Molecular Genetics and Genomics (2020) 295:1079–1089
https://doi.org/10.1007/s00438-020-01684-8

REVIEW

Prospective avenues for human population genomics and disease mapping in southern Africa

Yolandi Swart[1] · Gerald van Eeden[1] · Anel Sparks[1] · Caitlin Uren[1] · Marlo Möller[1]

Received: 13 November 2019 / Accepted: 6 May 2020 / Published online: 21 May 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract

Population substructure within human populations is globally evident and a well-known confounding factor in many genetic studies. In contrast, admixture mapping exploits population stratification to detect genotype–phenotype correlations in admixed populations. Southern Africa has untapped potential for disease mapping of ancestry-specific disease risk alleles due to the distinct genetic diversity in its populations compared to other populations worldwide. This diversity contributes to a number of phenotypes, including ancestry-specific disease risk and response to pathogens. Although the 1000 Genomes Project significantly improved our understanding of genetic variation globally, southern African populations are still severely underrepresented in biomedical and human genetic studies due to insufficient large-scale publicly available data. In addition to a lack of genetic data in public repositories, existing software, algorithms and resources used for imputation and phasing of genotypic data (amongst others) are largely ineffective for populations with a complex genetic architecture such as that seen in southern Africa. This review article, therefore, aims to summarise the current limitations of conducting genetic studies on populations with a complex genetic architecture to identify potential areas for further research and development.

Keywords Southern Africa · Population genetics · Admixture mapping · Disease risk alleles

Introduction

Genetics entered an exciting era of discovery with the advent of next-generation sequencing (NGS) technology, improved bioinformatic techniques and increased international collaboration to include underrepresented diversity in genetic studies. Collaborative initiatives, such as the Human Heredity and Health in Africa (H3Africa) Consortium and, African Genome Variation Project (AGVP) are rapidly obtaining and investigating valuable genetic data previously unattainable (Gurdasani et al. 2015; Zheng-Bradley and Flicek 2017; Fortes-Lima et al. 2017; Mulder et al. 2018). These studies have enabled novel genetic investigations, however large sample sizes and high-quality whole-genome data are still lacking for most populations from particularly southern Africa. Association studies often yield no significant single nucleotide polymorphisms (SNPs) associated with multifactorial diseases and fail to detect associations with rare genetic variants [minor allele frequency (MAF) of $< 1\%$] in southern African populations due to a lack of predictive power. Furthermore, the vast majority of association studies continue to be focused on populations of European ancestry and simple admixture scenarios (Wojcik et al. 2019).

Genetic regions associated with multifactorial diseases could be identified by investigating the allelic architecture of highly complex admixed individuals, since they received haplotypes from diverse continental populations previously exposed to various environments and pathogens (Dias-Alves et al. 2018; Mazandu et al. 2019). If such gene regions could be successfully identified, it will aid in the advancement of drug therapies, implementation of personalized medicine and vaccine development in underdeveloped countries such as South Africa. However, individuals from South Africa can be up to five-way admixed, arguably the most complex global example of admixture (Daya et al. 2013; Uren et al.

Communicated by S. Hohmann.

Caitlin Uren and Marlo Möller: co-senior authors.

✉ Marlo Möller
marlom@sun.ac.za

[1] DSI-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, Cape Town, South Africa

2017b). The history of South Africa contributed to this observed population substructure, one of which includes ancestral contributions from predominantly the indigenous hunter-gatherers of southern Africa and Bantu-speaking Africans, as well as European-descent groups, South East Asians and East Asians (de Wit et al. 2010; Chimusa et al. 2014; Daya et al. 2014b; Uren et al. 2016).

The frequency of disease risk alleles differs between populations (Secolin et al. 2019). These disparities are exploited to map disease-causing variants of multifactorial diseases in admixed genomes, better known as admixture mapping (Shriner 2013). However, additional modifications are required to conduct admixture mapping studies for individuals from southern Africa, since most computational tools are designed to infer local ancestry for two-or three-way admixed populations only (Chimusa et al. 2018; Schurz et al. 2019; Mazandu et al. 2019). In addition, statistical methods assume homogeneity and may not be applicable for Africans with more complex haplotype structures and mosaic patterns present on chromosomes generated by recent admixture events across the African continent (Fan et al. 2019). The continuous increase in non-communicable diseases in Africa and the persistent threat of emerging and re-emerging infectious diseases could in part be countered by the development of comprehensive research pipelines for disease mapping in highly admixed individuals from Africa. This review, therefore, aims to summarise the current limitations of and prospective avenues for population genomics research in relation to disease mapping in southern African populations.

Admixture mapping

The conventional method of genome-wide association studies (GWAS), which is a hypothesis-free method of detecting SNPs associated with a certain disease, is not sufficient for detecting SNPs associated with a disease in a population with admixed genomes (Visscher et al. 2017). In contrast, admixture mapping (study design summarised in Fig. 1) acknowledges biogeographic ancestry associated with specific phenotypes by including ancestry proportions (globally or locally inferred from dense genotypic data) as covariates in epidemiological studies (Thornton and Bermejo 2014; Duan et al. 2018). Instead of relying on the association between a genotype and a phenotype, as in GWAS, it considers the associations of the number of haplotypes (0, 1 or 2) of a specific ancestry with the phenotype of interest (Hoggart et al. 2004; Duan et al. 2018). The study design has recently been successfully used in a variety of complex diseases, e.g. hypertension (Zhu et al. 2005), prostate cancer in African Americans (Freedman et al. 2006), asthma-associated variants in Latinos (Gignoux et al. 2019), obstructive sleep apnea in Hispanic/Latino Americans (Wang et al. 2019), tuberculosis (TB) associated variants in South Africans (Daya et al. 2014a) and multiple sclerosis in 3692 African Americans, 3777 Hispanics and 4915 Asian Americans (Chi et al. 2019).

