individuals, including 33 YRI genotyped in the HapMap Project. They identified 247 proteins that had significantly different levels between CEU and YRI at a false discovery rate (FDR) of 10%. One such DE gene is Parathymosin (*PTMS*), which may confer resistance to certain opportunistic infections.

By characterizing the most and least variable proteins, they found that the most variable proteins were enriched for playing a role in immune response, whereas the least variable proteins were enriched for playing a role in housekeeping processes. Given the considerable infectious disease load in indigenous African populations, it will be important to include these populations in proteomics studies in the future to understand differences in immune response among populations.

African studies of integrative metabolomics

Metabolomics is the study of metabolites found in the body, broadly defined as the molecular substrates and products of cellular functions [72]. Exogenous compounds like food and medicine contribute to the metabolome, as they are also metabolized by the body and can impact a person's physiological state. The full metabolome, which most likely exceeds 1 million metabolites at any given time, is a snapshot of the chemical state of an individual at a single point in time [73]. Most current methods for measuring the metabolome are based on MS (Box 4). Metabolite variation tends to be moderately heritable, with SNP-heritabilities of ~10–50% commonly observed in European GWAS [74,75]) (see review [76]).

Genetic variation that impacts metabolism may be variable among ethnically diverse populations, which can lead to large inter-ethnic differences in metabolites and downstream disease risks. Most African ancestry cohorts used in integrative metabolomics studies are African-Americans (e.g., [77–85]). There have been few discovery-stage metabolite GWAS studies in African population cohorts [86]. Several online resources track this information, based on manual curation of the literature [87] (http://www.metabolomix.com/list-of-all-published-gwas-with-metabolomics; http://mips.helmholtz-muenchen.de/proj/GWAS/gwas) and automated querying of the EMBL-EBI GWAS catalogue [88]. For example, the GWAS Discovery Monitor [88], a GWAS catalog database tool, reports only one discovery-stage GWAS study based on WGS or African WGS-imputed genotype data with more than 10 000 Africans that includes metabolomic measurement [81,82,89,90]. This lack of discovery GWAS cohorts from African populations is a problem because GWAS results do not necessarily transfer well between genetically differentiated populations (Box 2). The transferability of lipid GWAS from Europeans to Asian and African populations has been examined recently [91]. They find that roughly 3/4 of the highly significant European GWAS loci show evidence of replication in a Ugandan cohort. The lack of replication for roughly 1/4 of the lipid loci shows that European GWAS do not completely predict African lipid genetic associations. This study also sets the baseline expectation that African discovery cohorts of equal size would uncover at least as many African-specific associated loci, and probably many more such loci, due to the higher levels of genetic and phenotypic diversity within Africa and lower levels of LD.

Microbiome sequencing and integration with host properties

The human **microbiome** comprises all microscopic organisms that live in or on the human body. The microbiome is a complex web of living material that can be highly adapted to living within specific human body site niches. These microbes can be long-term or transitory residents, they can be in a mutually beneficial relationship with the host, benign to the

host, or pathogenic to the host, and their roles may change in time as conditions change. In practice, the microbiome is often treated like an intermediate phenotype of the host and diseases can sometimes be linked to dysbiosis of the microbiome (i.e., deviations from the usual range found in 'healthy' hosts). Modern methods of profiling the microbiome are based on sequencing (Box 4). As with other omics research, most studies of the microbiome in African ancestry cohorts are based on African-Americans [84,92–96].

Disease-motivated microbiome studies in African populations—There have been a handful of recent microbiome studies in African populations that focus on links between regional endemic disease susceptibilities or progression with microbiome composition. The scope for most of these studies is to establish whether or not any associations exist between microbiome composition, or the abundance of specific taxa, and disease status. In most cases, future work is required to establish whether the observed correlations are a (partial) cause or consequence of the disease. For example, bacterial vaginosis (BV) is a vaginal microbiome state that is characterized by a shift from a *Lactobacillus*-dominated composition to a more diversified composition and has been linked to increased susceptibility to HIV acquisition and various reproductive health issues. BV is common globally but is most prevalent in women of sub-Saharan African descent [97]. Recent work has confirmed the prevalence of BV, which is correlated with increased proinflammatory cytokines, in different African populations [96–101]. Most genetic associations with BV are enriched for pathways of the innate immune system, including Toll-like receptors and cytokine regulation [99]. Future studies with more participants will be needed to pin down specific host genetic factors that may regulate the vaginal microbiome. Other recent work includes studies of the nasopharyngeal microbiome in children with pneumonia infection [102], studies of correlations between the gut microbiome (GM) and helminth infection [103], as well as type 2 diabetes [104]. Related recent work has focused on establishing a better baseline for normal GM compositions among healthy adults from West Africa (Ghana) [105].

Lifestyle- and diet-motivated microbiome studies in African populations—The connection between diet and GM has been a longstanding research focus due to the ramifications for nutritional health, socio-economic disparities, and human coevolutionary history with microbiota [106]. Comparisons of GMs between global populations with diverse dietary modes is one avenue to probe the impact of diet [107]. A consistent finding from global diversity studies is that the GM becomes less diverse in more urban and industrialized populations compared with many populations practicing traditional subsistence practices [108–114]. The bacterial diversity largely coincides with the abundance of either *Prevotella* (higher diversity GM) or *Bacteroides* (lower diversity GM). The reason for this discrepancy is not fully understood but is possibly due to the amount of fiber in the diet: evidence suggests that *Prevotella*-dominant GMs catabolize plant fibers more efficiently than *Bacteroides*-dominant GMs (see review [115]). Among populations practicing traditional lifestyles in sub-Saharan Africa, the GM of hunter-gatherer populations tend to be the most diverse, while settled agriculturalists have the least diverse GMs [109,113]. Indeed, the GMs of some rural agriculturalists from Botswana are nearly indistinguishable from the GMs of Western populations [113]. The impact of

genetics, diet, or other aspects of the environment on the GM in indigenous populations is still not clear.

African GWAS of variable traits and disease risk

Similar to other omics studies, GWAS for human physiological traits and clinical disease risks are under-represented in African populations. As of July 2021, no more than 15% of studies in the GWAS catalogue[i] are based on individuals with African ancestries. In the GWAS including African ancestries, only ~20% of them included indigenous African populations, of which ~60% were launched in the last 5 years (since 2017). One of the largest African GWAS to date is from the Ugandan Genome Resource (UGR) project, based on the Ugandan General Population Cohort and associated biobank [90] (Box 3). The UGR has analyzed genetic associations with many clinical phenotypes, including height, body mass index, circulating lipids (low- and high-density lipoprotein cholesterol), and white blood cell counts. Other phenotypes in GWAS involving African ancestries include malaria susceptibility [116–119], cholesterol levels [89,90], height [120], type I/II diabetes [121,122], and skin pigmentation [123–126]. Here, we highlight recent African GWAS studies on one of the best characterized human traits, skin pigmentation.

GWAS of skin pigmentation in African populations—Skin pigmentation is a highly heritable human trait and one of the most strikingly variable phenotypes among human populations [127]. Typically, darker skin pigmentation is observed closer to the equator and lighter pigmentation observed at high latitudes. Skin pigmentation also varies widely across Africa. Khoesan hunter-gatherers and pastoralists in and near the Kalahari Desert exhibit light skin pigmentation, while Nilo-Saharan-speaking populations from East Africa have some of the darkest pigmented skin among humans. The global variation in skin pigmentation has been shaped by natural selection, migration, and admixture (reviewed by Feng *et al.* [128]). Studies have identified dozens of genes (e.g., *SLC24A5*, *SLC45A2*, *MC1R*, *TYR*, *TYRP1*, and *OCA2*) associated with skin color differences in humans. Most of these GWAS have primarily focused on Eurasian and admixed African-American populations [126,127,129–131]. There have been only a few GWAS of skin pigmentation that include indigenous African populations [123–125].