The successful implementation of admixture mapping relies on suitable proxy haplotype reference panels that represent each ancestral group. The reference panels are used to infer the number of haplotypes that originated from a specific ancestral source population at a given locus, better known as local ancestry inference (LAI) (Paşaniuc et al. 2009; Dias-Alves et al. 2018). However, limited haplotype reference panels are available for southern African populations. Furthermore, admixture between ethnic groups creates long-range linkage disequilibrium (LD) between variants from different ethnic groups with different allelic frequencies. This subsequently results in differing ancestral haplotype LD blocks, which holds implications for the use of tagging SNPs when working with admixed individuals (Skotte et al. 2019). Tagging SNPs are normally genotyped whilst conducting association studies and act as a proxy for the underlying common disease-causing variants. Association signals depend on how strongly tagging SNPs correlate with the presence of the disease-causing variant located in a haplotype LD block (Hellwege et al. 2017). The tagging SNPs could be in the same ancestral LD block, but due to admixture induced LD, be located in a different haplotype LD block and no longer tag the original LD block containing the causal variant. Therefore no association signal would be detected (Skotte et al. 2019). This is predominantly evident when genetic heterogeneity is present in the admixed population (Duan et al. 2018). Genetic heterogeneity within a population in this context refers to the individuals of the population having different proportions of global ancestry and/or differing local ancestry at a given locus (Duan et al. 2018).

The increasing number of admixed individuals is an evolving obstacle for epidemiological association studies. Even populations assumed to be unadmixed may harbour fine-scale admixture. A different admixture mapping approach is required for southern African populations since most individuals will have ancestral contributions from more than two different non-intermating admixed subpopulations of unknown origin with different effect sizes (Bostoen 2018). The Bantu expansion shaped the genetic composition of most populations from southern Africa, which consists of the countries Namibia, Botswana, Eswatini, Lesotho and South Africa, the latter having 11 official languages which reflect the main ethnic groups (Uren et al. 2016). Differential admixture dynamics were experienced by Bantu-speaking communities in different areas of south-eastern Africa and the indigenous populations were most affected by this event (Pickrell et al. 2012; González-Santos et al.