By genotyping 1570 samples from Ethiopia, Tanzania, and Botswana, Crawford *et al.* [123] identified eight loci at four regions of the genome (i.e., *SLC24A5*, *MFSD12*, *DDB1/TMEM138*, and *OCA2/HERC2*) that are associated with skin pigmentation. All but one of these associations are novel loci and one of the genes (*MFSD12*) had never previously been characterized and was unknown to impact pigmentation. This study showed that both dark and light pigmentation alleles arose before the origin of modern humans and that both light and dark pigmented skin has continued to evolve throughout hominid history. For example, alleles near *DDB1* associated with light pigmentation swept to near fixation outside of Africa due to positive selection and they showed that these lineages coalesce ~60 ka, corresponding with the time of migration of modern humans out of Africa. They found that variants associated with dark pigmentation in Africans are identical by descent in South

[i] www.ebi.ac.uk/gwas/

Asian and Australo-Melanesian populations. Another study by Martin *et al.* [124] examined genetic association with pigmentation in the Nama and ‡Khomani San populations. They identified canonical and noncanonical skin pigmentation loci, including near *SLC24A5*, *TYRP1*, *SMARCA2/VLDLR*, and *SNX13*.

Both Martin *et al.* and Crawford *et al.* [123,124] identified novel genetic loci that are associated with skin pigmentation, suggesting the value of including more ethnically diverse African populations in GWAS for human traits. More variable loci in African populations, together with high levels of phenotypic variation, improves the power of GWAS; it is difficult to identify genetic associations in populations where the causal allele is at low frequency or fixed (meaning all individuals carry the same allele) due to demographic history or evolutionary pressures.

Concluding remarks

The lack of representation of ethnically diverse African ancestries in genomic studies leads to healthcare inequities in Africans, as integrative genomics analysis based on non-African populations do not necessarily transfer well (e.g., replication rates <95%) to African populations (Box 2). Current African integrative genomics research is limited in several respects. One is that many studies are based on international reference panels that only include a small number of populations with African ancestry. For example, four of the five continental African populations in the TGP are from West Africa, providing little coverage of northern, eastern, and southern African ancestries. Second, current integrative studies are often based on LCLs or whole blood, which is not the appropriate cell or tissue type for many phenotypes and diseases. Moreover, environmental factors were not considered in many studies, limiting the investigation of gene–environment interaction on phenotypes. Lastly, many studies were based on small sample sizes with limited amounts of integrative genomics data generation.

There are several challenges to conducting integrative genomics research in many parts of Africa. These studies involve different types of large biological datasets and rely on large numbers of specialized experimental and computational analyses, which requires a considerable investment of time and money sustained over a decade-long time scale. Additionally, ethical concerns such as privacy, data ownership and sharing, community and individual consent, and local cultural concerns require careful consideration [132].

In order to meet these challenges, international collaborations built on a common guideline and a comprehensive ethics and governance framework will be key (see Box 3 for large-scale genomics consortiums that include African ancestry cohorts). Partnerships with African scientists and scientific institutions helps with cost sharing and provides a vital layer of ethical oversight for international research. From the perspective of non-African scientific bodies with a stake in African omics research, greater funding emphasis on training African genomic research scientists on data transfer, storage, integration, and analysis would allow the genomic data research to be conducted locally. The development of online tools and computational resources that make omics data and analyses accessible in Africa should be an area of emphasis [133]. Efforts to improve local African research capabilities should also

be a priority, as this allows for greater and more fruitful international partnerships with local research scientists [134]. Despite the challenges, African integrative omics research could greatly benefit African healthcare as well as advance precision medicine for all people.

Acknowledgments

CZ, MEBH and SAT are supported by NIH 1R35GM134957, R01AR076241, and ADA 1-19-VSN-02

Glossary

caQTLs
chromatin accessibility quantitative trait loci (caQTLs) are genomic loci that are associated with chromatin accessibility

Genetic architecture
the characteristics of genetic variation that influences heritable phenotypic variability. The genetic architecture of a trait depends on the number of genetic variants affecting a trait, their frequencies in the population, the magnitude of their effects, and their interactions with each other and the environment

Genomics
the study of whole genomes of organisms. Genomics uses a combination of recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze the structure and function of genomes

Graph genome
a nonlinear representation of a genome, in which paths in the graph represent genetic variants from diverse individuals

Genome-wide association studies (GWAS)
a statistical association between genetic variation and measured phenotypes in an effort to identify functional variation

Integrative genomics
identification of the heritable molecular and latent intermediate biological causes underpinning higher-level phenotypes. Methodologically, omics research typically focuses on measuring large numbers of molecular traits at once using high-throughput techniques and then associating this high-dimensional data either with the higher-level phenotypes of interest or with underlying genetic variation

Intermediate phenotype
a quantitative biological trait that is reliable and reasonably heritable. In integrative genomics, intermediate phenotypes are those biological traits, such as the transcriptome, proteome, and metabolome, linking genomes and phenotypes or diseases

Linkage disequilibrium (LD)
the nonrandom association of alleles at different loci in a given population

Metabolomics
the study of metabolites found in the body, broadly defined as the molecular substrates and products of cellular functions

Microbiome
the human microbiome comprises all microscopic (smaller than ~1 micron) organisms that live in or on the human body

Pan-genome
a collection of multiple genomes from a single species. It emphasizes the completeness and diversity of genomes in a species

Proteomics
the study of the complete set of proteins (proteomes) that are produced by the genome under specific conditions or in a specific cell using high-throughput methods such as mass spectrometry and antibody-based methods

Precision medicine
also called personalized medicine, this is an approach for protecting health and treating disease, taking into account a person's genes, behaviors, and environment

Structural variation (SV)
large-scale structural differences in the genomic DNA (1 kb and larger in size) that are inherited and polymorphic in a species. They are a result of chromosomal rearrangement: deletion, insertion, duplication, inversion, and translocations

Polygenic risk score
an effect-weighted sum of the number of risk alleles an individual carries, calculated according to their genotypes and the allelic effects estimated from relevant GWAS summary statistics: $\sum_{\ell} G_{i\ell}\beta_{\ell}$ for individual i with genotypes $G_{i\ell}$ at variant ℓ with effects β_{ℓ}

Transcriptomics
the study of the complete set of RNA transcripts (transcriptomes) that are produced by the genome under specific conditions or in a specific cell type using high-throughput methods, such as microarray analysis and RNA-seq analysis.

References

1. Pereira L et al. (2021) African genetic diversity and adaptation inform a precision medicine agenda. Nat. Rev. Genet 22, 284–306 [PubMed: 33432191]
2. Hublin J-J et al. (2017) New fossils from Jebel Irhoud, Morocco and the pan-African origin of *Homo sapiens*. Nature 546, 289–292 [PubMed: 28593953]
3. Tishkoff SA et al. (2009) The genetic structure and history of Africans and African Americans. Science 324, 1035–1044 [PubMed: 19407144]
4. Gurdasani D et al. (2015) The African Genome Variation Project shapes medical genetics in Africa. Nature 517, 327–332 [PubMed: 25470054]
5. Snow RW et al. (2017) The prevalence of *Plasmodium falciparum* in sub-Saharan Africa since 1900. Nature 550, 515–518 [PubMed: 29019978]

6. Dwyer-Lindgren L et al. (2019) Mapping HIV prevalence in sub-Saharan Africa between 2000 and 2017. Nature 570, 189–193 [PubMed: 31092927]
7. Wilson H (2014) Pharmacogenomics failing to reach developing countries. Pharmacogenomics 15, 731–732 [PubMed: 24897280]
8. Martin AR et al. (2017) Human demographic history impacts genetic risk prediction across diverse populations. Am. J. Hum. Genet 100, 635–649 [PubMed: 28366442]
9. Duncan L et al. (2019) Analysis of polygenic risk score usage and performance in diverse human populations. Nat. Commun 10, 3328 [PubMed: 31346163]
10. United Nations (2019) World Population Prospects 2019: Highlights, United Nations Publications
11. Sirugo G et al. (2019) The missing diversity in human genetic studies. Cell 177, 1080 [PubMed: 31051100]
12. Zhang C et al. (2019) PGG.SNV: understanding the evolutionary and medical implications of human single nucleotide variations in diverse populations. Genome Biol. 20, 215 [PubMed: 31640808]
13. Bentley DR et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53–59 [PubMed: 18987734]
14. Schuster SC et al. (2010) Complete Khoisan and Bantu genomes from southern Africa. Nature 463, 943–947 [PubMed: 20164927]
15. 1000 Genomes Project Consortium et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073 [PubMed: 20981092]
16. Byrska-Bishop M et al. (2021) High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. bioRxiv Published online February 7, 2021. 10.1101/2021.02.06.430068
17. Hsieh P et al. (2016) Whole-genome sequence analyses of Western Central African Pygmy hunter-gatherers reveal a complex demographic history and identify candidate genes under positive natural selection. Genome Res. 26, 279–290 [PubMed: 26888263]
18. Lachance J et al. (2012) Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. Cell 150, 457–469 [PubMed: 22840920]
19. Schlebusch CM et al. (2020) Khoe-San genomes reveal unique variation and confirm the deepest population divergence in *Homo sapiens*. Mol. Biol. Evol 37, 2944–2954 [PubMed: 32697301]
20. Lopez M et al. (2019) Genomic evidence for local adaptation of hunter-gatherers to the African rainforest. Curr. Biol 29, 2926–2935 [PubMed: 31402299]
21. Retshabile G et al. (2018) Whole-exome sequencing reveals uncaptured variation and distinct ancestry in the Southern African population of Botswana. Am. J. Hum. Genet 102, 731–743 [PubMed: 29706352]
22. Auer PL et al. (2012) Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. Am. J. Hum. Genet 91, 794–808 [PubMed: 23103231]
23. Choudhury A et al. (2017) Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. Nat. Commun 8, 2062 [PubMed: 29233967]
24. Mallick S et al. (2016) The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 538, 201–206 [PubMed: 27654912]
25. Fan S et al. (2019) African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. Genome Biol. 20, 82 [PubMed: 31023338]
26. Lorente-Galdos B et al. (2019) Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-Saharan populations. Genome Biol. 20, 77 [PubMed: 31023378]
27. Choudhury A et al. (2020) High-depth African genomes inform human migration and health. Nature 586, 741–748 [PubMed: 33116287]
28. Ballouz S et al. (2019) Is it time to change the reference genome? Genome Biol. 20, 1–9 [PubMed: 30606230]
29. Sherman RM et al. (2019) Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat. Genet 51, 30–35 [PubMed: 30455414]

30. Duan Z et al. (2019) HUPAN: a pan-genome analysis pipeline for human genomes. Genome Biol. 20, 149 [PubMed: 31366358]
31. Ho SS et al. (2020) Structural variation in the sequencing era. Nat. Rev. Genet 21, 171–189 [PubMed: 31729472]
32. Garrison E et al. (2018) Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat. Biotechnol 36, 875–879 [PubMed: 30125266]
33. Rakocevic G et al. (2019) Fast and accurate genomic analyses using genome graphs. Nat. Genet 51, 354–362 [PubMed: 30643257]
34. Eggertsson HP et al. (2019) GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. Nat. Commun 10, 5402 [PubMed: 31776332]
35. ENCODE Project Consortium et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447, 799–816 [PubMed: 17571346]
36. GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. Nat. Genet 45, 580–585 [PubMed: 23715323]
37. Astle WJ et al. (2016) The allelic landscape of human blood cell trait variation and links to common complex disease. Cell 167, 1415–1429 [PubMed: 27863252]
38. Uhlén M et al. (2015) Proteomics. Tissue-based map of the human proteome. Science 347, 1260419 [PubMed: 25613900]
39. Bycroft C et al. (2018) The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203–209 [PubMed: 30305743]
40. Jiang Y et al. (2021) Lifestyle, multi-omics features, and pre-clinical dementia among Chinese: the Taizhou Imaging study. Alzheimers Dement. 17, 18–28 [PubMed: 32776666]
41. Gaziano JM et al. (2016) Million Veteran Program: a mega-biobank to study genetic influences on health and disease. J. Clin. Epidemiol 70, 214–223 [PubMed: 26441289]
42. Nagai A et al. (2017) Overview of the BioBank Japan Project: study design and profile. J. Epidemiol 27, S2–S8 [PubMed: 28189464]
43. Lappalainen T et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506–511 [PubMed: 24037378]
44. Kelly DE et al. (2017) Global variation in gene expression and the value of diverse sampling. Curr. Opin. Syst. Biol 1, 102–108 [PubMed: 28596996]
45. Quach H et al. (2016) Genetic adaptation and Neandertal admixture shaped the immune system of human populations. Cell 167, 643–656 [PubMed: 27768888]
46. Nédélec Y et al. (2016) Genetic ancestry and natural selection drive population differences in immune responses to pathogens. Cell 167, 657–669 [PubMed: 27768889]
47. Quin JE et al. (2017) Major transcriptional changes observed in the Fulani, an ethnic group less susceptible to malaria. Elife 6, e29156 [PubMed: 28923166]
48. Harrison GF et al. (2019) Natural selection contributed to immunological differences between hunter-gatherers and agriculturalists. Nat. Ecol. Evol 3, 1253–1264 [PubMed: 31358949]
49. Modiano D et al. (1996) Different response to *Plasmodium falciparum* malaria in west African sympatric ethnic groups. Proc. Natl. Acad. Sci. U. S. A 93, 13206–13211 [PubMed: 8917569]
50. Wiley JS et al. (2011) The human P2X7 receptor and its role in innate immunity. Tissue Antigens 78, 321–332 [PubMed: 21988719]
51. Wolfe ND et al. (2007) Origins of major human infectious diseases. Nature 447, 279–283 [PubMed: 17507975]
52. Bell JT et al. (2012) Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. PLoS Genet. 8, e1002629 [PubMed: 22532803]
53. Gopalan S et al. (2017) Trends in DNA methylation with age replicate across diverse human populations. Genetics 206, 1659–1674 [PubMed: 28533441]
54. Kader F and Ghai M (2017) DNA methylation-based variation between human populations. Mol. Gen. Genomics 292, 5–35
55. Dominguez-Salas P et al. (2014) Maternal nutrition at conception modulates DNA methylation of human metastable epialleles. Nat. Commun 5, 3746 [PubMed: 24781383]