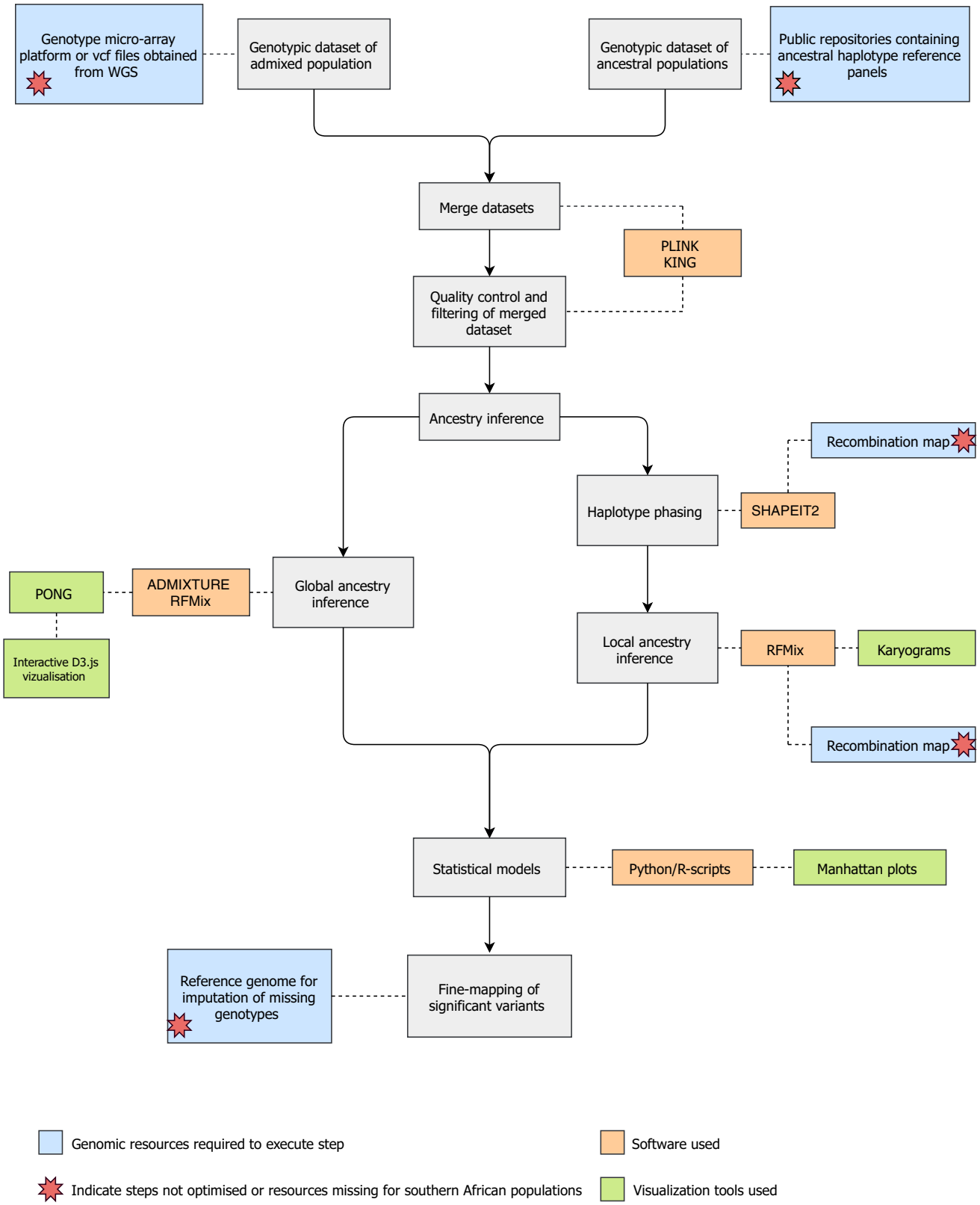

Fig. 1 Flow diagram indicating resources and software used for admixture mapping. Black blocks indicate the analysis steps, orange blocks represent the software used to conduct the relevant step, blue blocks indicate the resource required for the step and green blocks indicate software or approaches used for visualization. The red stars indicate missing or inadequate resources for executing the analysis step in South African populations

2015; Beltrame et al. 2016; Fan et al. 2019; Tucci and Akey 2019). The migration from west to eastern- and southern Africa of Bantu-speaking Africans disrupted local communities by displacement or admixture which led to genetic and cultural exchange. This caused agricultural expansion (González-Santos et al. 2015). Several aspects, therefore, require consideration before conducting genetic studies in southern African populations. Firstly, the extensive population heterogeneity amongst Africans caused by complex genetic population sub-structure and differing levels of admixture (Choudhury et al. 2018; Fan et al. 2019). These groups can therefore not be grouped as one population in genomic studies (Patin et al. 2017). Secondly, differences in LD between populations of African descent causes different haplotype structures resulting in reduced power to detect untyped causal loci (Campbell and Tishkoff 2008). Thirdly, derived alleles are more likely to be heterozygous instead of homozygous (Barnes et al. 2007).

Although admixture mapping increases power to detect disease-associated variants, due to longer ancestral LD blocks than haplotype blocks, differing amounts of ancestry and LD patterns from unknown ancestral populations could be present in each individual in the cohort under study (Zhu and Wang 2017). Adjusting by local ancestry, reflecting the admixture induced LD within admixed populations, will significantly improve the detection of genetic variants with small or moderate effect when extensive genetic heterogeneity is present in the admixed population under study (Duan et al. 2018). However, it is important to also correct for global ancestry proportions, whilst correcting for local ancestry, since association testing will take place in the context of the admixture induced LD blocks in the admixed genome (Duan et al. 2018). This emphasizes the importance of characterising fine-scale genome variation among underrepresented southern African populations to facilitate complex-trait mapping amongst those who harbour a significant burden of global disease.

Currently, no universal standard operating procedures for admixture mapping exists due to the unique admixture scenarios of each project (Zhang and Stram 2014). Considering finer details of population history and ancestry locally inferred for each haplotype will become a requirement for genomic studies among admixed individuals (Duan et al. 2018). This is especially true for populations with complex histories, such as those found in southern Africa.

Genomic resources

Population-specific reference genomes

The current human reference genome was essential in advancing genetic analysis such as imputation of missing genotypes required for GWAS and admixture mapping studies, as well as the identification of rare variants in mostly European and Asian populations (Ikegawa 2012). It is estimated that any two humans will share approximately 99.9% of their genomes (Li and Sadler 1991). A 0.1% difference might seem insignificant, although in a human genome this translates to approximately 3 million base pairs (Suwinski et al. 2019). Since the current reference systematically under represents the tremendous global sequence diversity it is necessary to focus on the individual, geographically defined populations.

Reference genomes have been assembled for multiple distinct human populations and have empirically proven their importance by improving both short read mapping and genotype calling. It facilitates the assessment and meta-analysis of studies done on different microarrays by imputing missing genotypes and is regularly utilized for improving imputation of low frequency and rare variants (Vergara et al. 2018). Despite the decrease in the cost of sequencing, microarrays are still the technology of choice, although these do have certain limitations. Most notably, probes on the microarray are designed from publicly available data to be an exact complement to the desired genomic region to maximize genotyping rate. If not, it will result in a lower genotyping rate and is especially problematic when considering the higher base-pair substitution rate of intergenic versus protein-coding regions (Cargill et al. 1999; Halushka et al. 1999; Leabman et al. 2003). Another limitation is the implication of a probe binding within structural variants (SV) which have been shown to be greatly population-specific (Rosenfeld et al. 2012; Sudmant et al. 2015). If sequencing is selected, the raw data still requires conversion to genomic data either through de novo assembly or by alignment to a reference genome. The latter approach, which is less computationally demanding, has not always allowed the detection of SVs. Due to the read length cut-off of standard NGS, this method of variant detection has a resolution perfect for single nucleotide variants (SNV), but is too small to accurately detect larger events (Merker et al. 2018).