56. Waterland RA et al. (2010) Season of conception in rural Gambia affects DNA methylation at putative human metastable epialleles. PLoS Genet. 6, e1001252 [PubMed: 21203497]
57. Czamara D et al. (2019) Integrated analysis of environmental and genetic influences on cord blood DNA methylation in new-borns. Nat. Commun 10, 2548 [PubMed: 31186427]
58. Fraser HB et al. (2012) Population-specificity of human DNA methylation. Genome Biol. 13, R8 [PubMed: 22322129]
59. Heyn H et al. (2013) DNA methylation contributes to natural human variation. Genome Res. 23, 1363–1372 [PubMed: 23908385]
60. Fagny M et al. (2015) The epigenomic landscape of African rainforest hunter-gatherers and farmers. Nat. Commun 6, 10047 [PubMed: 26616214]
61. Galanter JM et al. (2017) Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures. Elife 6, e20532 [PubMed: 28044981]
62. Natri HM et al. (2020) Genome-wide DNA methylation and gene expression patterns reflect genetic ancestry and environmental differences across the Indonesian archipelago. PLoS Genet. 16, e1008749 [PubMed: 32453742]
63. Tehranchi A et al. (2019) Fine-mapping cis-regulatory variants in diverse human populations. eLife 8, e39595 [PubMed: 30650056]
64. Garagnani P et al. (2012) Methylation of ELOVL2 gene as a new epigenetic marker of age. Aging Cell 11, 1132–1134 [PubMed: 23061750]
65. Thul PJ et al. (2017) A subcellular map of the human proteome. Science 356, eaal3321 [PubMed: 28495876]
66. Robins C et al. (2021) Genetic control of the human brain proteome. Am. J. Hum. Genet 108, 400–410 [PubMed: 33571421]
67. Zhernakova DV et al. (2018) Individual variations in cardiovascular-disease-related protein levels are driven by genetics and gut microbiome. Nat. Genet 50, 1524–1532 [PubMed: 30250126]
68. Yao C et al. (2018) Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. Nat. Commun 9, 3268 [PubMed: 30111768]
69. Sun W et al. (2016) Common genetic polymorphisms influence blood biomarker measurements in COPD. PLoS Genet. 12, e1006011 [PubMed: 27532455]
70. Suhre K et al. (2021) Genetics meets proteomics: perspectives for large population-based studies. Nat. Rev. Genet 22, 19–37 [PubMed: 32860016]
71. Wu L et al. (2013) Variation and genetic control of protein abundance in humans. Nature 499, 79–82 [PubMed: 23676674]
72. Johnson CH et al. (2016) Metabolomics: beyond biomarkers and towards mechanisms. Nat. Rev. Mol. Cell Biol 17, 451–459 [PubMed: 26979502]
73. Uppal K et al. (2016) Computational metabolomics: a framework for the million metabolome. Chem. Res. Toxicol 29, 1956–1975 [PubMed: 27629808]
74. Kettunen J et al. (2012) Genome-wide association study identifies multiple loci influencing human serum metabolite levels. Nat. Genet 44, 269–276 [PubMed: 22286219]
75. Shin S-Y et al. (2014) An atlas of genetic influences on human blood metabolites. Nat. Genet 46, 543–550 [PubMed: 24816252]
76. Rhee EP (2020) Genetic influence on the metabolome. In Metabolomics for Biomedical Research (Adamski J, ed.), pp. 105–121, Elsevier
77. Yu B et al. (2014) Genetic determinants influencing human serum metabolome among African Americans. PLoS Genet. 10, e1004212 [PubMed: 24625756]
78. Yu B et al. (2016) Loss-of-function variants influence the human serum metabolome. Sci. Adv 2, e1600800 [PubMed: 27602404]
79. de Vries PS et al. (2017) Whole-genome sequencing study of serum peptide levels: the Atherosclerosis Risk in Communities study. Hum. Mol. Genet 26, 3442–3450 [PubMed: 28854705]
80. Feofanova EV et al. (2018) Sequence-based analysis of lipid-related metabolites in a multiethnic study. Genetics 209, 607–616 [PubMed: 29610217]

81. Peloso GM et al. (2014) Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. Am. J. Hum. Genet 94, 223–232 [PubMed: 24507774]
82. Liu DJ et al. (2017) Exome-wide association study of plasma lipids in >300,000 individuals. Nat. Genet 49, 1758–1766 [PubMed: 29083408]
83. Signorello LB et al. (2010) The Southern Community Cohort study: investigating health disparities. J. Health Care Poor Underserved 21, 26–37
84. Walejko JM et al. (2018) Gut microbiota and serum metabolite differences in African Americans and white Americans with high blood pressure. Int. J. Cardiol 271, 336–339 [PubMed: 30049487]
85. Tahir UA et al. (2021) Metabolomic profiles and heart failure risk in Black adults: insights from the Jackson Heart study. Circ. Heart Fail 14, e007275 [PubMed: 33464957]
86. Abdrabou W et al. (2021) Metabolome modulation of the host adaptive immunity in human malaria. Nat Metab. 3, 1001–1016 [PubMed: 34113019]
87. Kastenmüller G et al. (2015) Genetics of human metabolism: an update. Hum. Mol. Genet 24, R93–R101 [PubMed: 26160913]
88. Mills MC and Rahal C (2020) The GWAS Diversity Monitor tracks diversity by disease in real time. Nat. Genet 52, 242–243 [PubMed: 32139905]
89. Bentley AR et al. (2019) Multi-ancestry genome-wide gene–smoking interaction study of 387,272 individuals identifies new loci associated with serum lipids. Nat. Genet 51, 636–648 [PubMed: 30926973]
90. Gurdasani D et al. (2019) Uganda genome resource enables insights into population history and genomic discovery in Africa. Cell 179, 984–1002 [PubMed: 31675503]
91. Kuchenbaecker K et al. (2019) The transferability of lipid loci across African, Asian and European cohorts. Nat. Commun 10, 4330 [PubMed: 31551420]
92. Yang Y et al. (2019) Prospective study of oral microbiome and colorectal cancer risk in low-income and African American populations. Int. J. Cancer 144, 2381–2389 [PubMed: 30365870]
93. Serrano MG et al. (2019) Racioethnic diversity in the dynamics of the vaginal microbiome during pregnancy. Nat. Med 25, 1001–1011 [PubMed: 31142850]
94. Fettweis JM et al. (2019) The vaginal microbiome and preterm birth. Nat. Med 25, 1012–1021 [PubMed: 31142849]
95. Faucher MA et al. (2020) Exploration of the vaginal and gut microbiome in African American women by body mass index, class of obesity, and gestational weight gain: a pilot study. Am. J. Perinatol 37, 1160–1172 [PubMed: 31242511]
96. Florova V et al. (2021) Vaginal host immune-microbiome interactions in a cohort of primarily African-American women who ultimately underwent spontaneous preterm birth or delivered at term. Cytokine 137, 155316 [PubMed: 33032107]
97. van de Wijgert J and Jespers V (2017) The global health impact of vaginal dysbiosis. Res. Microbiol 168, 859–864 [PubMed: 28257809]
98. Kyongo JK et al. (2015) Cross-sectional analysis of selected genital tract immunological markers and molecular vaginal microbiota in sub-Saharan African women, with relevance to HIV risk and prevention. Clin. Vaccine Immunol 22, 526–538 [PubMed: 25761460]
99. Mehta SD et al. (2020) Host genetic factors associated with vaginal microbiome composition in Kenyan women. mSystems 5, e00502–20 [PubMed: 32723796]
100. Sivro A et al. (2020) Sex work is associated with increased vaginal microbiome diversity in young women from Mombasa, Kenya. J. Acquir. Immune Defic. Syndr 85, 79–87 [PubMed: 32433252]
101. Bayigga L et al. (2020) Diverse vaginal microbiome was associated with pro-inflammatory vaginal milieu among pregnant women in Uganda. Human Microbiome J. 18, 100076
102. Kelly MS et al. (2018) Pneumococcal colonization and the nasopharyngeal microbiota of children in Botswana. Pediatr. Infect. Dis. J 37, 1176–1183 [PubMed: 30153231]
103. Rubel MA et al. (2020) Lifestyle and the presence of helminths is associated with gut microbiome composition in Cameroonians. Genome Biol. 21, 122 [PubMed: 32450885]
104. Doumatey AP et al. (2020) Gut microbiome profiles are associated with type 2 diabetes in urban Africans. Front. Cell. Infect. Microbiol 10, 63 [PubMed: 32158702]

105. Parbie PK et al. (2021) Fecal microbiome composition in healthy adults in Ghana. Jpn. J. Infect. Dis 74, 42–47 [PubMed: 32611986]
106. Crittenden AN and Schnorr SL (2017) Current views on hunter-gatherer nutrition and the evolution of the human diet. Am. J. Phys. Anthropol 162, 84–109 [PubMed: 28105723]
107. Wilson AS et al. (2020) Diet and the human gut microbiome: an international review. Dig. Dis. Sci 65, 723–740 [PubMed: 32060812]
108. Schnorr SL et al. (2014) Gut microbiome of the Hadza hunter-gatherers. Nat. Commun 5, 3654 [PubMed: 24736369]
109. Gomez A et al. (2016) Gut microbiome of coexisting BaAka pygmies and Bantu reflects gradients of traditional subsistence patterns. Cell Rep. 14, 2142–2153 [PubMed: 26923597]
110. Turroni S et al. (2016) Enterocyte-associated microbiome of the Hadza hunter-gatherers. Front. Microbiol 7, 865 [PubMed: 27375586]
111. Smits SA et al. (2017) Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. Science 357, 802–806 [PubMed: 28839072]
112. Ayeni FA et al. (2018) Infant and adult gut microbiome and metabolome in rural Bassa and urban settlers from Nigeria. Cell Rep. 23, 3056–3067 [PubMed: 29874590]
113. Hansen MEB et al. (2019) Population structure of human gut bacteria in a diverse cohort from rural Tanzania and Botswana. Genome Biol. 20, 16 [PubMed: 30665461]
114. Oduaran OH et al. (2020) Gut microbiome profiling of a rural and urban South African cohort reveals biomarkers of a population in lifestyle transition. BMC Microbiol. 20, 330 [PubMed: 33129264]
115. Tett A et al. (2021) Prevotella diversity, niches and interactions with the human host. Nat. Rev. Microbiol 19, 585–599 [PubMed: 34050328]
116. Jallow M et al. (2009) Genome-wide and fine-resolution association analysis of malaria in West Africa. Nat. Genet 41, 657–665 [PubMed: 19465909]
117. Timmann C et al. (2012) Genome-wide association study indicates two novel resistance loci for severe malaria. Nature 489, 443–446 [PubMed: 22895189]
118. Band G et al. (2013) Imputation-based meta-analysis of severe malaria in three African populations. PLoS Genet. 9, e1003509 [PubMed: 23717212]
119. Malaria Genomic Epidemiology Network (2019) Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. Nat. Commun 10, 5732 [PubMed: 31844061]
120. Graff M et al. (2021) Discovery and fine-mapping of height loci via high-density imputation of GWASs in individuals of African ancestry. Am. J. Hum. Genet 108, 564–582 [PubMed: 33713608]
121. Robertson CC et al. (2021) Fine-mapping, trans-ancestral and genomic analyses identify causal variants, cells, genes and drug targets for type 1 diabetes. Nat. Genet 53, 962–971 [PubMed: 34127860]
122. Vujkovic M et al. (2020) Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. Nat. Genet 52, 680–691 [PubMed: 32541925]
123. Crawford NG et al. (2017) Loci associated with skin pigmentation identified in African populations. Science 358, eaan8433 [PubMed: 29025994]
124. Martin AR et al. (2017) An unexpectedly complex architecture for skin pigmentation in Africans. Cell 171, 1340–1353 [PubMed: 29195075]
125. Lona-Durazo F et al. (2019) Meta-analysis of GWA studies provides new insights on the genetic architecture of skin pigmentation in recently admixed populations. BMC Genet. 20, 59 [PubMed: 31315583]
126. Beleza S et al. (2013) Genetic architecture of skin and eye color in an African-European admixed population. PLoS Genet. 9, e1003372 [PubMed: 23555287]
127. Sturm RA and Duffy DL (2012) Human pigmentation genes under environmental selection. Genome Biol. 13, 248 [PubMed: 23110848]

128. Feng Y et al. (2021) Evolutionary genetics of skin pigmentation in African populations. Hum. Mol. Genet 30, R88–R97 [PubMed: 33438000]
129. Beleza S et al. (2013) The timing of pigmentation lightening in Europeans. Mol. Biol. Evol 30, 24–35 [PubMed: 22923467]
130. Candille SI et al. (2012) Genome-wide association studies of quantitatively measured skin, hair, and eye pigmentation in four European populations. PLoS One 7, e48294 [PubMed: 23118974]
131. Sulem P et al. (2008) Two newly identified genetic determinants of pigmentation in Europeans. Nat. Genet 40, 835–837 [PubMed: 18488028]
132. Bentley AR et al. (2020) Evaluating the promise of inclusion of African ancestry populations in genomics. NPJ Genom. Med 5, 5 [PubMed: 32140257]
133. Jongeneel CV et al. (2017) Assessing computational genomics skills: our experience in the H3ABioNet African bioinformatics network. PLoS Comput. Biol 13, e1005419 [PubMed: 28570565]
134. Ras V et al. (2021) Using a multiple-delivery-mode training approach to develop local capacity and infrastructure for advanced bioinformatics in Africa. PLoS Comput. Biol 17, e1008640 [PubMed: 33630830]
135. Beichman AC et al. (2018) Using genomic data to infer historic population dynamics of nonmodel organisms. Annu. Rev. Ecol. Evol. Syst 49, 433–456
136. Mather N et al. (2020) A practical introduction to sequentially Markovian coalescent methods for estimating demographic history from genomic data. Ecol. Evol 10, 579–589 [PubMed: 31988743]
137. Schiffels S and Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. Nat. Genet 46, 919–925 [PubMed: 24952747]
138. Chen L et al. (2020) Identifying and interpreting apparent Neanderthal ancestry in African individuals. Cell 180, 677–687 [PubMed: 32004458]
139. Ng PC and Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 31, 3812–3814 [PubMed: 12824425]
140. Adzhubei I et al. (2013) Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet Chapter 7, Unit7.20
141. Cooper GM et al. (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 15, 901–913 [PubMed: 15965027]
142. Rentzsch P et al. (2019) CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 47, D886–D894 [PubMed: 30371827]
143. McLaren W et al. (2016) The Ensembl variant effect predictor. Genome Biol. 17, 122 [PubMed: 27268795]
144. Vitti JJ et al. (2013) Detecting natural selection in genomic data. Annu. Rev. Genet 47, 97–120 [PubMed: 24274750]
145. Tishkoff SA et al. (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nat. Genet 39, 31–40 [PubMed: 17159977]
146. Patin E et al. (2017) Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. Science 356, 543–546 [PubMed: 28473590]
147. Ko W-Y et al. (2013) Identifying Darwinian selection acting on different human APOL1 variants among diverse African populations. Am. J. Hum. Genet 93, 54–66 [PubMed: 23768513]
148. Fan S et al. (2016) Going global by adapting local: a review of recent human adaptation. Science 354, 54–59 [PubMed: 27846491]
149. Kudaravalli S et al. (2009) Gene expression levels are a target of recent natural selection in the human genome. Mol. Biol. Evol 26, 649–658 [PubMed: 19091723]
150. Wang Y et al. (2020) Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. Nat. Commun 11, 3865 [PubMed: 32737319]
151. Berg JJ and Coop G (2014) A population genetic signal of polygenic adaptation. PLoS Genet. 10, e1004412 [PubMed: 25102153]
152. Mulder N et al. (2018) H3Africa: current perspectives. Pharmgenomics Pers. Med 11, 59–66 [PubMed: 29692621]

153. Asiki G et al. (2013) The general population cohort in rural south-western Uganda: a platform for communicable and noncommunicable disease studies. Int. J. Epidemiol 42, 129–141 [PubMed: 23364209]

154. Zar HJ et al. (2015) Investigating the early-life determinants of illness in Africa: the Drakenstein Child Health study. Thorax 70, 592–594 [PubMed: 25228292]

155. Shah-Williams E et al. (2020) Enrollment of diverse populations in the INGENIOUS pharmacogenetics clinical trial. Front. Genet 11, 571 [PubMed: 32670350]

156. Lloyd-Price J et al. (2017) Erratum: strains, functions and dynamics in the expanded Human Microbiome Project. Nature 551, 256

157. Signorello LB et al. (2005) Southern Community Cohort study: establishing a cohort to investigate health disparities. J. Natl. Med. Assoc 97, 972–979 [PubMed: 16080667]

158. Jensen ET et al. (2020) Rationale, design and baseline characteristics of the Microbiome and Insulin Longitudinal Evaluation Study (MILES). Diabetes Obes. Metab 22, 1976–1984 [PubMed: 32687239]

159. Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet 10, 669–680 [PubMed: 19736561]

160. Song L and Crawford GE (2010) DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb. Protoc 2010 db.prot5384

161. Gu H et al. (2011) Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat. Protoc 6, 468–481 [PubMed: 21412275]

162. Wang Z et al. (2009) RNA-seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet 10, 57–63 [PubMed: 19015660]

163. Gawad C et al. (2016) Single-cell genome sequencing: current state of the science. Nat. Rev. Genet 17, 175–188 [PubMed: 26806412]