The failure to detect variants of any size in genomic data due to a lack of a population-specific reference genome has multiple consequences. Firstly, it has an impact on the ability to identify potentially disease-associated variants in patients with underlying genetic conditions. This is applicable not only to the discovery of the variant but extends to the precise clinical diagnosis and subsequent treatment. Detecting functionally relevant genetic variants with increased accuracy with the help of a population-specific reference genome brings precision medicine to the forefront of diagnostic/treatment options. Secondly, the failure to detect variants has consequences for the generation and maintenance of allele frequency databases as the failure to accurately curate these findings overburden variant prioritization pipelines. Whilst

considering genomic variants linked to a condition that is more prevalent in a certain population, it is preferable to compare the genomes to a reference genome more representative of that population.

Currently, no reference genome exists that adequately represents African genetic diversity and it is unlikely that a quality reference genome representative of all populations will be achieved. Rare genetic variants could be missed when conducting studies on participants from African origins since Africans contain ± 10% more DNA (± 3 million base pairs) than the presently available human reference genome (Sherman et al. 2019). Regions as large as 100,000 base pairs were identified and 387 novel contigs were in 315 distinct protein-coding regions (Sherman et al. 2019). Furthermore, the current $2.7 billion Human Genome Reference build 38 (GRCh38) lacks genetic variation from individuals worldwide. A study conducted by Yang et al. identified certain biases in GRCh38 by sequencing three Africans, three Asians, two Europeans and three Americans with PacBio single molecule, real-time (SMRT) sequencing and comparing these to GRCh38, 174 individuals from the 1000 Genome Project (1000GP) and 266 individuals from the Simons Genomes Variation Project (SGDP). A total of 40.8% (99,604 nonredundant SVs) were novel compared to previously published large-scale projects (Yang et al. 2019). The SGDP obtained high quality (average coverage of 43-fold) whole genomes from 142 diverse populations and indicated that these genomes include at least 5.8 million base pairs that are not present in GRCh38 (Mallick et al. 2016). Additionally, a study in Sweden identified 61,000 novel genetic sequences from 1000 individuals that were missing in GRCh38 and nearly 40% of the genetic material couldn't be mapped (Eisfeldt et al. 2020).

Previous attempts to capture underrepresented southern African genetic variation have been made by the 1000GP, AGVP and SGDP to include more genome-wide genetic markers for broader groups of southern African populations (Gurdasani et al. 2015; 1000 Genomes Project Consortium et al. 2015). The 1000GP, the largest whole-genome sequencing survey, analysed 26 populations from Europe, East Asian, South Asian, the Americas and Africa. Low coverage sequencing was used and the focus was on demographically large populations, while smaller populations were excluded, despite their respective contributions to human diversity. Although five African populations were included in the analysis, most of these populations are of recent Niger-Kordofanian ancestry (West-and East African) and do not reflect the diversity present in southern Africa (1000 Genomes Project Consortium et al. 2015). Furthermore, 11%, 5% and 5% of heterozygous positions in KhoeSan, New Guineans and Australians respectively were not identified by the 1000GP. This study also validated that populations from southern Africa contain the highest genetic diversity amongst modern humans (Mallick et al. 2016). Addressing this diversity, the AGVP included whole-genome sequencing across individuals belonging to ten language subgroups in southern Africa, however, the low coverage sequencing (4 × coverage) risks misclassifying both observed and imputed rare variants (Gurdasani et al. 2015). Although efforts by multiple consortiums are currently expanding, the risk of eliminating genuine pathogenic variants that are segregating in the population will not be improved in the absence of comprehensive knowledge of human genetic architecture including rare variant frequencies.

Population-specific recombination maps

Recombination maps are often used for admixture mapping (Browning and Browning 2007). A recombination map is a genetic map that illustrates the variation of the recombination rate across a region of the genome or the entire genome (Myers et al. 2005). It is dependent on the underlying distribution of recombination events that occur between successive generations within a given population (Kong et al. 2010). The presence and activity of the PRDM9 zinc finger protein in the population under study, the ratio of males to females and the population's genetic substructure are some of the known factors that have an effect on these recombination events. Population substructure is affected by the migratory history, the evolutionary history and the common ancestry of the population (Manu et al. 2018). The extent to which the population substructure impacts the utility of a recombination map is yet to be determined.

Currently, there is a lack of high-resolution population-specific recombination maps for southern African populations. This has inevitably led to inaccuracies in studies that make use of a recombination map. These inaccuracies are exacerbated when no recombination maps for closely related populations are available. Research being done in southern African populations have thus been forced to make use of ancestral maps (such as European and West African maps) (Uren et al. 2017a) or they have to rely on ancestry informative markers to mitigate potential bias when genome-wide data is not available (Daya et al. 2013). There is thus a need for accurate, high-resolution recombination maps for southern African populations.