164. Aebersold R and Mann M (2003) Mass spectrometry-based proteomics. Nature 422, 198–207 [PubMed: 12634793]

165. Stoevesandt O and Taussig MJ (2012) Affinity proteomics: the role of specific binding reagents in human proteome analysis. Expert Rev. Proteomics 9, 401–414 [PubMed: 22967077]

166. Timp W and Timp G (2020) Beyond mass spectrometry, the next step in proteomics. Sci. Adv 6, eaax8978 [PubMed: 31950079]

167. Emwas A-H et al. (2019) NMR spectroscopy for metabolomics research. Metabolites 9, 123

168. Fricker AM et al. (2019) What is new and relevant for sequencing-based microbiome research? A mini-review. J. Advert. Res 19, 105–112

169. Chaudhary DK and Dahal RH (2017) DNA bar-code for identification of microbial communities: a mini-review. EC Microbiol. 6, 219–224

170. Breitwieser FP et al. (2019) A review of methods and databases for metagenomic classification and assembly. Brief. Bioinform 20, 1125–1136 [PubMed: 29028872]

171. Roumpeka DD et al. (2017) A review of bioinformatics tools for bio-prospecting from metagenomic sequence data. Front. Genet 8, 23 [PubMed: 28321234]

Highlights

Africans harbor great genetic, phenotypic, cultural and linguistic diversity. Disparity remains in genetic studies representing African ancestries.

Whole-genome sequencing (WGS) for African populations lags behind European and Asian ancestries. More WGS efforts to capture genetic variation in Africa are warranted.

The human reference genome (HRG) lacks diversity, which can bias downstream analysis. Population-specific reference genomes for African populations would enable the detection of more complex variation and variation that would be difficult to map to the current HRG.

Combining genomic data with intermediate phenotypes is required to understand the biological mechanisms underlying phenotype and disease in the post-genomic era.

African integrative genomics is in its infancy. The scientific rewards and translational health benefits of dramatically expanded integrative omics studies in African populations is high.

Box 1.

Inferring African evolutionary history and functional variants from WGS data

One can reconstruct the demographic history of a population from analysis of genomes, as genomes carry a record of the evolutionary forces that have shaped the corresponding populations (see methods reviewed elsewhere [135,136]). Using the multiple sequentially Markovian coalescent method [136,137], Fan *et al.* constructed an evolutionary history for modern African populations with 30× WGS data from diverse ancestries [25]. The San lineage, comprising hunter-gatherer populations in southern Africa, is basal to all modern human lineages. It diverged from Niger-Congo, Afroasiatic, and Nilo-Saharan lineages as early as 160 kya (thousand years ago). The San and other hunter-gatherer lineages, such as Central African rainforest hunter-gatherer (CRHG), Hadza hunter-gatherer, and Sandawe hunter-gatherer, were diverged by ~100 kya, followed by more recent divergence of non-hunter-gatherer lineages; Niger-Congo, Nilo-Saharan, and Afroasiatic lineages diverged by ~54–16 kya. Eastern and western CRHG lineages diverged by ~50–31 kya and the western CRHG lineages diverged by ~18–12 kya. The San and CRHG populations maintained the largest effective population size compared with other populations prior to 60 kya [25].

Analysis based on WGS data confirmed signatures of multiple waves of recent migration events within Africa, such as the expansion of Bantu-speaking agriculturists from western Africa to eastern and southern Africa by 5–3 kya [25,27]. Ancient archaic introgression events from archaic hominids to indigenous Africans have also been identified [18,26,138].

Tests for natural selection are also useful for identifying functionally important variation in the genome. Genomic regions that show signatures of evolutionary constraint are likely to be functional, as deleterious variants will be removed from the population/species. Many constraint-based algorithms have been developed to measure the deleteriousness of protein-coding variants (e.g., SIFT [139] and PolyPhen2 [140]) and noncoding variants (e.g., CADD [141] and GERP [142]), which have been widely embedded in annotation tools [143]. Additionally, approaches for detecting positive selection enable the identification of adaptive variants [144]. Signatures of positive selection observed in Africans include genes associated with diet (e.g., *LCT* [145]), skin color (e.g., *SLC24A5* and *MFSD12* [123]), short stature (e.g., *PITX1* and *THRB* [25]), and immunity (e.g., *HLA* locus [146] and *APOL1* [27,147]) [148]. However, the causative variants and the target phenotypes under natural selection are not always clear. Regulatory variants influencing gene expression are enriched for targets of recent natural selection [149]. Integrative genomic research will shed light on the functional impacts of variants under natural selection.

Box 2.

Transferability of polygenic risk score (PRS) between populations

Transferability refers to the ability that the research study's findings are applicable in other contexts, situations, times, and populations. A **polygenic risk score**, or PRS, is an effect-weighted sum of the number of risk alleles an individual carries, calculated according to their genotypes and the allelic effects estimated from relevant GWAS summary statistics.

The transferability of PRS is often quite low between human populations for complex traits [9]. This is mainly attributed to population differences in linkage disequilibrium (LD) and allele frequencies (Figure I) [150]. Specifically, the extent of LD in Africans is lower than non-Africans because non-Africans experienced an out-of-Africa bottleneck. The variants tagging causal variants in one population may not be in LD with the causal variants in other populations. Differences in allele frequency can lead to low transferability if a common causal variant in a target population is too rare in the GWAS discovery cohort to detect via association tests. For example, causal alleles of height in the Central African rainforest hunter-gatherer (CRHG) population commonly referred to as 'pygmy', who have an average adult male height (~150 cm) may be common in that population but are rare in most other populations. This may explain why height PRS based on European GWAS do not always predict extremely low values for height when applied to CRHG populations [151]. Therefore, predicting disease risk based on PRS from GWAS results of one population may not apply to another (reviewed by Sirugo *et al.* [11]).

The low transferability of PRS may also be due to gene-by-environment effects that differ between populations (Figure I). Such factors may include regional climate, communicable diseases, and traditional subsistence practices. Considering the aforementioned factors, including ethnically diverse African populations in GWAS analysis will be necessary to establish unbiased estimates of genetic risk across global populations.

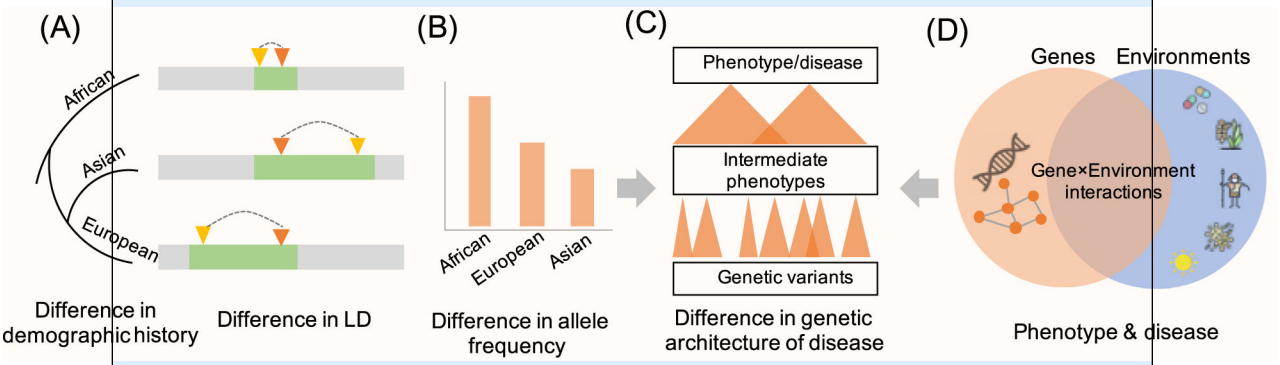

Figure I. Factors that shape the genetic architecture of traits and disease risk.

The different demographic histories of populations lead to genetic differences due to drift and local adaptation, which in turn shapes population patterns of linkage disequilibrium (LD) (A) and allele frequencies (B). In trait mapping studies, causative variants (dark

yellow triangle in A) may be tagged by different variants in different populations (yellow triangles in A). Intermediate phenotypes are the building blocks of overarching complex traits and are impacted by underlying genetic variation (C). Environmental and lifestyle factors can also interact with underlying genetic variation to influence phenotypes (D).

Box 3.

Large-scale African genomics consortiums

Recently, there have been large-scale integrative genomics projects, which include a large number of African ancestry individuals, that are the current models for international collaborative omics research. The H3Africa consortium is a prime example of an international scientific institution working to both create cutting edge genomics research in an ethical manner and to help build African scientific infrastructure. H3Africa was funded by the National Institutes of Health (NIH) and the Wellcome Trust to facilitate innovative research into the genetic and environmental basis for diseases affecting African populations[ii]. The consortium is collectively processing samples and data for over 70 000 participants across the African continent, accompanied in most cases by rich clinical information on a variety of noncommunicable and infectious diseases such as heart and renal disease, as well as communicable diseases such as tuberculosis [152]. The initiative consists of 51 African projects that are increasingly providing novel insights into the genetic basis of diseases in indigenous populations; insights that have the potential to drive the development of new diagnostics and treatments. H3Africa has established comprehensive guidelines for informed consent, community engagement, the return of individual genetics research findings, and publication policy. It also provided a comprehensive ethics and governance framework for best practice in genomic research and biobanking in Africa[iii].

However, H3Africa currently is not designed specifically for integrating genomics across multiple omics modalities. One of the primary global integrative omics projects is the Trans-Omics for Precision Medicine (TOPMed) program, sponsored by the NIH National Heart, Lung, and Blood Institute (NHLBI), which aims to provide disease treatments tailored to an individual's unique genes and environment by integrating WGS and other omics data[iv]. Most participants have rich phenotype data related to cardiovascular function. The current TOPMed cohort contains over 50 000 individuals with African ancestry. Ultimately, integrative African omics will require the scientific infrastructure and oversight of an organization like H3A combined with the technological integrations being pioneered by TOPMed.

In addition to H3Africa and TOPMed, other notable integrative genomics cohorts for Africans are the Ugandan General Population Cohort, which is a population-scale biomedical cohort containing clinical phenotyping, including many blood biomarkers [90,153], and the Drakenstein Child Health Survey [154], which is a longitudinal population-based birth cohort from peri-urban South African populations. African biomedical research cohorts with a metabolomics and genetic component include the INGENIOUS pharmacology-focused cohort [155], the Human Microbiome Project phase 1 and 2 for microbiome data [156], the Southern Community Cohort Study [84,157]

[ii] https://h3africa.org/
[iii] https://h3africa.org/wp-content/uploads/2018/05/Final-Framework-for-African-genomics-and-biobanking_SC-.pdf
[iv] www.nhlbiwgs.org/

for metabolomic and microbiome data, and the Microbiome and Insulin Longitudinal Evaluation Study [158] which is focused on type 2 diabetes research.

Box 4.

Advanced integrative genomic technologies

NGS technologies have advanced large-scale 'omics' research in the past decade (Figure 2). Powerful new techniques for mapping gene regulatory networks include ChIP-seq [159] for identifying the binding sequences of transcription factors and the genomic location of histone modification, DNase-seq for identifying open regions of chromatin [160], and bisulfite sequencing to determine the location of DNA inactivation from CpG methylation [161]. In transcriptomic analysis, RNA-seq, lncRNA-seq, and smRNA-seq data are commonly used to measure the expression of mRNA, long noncoding RNA, and small RNA, respectively [162]. Single-cell RNA sequencing (scRNA-seq) provides the expression profiles of individual cells [163].

In both proteomic and metabolomic research, the most commonly used technology is mass spectrometry (MS) [164], which aims to measure the mass-to-charge ratio and relative abundances of all proteins present in a sample at once. The two main types of MS are targeted and untargeted. Targeted MS measures a smaller number of preselected proteins or molecules, while untargeted MS measures all proteins or molecules in the sample [70]. Complementary approaches use affinity-based assays (e.g., Olink, and SomaLogic), which are based on the use of antibodies and other binding reagents as protein-specific detection probes [165]. Emerging protein sequencing methods can both identify whole proteins and also sequence them in one experiment [166]. Other methods used to characterize metabolomic variation include high or ultrahigh performance liquid chromatography coupled to UV or fluorescent detection and nuclear magnetic resonance spectroscope [167].

In microbiomic analysis, microbiomes are typically measured based on either amplicon sequencing of specific genes or genic subdomains, suitable for phylogenetic profiling, or metagenomic (shotgun) sequencing of all DNA present in a sample [168]. Commonly used genes are 16S ribosomal RNA for bacteria, 18S or 28S ribosomal RNA for archaea, and fungal internal transcribed spacers (ITS) for fungi [169]. In contrast to focusing on specific genes, shotgun sequencing of short reads (~100–300 bp) provides information on all gene content of a sample and can be used to infer taxonomic abundances, identify new genes, and study gene function [170,171].

Outstanding questions

How much genetic diversity in African genomes are missing in the current global whole-genome sequencing (WGS) reference datasets? To what extent does the current human reference genome (HRG) bias African variant calling?

What is an effective strategy to construct a population-specific reference genome for African populations?

How many functional sequences (e.g., genes, enhancers) are in the unmapped reads from African WGS data and what are their functions?

Does the high degree of environmental diversity between African populations cause population-level differences in intermediate phenotypes (transcriptomes, proteomes, metabolomes, and microbiomes)? For example, do metabolomes differ by broad dietary types? Are immune-related genes differentially expressed based on regional disease burdens?

Are there African-specific genetic determinants underlying differences in 'intermediate phenotypes' such as gene expression and metabolome?

What are the genetic determinants for the high prevalence of communicable (e.g., malaria) and noncommunicable (e.g., type 2 diabetes) diseases in Africa?

How have genetic adaptations to diverse environments (e.g., tropical rainforests) and a wide variety of subsistence patterns (e.g., hunting-gathering) impacted risk for disease in individuals with African ancestry?

How can we expand the African biomedical research community, including increasing training and capacity for genomics research within the African continent and developing online tools and computational resources for African research scientists?