However, there are several uncertainties to be addressed before such a map can be handled with confidence. Firstly, the accuracy of the map needs to be established. Software used to construct recombination maps has been developed and tested on populations with homogeneous ancestry (Auton and McVean 2007). Secondly, testing the accuracy of a recombination map of an admixed population is difficult, because there are variable recombination rates between ancestries. Any one segment of a recombination map would have a recombination rate that closely resembles the average

rate of the rates of all the ancestries represented in the population. Thirdly, the method used to develop the map and the map itself would then have to be validated against currently available recombination maps (Kong et al. 2010). It should also be noted that the resolution of a given map relies strongly on the method used to construct the map and the number of individuals used to construct the map (Halldorsson et al. 2019). The most common methods used to build recombination maps are pedigree-based methods, LD-based methods and admixture based-inference (Halldorsson et al. 2019). Of these methods, the LD-based method produces the highest resolution if there are a limited number of individuals available. However, the pedigree-based and the admixture-based method can produce sub-kilobase resolutions when a few thousand individuals are available (Halldorsson et al. 2019). The problem with using the admixture-based method on a population for which no recombination map exists is that many methods that infer ancestry rely on a recombination map for the inference. Thus the resulting recombination map could be inaccurate because the map used for the ancestry inference might be based on a population that is distantly related to the population in question. When dealing with admixed populations, the pedigree-based method would produce the least amount of bias due to admixture, since the algorithms employed rely on direct observations of recombination events between parent–offspring pairs (Halldorsson et al. 2019). Because of the aforementioned reasons, the pedigree-based method should be the method of choice when a large enough sample from a population is available. The theoretical benefit of a population-specific recombination map has yet to be proven in practice, but one can expect such a map to improve the accuracy of admixture mapping and this improved accuracy could result in the discovery of novel variants associated with numerous phenotypes.

Future prospects for disease mapping

Novel loci identified in African populations

Novel genetic regions associated with multifactorial diseases could be identified by investigating the allelic architecture of highly admixed individuals from southern Africa, along with fine-mapping previous genomic loci associated with complex traits (Narang et al. 2011; Gurdasani et al. 2019). The first meta-analysis conducted in western Africa identified a novel locus (ZRANB3) significantly associated with type 2 diabetes (T2D) in a study investigating 5231 individuals from Nigeria, Kenya and Ghana. The study also indicated the transferability of 32 established T2D loci from previous investigations and contributed to the disease aetiology of T2D (Adeyemo et al. 2019). Furthermore, Gulsuner et al. investigated 909 schizophrenia patients and 917 healthy controls from the Xhosa population of South Africa (residing mostly in the eastern cape of South Africa). Not only did they identify admixture between Bantu-speaking Africans and San individuals, but also identified more private damaging mutations in cases than in controls. Interestingly when the same analysis was replicated in a Swedish cohort, the Xhosa individuals generally had larger effect sizes than that of the Swedes (Gulsuner et al. 2020). Furthermore, a meta-analysis consisting of 14,100 African individuals concerning cardiometabolic traits, identified novel loci associated with lipid, blood cell, and also other traits that appear to be rare in populations from other parts of the world (Gurdasani et al. 2019). However, these are mostly concerning common genetic variants and not adequate to identify rare genetic variants.

High-throughput technologies, such as whole-exome sequencing (WES) and whole-genome sequencing (WGS), are required to locate rare population-specific variants (Uren et al. 2017a; Retshabile et al. 2018; De La Vega and Bustamante 2018). Although WES is a cost-effective approach for identifying coding sequence targets in resource-restricted settings, WGS includes the complete and unbiased information carried by an individual, and high coverage WGS can detect rare variants (Suwinski et al. 2019). The first deep sequencing experiment of southern African populations assessed the population substructure within a cohort of HIV positive children from Botswana. WES data of 164 individuals from Botswana were analysed and compared with 150 similarly sequenced HIV positive Ugandan children (Retshabile et al. 2018). Approximately 13–25% of genetic variation in populations from Botswana was not captured in current public databases. These missing variants were significantly enriched for coding variants with MAF between 1% and 5% and included predicted-damaging non-synonymous variants. This population also had more rare (< 1%) pathogenic and damaging variants (Retshabile et al. 2018). These studies highlight the untapped potential of these populations to contribute to the novel discovery of disease risk alleles in GWAS studies. Extending GWAS and sequencing studies to diverse populations will surely generate a rich harvest of novel risk alleles.

Population-specific allele frequency databases

Population-specific allele frequencies have been sparsely characterised for southern African populations. For rare disease genetics, reference databases are continuously used for filtering based on allele frequency with the idea that common alleles are unlikely to be responsible for rare, highly penetrant disorders (Visscher et al. 2017). Therefore, in the absence of appropriate population reference datasets, variants can be misclassified and may lead to false disease associations. For instance, a major allele for southern African

populations can be identified as minor, since the current reference genome indicates it is a minor allele (Yang et al. 2019). High-coverage whole-genome reference datasets are needed to characterize and catalog population-specific variation and facilitate genetic studies in admixed southern African populations to identify causal rare variants.

The clinical value contributed by the deep sequencing of whole genomes was demonstrated by The GenomeAsia 100 K project (GAsP) (Wall et al. 2019). The pilot phase, which included a WGS dataset of 1739 individuals from 219 populations and 64 countries across Asia, identified a total of 194,585 novel variants with a MAF of > 1%. Overall 23% of protein-coding altering variants in GAsP were not found in publicly available databases such as the Single Nucleotide Polymorphism Database (dbSNP), the Genome Aggregation Database (gnomAD), the Exome Aggregation Consortium (ExAC) and the Exome Sequencing Project (ESP) (Wall et al. 2019). Importantly, imputation accuracy using the GAsP reference panel was 93- 95% compared to < 90% utilising the 1000GP reference panel. GAsP discovered thirteen unique cancer risk variants and HBB, a variant associated with beta-thalassemia. HBB is found almost exclusively in South Asians and at a lower frequency in Southeast Asia. Ultimately the GAsP reference dataset improves the ability to filter out low–probability candidates for highly penetrant disorders to identify putatively pathogenic variants that are found at high frequency in particular populations and improves the ability to infer pathogenicity of identified variants (Wall et al. 2019). Not only did this study exceed the ability of publicly available sources to annotate protein-coding variants and capture low-frequency rare variants unique to Asian populations, but it also improved the imputation of missing genotypes.

The Ugandan 2000 Genomes Project (UG2G) consists of 1978 individuals from rural Uganda and is the largest sequence panel from Africa (Gurdasani et al. 2019). The investigators identified 41.5 million SNPs and 4.5 million insertions and deletions. Likewise, 29% of the SNPs discovered in the UG2G project were absent in gnomAD. Furthermore, 52 population clusters in the region of Uganda (home to 9 ethnolinguistic groups) were identified and revealed a mixture of complex ancient East African pastoralists (Gurdasani et al. 2019). A genetic study conducted by Higasa et al., which included 1208 Japanese individuals identified 156,622 previously unreported variants. Surprisingly, the allele frequencies were lower than 0.5% and functional deleterious. This study specifically emphasized the importance of constructing an ethnicity-specific reference genome for identifying rare variants (Higasa et al. 2016).

An existing catalogue of known variants, be it common or rare, will allow researchers to identify mutations in protein-coding regions, rare causal variants and track the small and discrete mutations at a genomic level at multiple loci. However, population-specific variants will only be accurately collected if a reference genome exists with a representative population consensus, instead of using the existing human reference genome (GRCh38) of European ancestry (currently employed as a proxy in all genomic studies) (Ballouz et al. 2019).

Avenues for future improvements

Genomic resources lack southern African representation which is impacting on research in these settings. Future investigations to address this could include the following:

1. A consensus southern African reference genome, obtained from high-throughput whole genomes, for southern African populations, is required to capture the major alleles present in the region. This will serve as a genetic toolbox to improve imputation of missing genotypes to standardize cohorts genotyped on different arrays for meta-analysis and minimise the possibility of misclassifying major and minor alleles for southern African populations.
2. A southern African recombination map might improve the phasing of haplotypes to increase the accuracy of local ancestry inference in highly admixed individuals. However, there still exists some uncertainty in this regard and further investigations are warranted.
3. A southern African population-specific catalogue is required to capture allele frequencies in this region. Rare variants could be shared amongst healthy individuals, but not be present in public databases. Only high throughput sequencing technologies will be able to capture population-specific rare variants, since a reference genome, which is used as consensus in disease mapping, would not necessarily contain a specific variant.
4. An electronic catalogue of phenotypic information and the associated genotypic information enables geneticists to accurately identify genetic variants associated with disease phenotypes. However, the complexity of sample collection (due to unique ethical, cultural and socio-economic factors) in southern Africa is frequently underestimated as is reviewed elsewhere (Martin et al. 2018). The United Kingdom Biobank is a recent example of how incorporating clinical data embedded in electronic health records combined with GWAS data and registries available for research, can benefit everyone and not just individuals from a specific region.

Conclusion

Current GWAS and admixture mapping study designs are failing to identify disease-causing loci or rare genetic variants in southern African individuals. This is largely the result of limited reference haplotype panels and in turn limited genetic and computational tools available for southern African populations. The majority of SNP genotyping arrays are selected from a small sample of individuals (predominantly of European ancestry) and imputation and phasing of genotypic data usually involve a human reference of European ancestry, missing ± 10% of the genomes of individuals from African descent (De La Vega and Bustamante 2018). The genome structures of future generations might develop in a similar way to that of a complex five-way admixed southern African populations as admixture between populations originating from more than two different continents are now considered a customary feature of human populations across the globe (Busby et al. 2016; Salter-Townshend and Myers 2018).

Existing methods to detect loci associated with the multifactorial disease are not optimized for southern African ancestral groups and innovative approaches are urgently needed to study lethal communicable diseases such as TB, as well as non-communicable diseases such as cardio-metabolic diseases and type 2 diabetes in Africa. This entails the systematic development of best practises for ancestry inference, imputation and association studies. The establishment of a publicly available southern African-specific consensus reference genome is required to capture novel genetic variants and to maximize imputation for southern African populations. This will benefit future genetic studies involving complex diseases and traits by capturing rare variants previously lost due to a lack of publicly available data. Admixture mapping studies will continue to be inconclusive for populations from Southern Africa if no reference panels are available to represent proxy ancestral populations contributing to their genomes. The accuracy of the local ancestry calls for southern African individuals could also be decreased if population-specific recombination maps are not available. This will, in turn, affect the accuracy of admixture mapping studies that make use of LAI.

Conducting genetic studies on admixed southern African populations, with varying ancestral contributions, could also be beneficial for genetic studies of communicable and non-communicable diseases not mentioned in this review. Without a proper representative reference genome and methodologies to analyse complex admixed southern African genomes, genomic medicine will never benefit these individuals in contrast to those of European descent. For southern African countries and ethnicities to benefit from large-scale GWAS, as most European countries have, disease variants associated with southern African-specific diseases have to be identified. This will allow precision medicine and polygenic risk scores to be implemented. Although several consortiums contributed immensely to the development and training of African genetic researchers to include more diverse populations that have traditionally been underrepresented, global collaboration is still essential to increase the genetic representation of southern African populations.

Funding 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 represent the official views of the SAMRC.