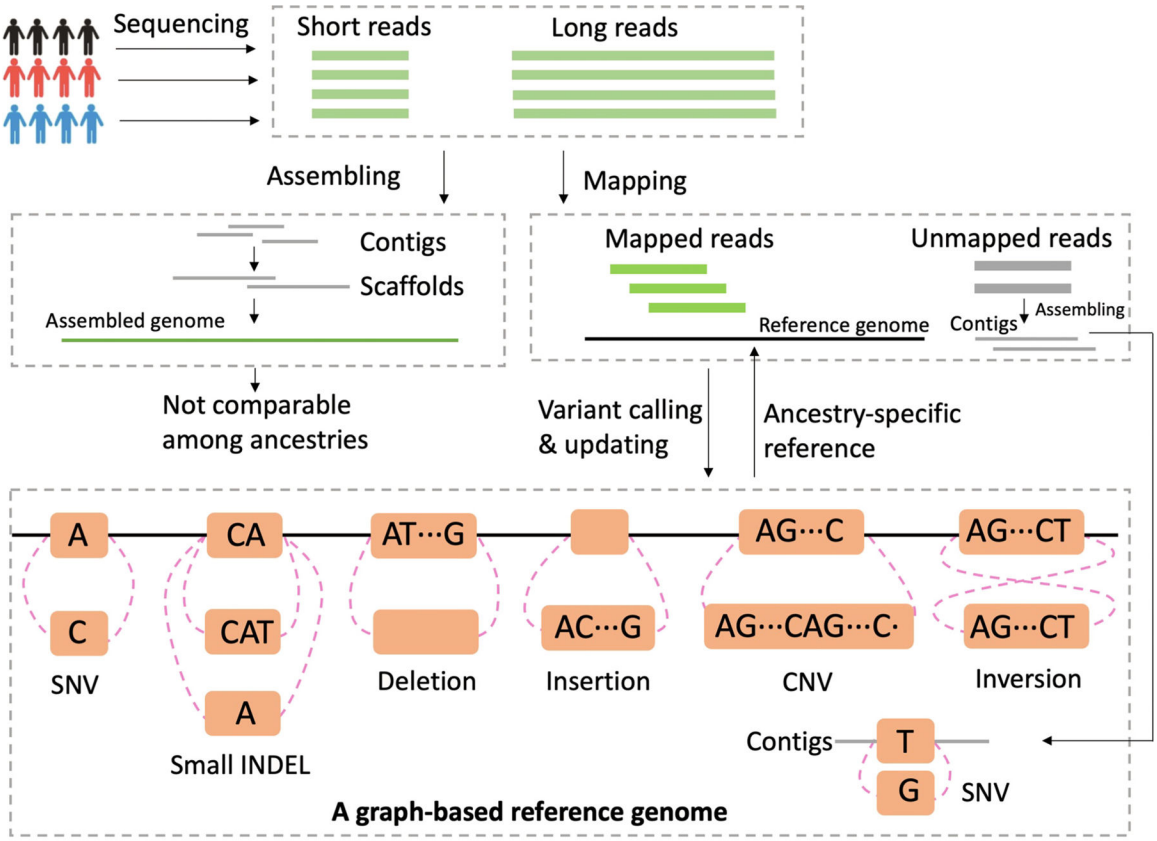

Figure 1. Illustration of graph-based population-specific reference genomes.

Constructing population-specific reference genomes will need the input of sequenced reads from the ancestry of interest. Both short- and long-read sequences are informative for graph reference genomes. Short-read sequencing is relatively inexpensive (~$1000–2000/genome at 30×), allowing greater sampling per total cost. Long-read sequencing, though more expensive (~$5000–20 000/genome at 30×), is required to detect complex genetic variation such as structural variants. Reads can be used for *de novo* assembly of scaffolds or can be mapped to a reference sequence for variant detection. *De novo* genome assembly is computationally intensive and is difficult to compare between populations because of the lack of genomic coordinates. Reads mapping to a reference genome that lacks genetic variation present in the ancestry of interest can result in biases in mapping and variant calling processes. Thus, a considerable proportion of unmapped reads may be observed. The traditional linear reference genome has limited power to accomodate genetic variation within a specific population, while a graph-based reference genome has the ability to accommodate complex individual genetic variation (dark orange boxes) as paths (pink dashed lines) through a graph. Detected variation can be easily updated to the original graph-based reference genome. The updated graph-based reference genome could provide a good reference for mapping and variant calling in future sequencing efforts. Unmapped reads could be assembled to contigs, which would be included in the graph-based reference genome to ensure the completeness of the reference genome. Abbreviations: CNV, copy number variant; SNV, single-nucleotide variant.

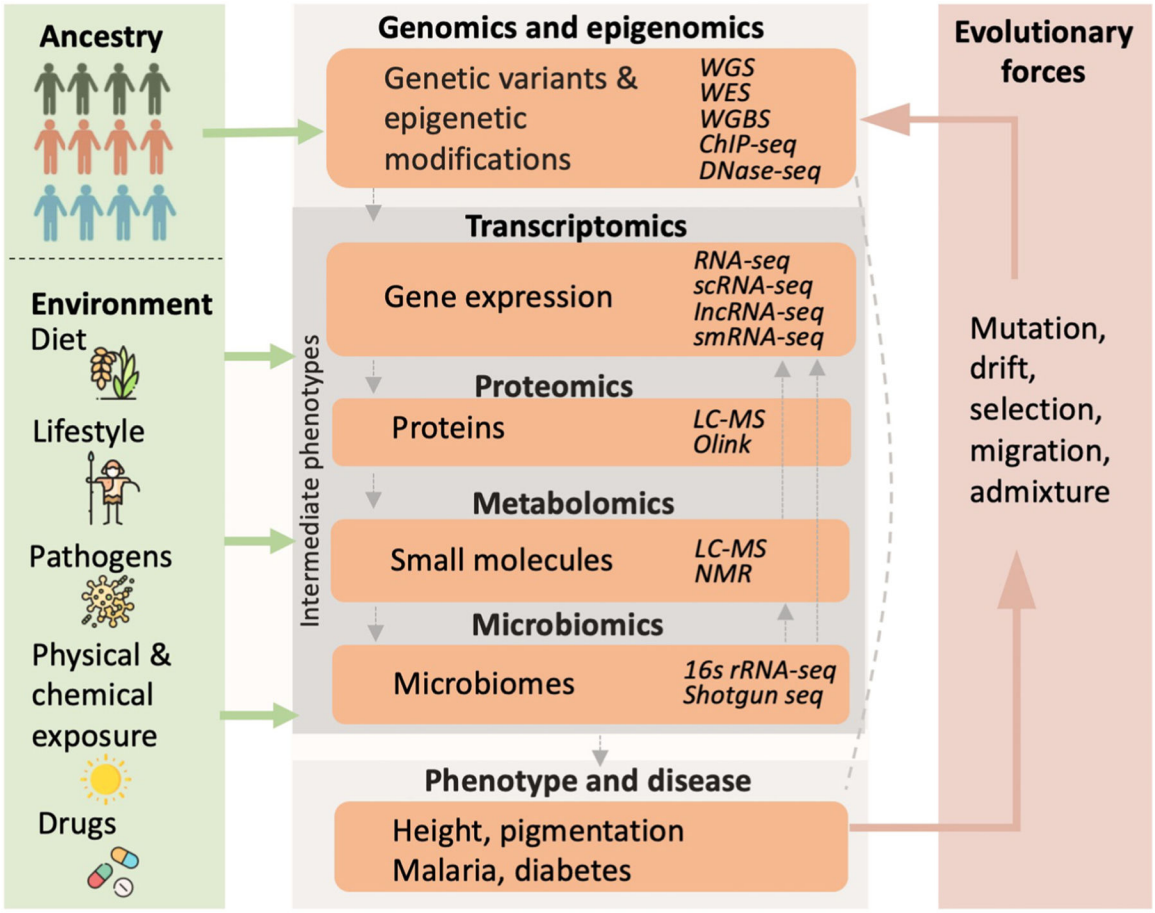

Figure 2. Integrative genomics approach to study variant function.

The relationships between genetic ancestry, the environment, fitness, individual genetics, omics data, intermediate phenotypes, and endpoint phenotypes, including diseases, are illustrated. Omics data generally represent measurements of intermediate phenotypes (dark gray box in the middle panel) that link underlying genetic variation to outcome phenotypes or disease. Integrating genomic data with intermediate phenotypes enables identification of the quantitative trait loci (QTLs), such as expression QTLs (eQTLs), protein QTLs (pQTLs), and methylation QTLs (mQTLs). Both genetic ancestry and environment can impact intermediate phenotypes and associations with specific variants, either because of gene–gene epistatic effects (G×G) or because of genetic effects that are only relevant in certain environmental contexts or given specific environmental triggers (G×E). Population differences in linkage disequilibrium and allele frequency could also result in decreased transferability of polygenic risk scores (Box 2). Evolutionary forces, such as drift and natural selection, can shape the genomic diversity of populations (right panel). Detecting signatures of natural selection can help identify functional variants, as selection only acts upon functional variation (Box 1). Abbreviations: ChIP-seq, chromatin immunoprecipitation sequencing; DNase-seq, DNase I hypersensitive sites sequencing; LC-MS, liquid chromatography–mass spectrometry; lncRNA-seq, long noncoding RNA sequencing; NMR, nuclear magnetic resonance; scRNA-seq, single-cell RNA sequencing;

smRNA-seq, small RNA sequencing; WES, whole-exome sequencing; WGBS, whole-genome bisulfite sequencing; WGS, whole-genome sequencing.