Compliance with ethical standards

Conflict of interest All authors declare no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR (2015) A global reference for human genetic variation. Nature 526:68–74

Adeyemo AA, Zaghloul NA, Chen G, Doumatey AP, Leitch CC, Hostelley TL, Nesmith JE, Zhou J, Bentley AR, Shriner D, Fasanmade O, Okafor G, Eghan B, Agyenim-Boateng K, Chandrasekharappa S, Adeleye J, Balogun W, Owusu S, Amoah A, Acheampong J, Johnson T, Oli J, Adebamowo C, Collins F, Dunston G, Rotimi CN (2019) *ZRANB3* is an African-specific type 2 diabetes locus associated with beta-cell mass and insulin response. Nat Commun 10:1–12

Auton A, McVean G (2007) Recombination rate estimation in the presence of hotspots. Genome Res 17:1219–1227

Ballouz S, Dobin A, Gillis JA (2019) Is it time to change the reference genome? Genome Biol 20:1–9

Barnes KC, Grant AV, Hansel NN, Gao P, Dunston GM (2007) African Americans with asthma: genetic insights. Proc Am Thorac Soc 4:58–68

Beltrame MH, Rubel MA, Tishkoff SA (2016) Inferences of African evolutionary history from genomic data. Curr Opin Genet Dev 41:159–166

Bostoen K (2018) The Bantu expansion. Oxford University Press, Oxford, UK

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084–1097

Busby GB, Band G, Le Si Q, Jallow M, Bougama E, Mangano VD, Amenga-Etego LN, Enimil A, Apinjoh T, Ndila CM, Manjurano A, Nyirongo V, Doumba O, Rockett KA, Kwiatkowski DP, Spencer CC (2016) Admixture into and within sub-Saharan Africa. eLife 5:e15266. https://doi.org/10.7554/eLife.15266

Campbell MC, Tishkoff SA (2008) African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet 9:403–433

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22:231–238

Chimusa ER, Zaitlen N, Daya M, Möller M, van Helden PD, Mulder NJ, Price AL, Hoal EG (2014) Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. Hum Mol Genet 23:796–809

Chimusa ER, Defo J, Thami PK, Awany D, Mulisa DD, Allali I, Ghaza H, Moussa A, Mazandu GK (2018) Dating admixture events is unsolved problem in multi-way admixed populations. Brief Bioinform. https://doi.org/10.1093/bib/bby112

Chi C, Shao X, Rhead B, Gonzales E, Smith JB, Xiang AH, Graves J, Waldman A, Lotze T, Schreiner T, Weinstock-Guttman B, Aaen G, Tillema J-M, Ness J, Candee M, Krupp L, Gorman M, Benson L, Chitnis T, Mar S, Belman A, Casper TC, Rose J, Moodley M, Rensel M, Rodriguez M, Greenberg B, Kahn L, Rubin J, Schaefer C, Waubant E, Langer-Gould A, Barcellos LF (2019) Admixture mapping reveals evidence of differential multiple sclerosis risk by genetic ancestry. PLoS Genet 15:e1007808. https://doi.org/10.1371/journal.pgen.1007808

Choudhury A, Aron S, Sengupta D, Hazelhurst S, Ramsay M (2018) African genetic diversity provides novel insights into evolutionary history and local adaptations. Hum Mol Genet 27:R209–R218

Daya M, van der Merwe L, Galal U, Möller M, Salie M, Chimusa ER, Galanter JM, van Helden PD, Henn BM, Gignoux CR, Hoal E (2013) A panel of ancestry informative markers for the complex five-way admixed South African coloured population. PLoS ONE 8:e82224. https://doi.org/10.1371/journal.pone.0082224

Daya M, van der Merwe L, van Helden PD, Möller M, Hoal EG (2014b) The role of ancestry in TB susceptibility of an admixed South African population. Tuberc Edinb Scotl 94:413–420

Daya M, van der Merwe L, Gignoux CR, van Helden PD, Möller M, Hoal EG (2014a) Using multi-way admixture mapping to elucidate TB susceptibility in the South African Coloured population. BMC Genomics 15:1021. https://doi.org/10.1186/1471-2164-15-1021

De La Vega FM, Bustamante CD (2018) Polygenic risk scores: a biased prediction? Genome Med 10:100. https://doi.org/10.1186/s13073-018-0610-x

de Wit E, Delport W, Rugamika CE, Meintjes A, Möller M, van Helden PD, Seoighe C, Hoal EG (2010) Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. Hum Genet 128:145–153

Dias-Alves T, Mairal J, Blum MGB (2018) Loter: a software package to infer local ancestry for a wide range of species. Mol Biol Evol 35:2318–2326

Duan Q, Xu Z, Raffield LM, Chang S, Wu D, Lange EM, Reiner AP, Li Y (2018) A robust and powerful two-step testing procedure for local ancestry adjusted allelic association analysis in admixed populations. Genet Epidemiol 42:288–302. https://doi.org/10.1002/gepi.22104

Eisfeldt J, Mårtensson G, Ameur A, Nilsson D, Lindstrand A (2020) Discovery of novel sequences in 1,000 swedish genomes. Mol Biol Evol 37:18–30. https://doi.org/10.1093/molbev/msz176

Fan S, Kelly DE, Beltrame MH, Hansen MEB, Mallick S, Ranciaro A, Hirbo J, Thompson S, Beggs W, Nyambo T, Omar SA, Meskel DW, Belay G, Froment A, Patterson N, Reich D, Tishkoff SA (2019) African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. Genome Biol 20:82. https://doi.org/10.1186/s13059-019-1679-2

Fortes-Lima C, Gessain A, Ruiz-Linares A, Bortolini M-C, Migot-Nabias F, Bellis G, Moreno-Mayar JV, Restrepo BN, Rojas W, Avendaño-Tamayo E, Bedoya G, Orlando L, Salas A, Helgason A, Gilbert MTP, Sikora M, Schroeder H, Dugoujon J-M (2017) Genome-wide ancestry and demographic history of african-descendant maroon communities from French Guiana and Suriname. Am J Hum Genet 101:725–736

Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, Waliszewska A, Penney K, Steen RG, Ardlie K, John EM, Oakley-Girvan I, Whittemore AS, Cooney KA, Ingles SA, Altshuler D, Henderson BE, Reich D (2006) Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. Proc Natl Acad Sci U S A 103:14068–14073

Gignoux CR, Torgerson DG, Pino-Yanes M, Uricchio LH, Galanter J, Roth LA, Eng C, Hu D, Nguyen EA, Huntsman S, Mathias RA, Kumar R, Rodriguez-Santana J, Thakur N, Oh SS, McGarry M, Moreno-Estrada A, Sandoval K, Winkler CA, Seibold MA, Padhukasahasram B, Conti DV, Farber HJ, Avila P, Brigino-Buenaventura E, Lenoir M, Meade K, Serebrisky D, Borrell LN, Rodriguez-Cintron W, Thyne S, Joubert BR, Romieu I, Levin AM, Sienra-Monge J-J, Del Rio-Navarro BE, Gan W, Raby BA, Weiss ST, Bleecker E, Meyers DA, Martinez FJ, Gauderman WJ, Gilliland F, London SJ, Bustamante CD, Nicolae DL, Ober C, Sen S, Barnes K, Williams LK, Hernandez RD, Burchard EG (2019) An admixture mapping meta-analysis implicates genetic variation at 18q21 with asthma susceptibility in Latinos. J Allergy Clin Immunol. 143:957–969

González-Santos M, Montinaro F, Oosthuizen O, Oosthuizen E, Busby GBJ, Anagnostou P, Destro-Bisol G, Pascali V, Capelli C (2015) Genome-wide SNP analysis of Southern African populations provides new insights into the dispersal of Bantu-speaking groups. Genome Biol Evol 7:2560–2568

Gulsuner S, Stein DJ, Susser ES, Sibeko G, Pretorius A, Walsh T, Majara L, Mndini MM, Mqulwana SG, Ntola OA, Casadei S, Ngqengelele LL, Korchina V, van der Merwe C, Malan M, Fader KM, Feng M, Willoughby E, Muzny D, Baldinger A, Andrews HF, Gur RC, Gibbs RA, Zingela Z, Nagdee M, Ramesar RS, King M-C, McClellan JM (2020) Genetics of schizophrenia in the South African Xhosa. Science 367:569–573

Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, Karthikeyan S, Iles L, Pollard MO, Choudhury A, Ritchie GRS, Xue Y, Asimit J, Nsubuga RN, Young EH, Pomilla C, Kivinen K, Rockett K, Kamali A, Doumatey AP, Asiki G, Seeley J, Sisay-Joof F, Jallow M, Tollman S, Mekonnen E, Ekong R, Oljira T, Bradman N, Bojang K, Ramsay M, Adeyemo A, Bekele E, Motala A, Norris SA, Pirie F, Kaleebu P, Kwiatkowski D, Tyler-Smith C, Rotimi C, Zeggini E, Sandhu MS (2015) The African Genome Variation Project shapes medical genetics in Africa. Nature 517:327–332

Gurdasani D, Carstensen T, Fatumo S, Chen G, Franklin CS, Prado-Martinez J, Bouman H, Abascal F, Haber M, Tachmazidou I, Mathieson I, Ekoru K, DeGorter MK, Nsubuga RN, Finan C, Wheeler E, Chen L, Cooper DN, Schiffels S, Chen Y, Ritchie GRS, Pollard MO, Fortune MD, Mentzer AJ, Garrison E, Bergström A, Hatzikotoulas K, Adeyemo A, Doumatey A, Elding H, Wain LV, Ehret G, Auer PL, Kooperberg CL, Reiner AP, Franceschini N, Maher D, Montgomery SB, Kadie C, Widmer C, Xue Y, Seeley J, Asiki G, Kamali A, Young EH, Pomilla C, Soranzo N, Zeggini E, Pirie F, Morris AP, Heckerman D, Tyler-Smith C, Motala AA, Rotimi C, Kaleebu P, Barroso I, Sandhu MS (2019) Uganda genome resource enables insights into population history and genomic discovery in Africa. Cell 179:984-1002.e36. https://doi.org/10.1016/j.cell.2019.10.004

Halldorsson BV, Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP, Gunnarsson B, Oddsson A, Halldorsson GH, Zink F, Gudjonsson SA, Frigge ML, Thorleifsson G, Sigurdsson A,

Stacey SN, Sulem P, Masson G, Helgason A, Gudbjartsson DF, Thorsteinsdottir U, Stefansson K (2019) Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science 363:eaaw1043. https://doi.org/10.1126/science.aaw8705

Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet 22:239–247

Hellwege J, Keaton J, Giri A, Gao X, Velez Edwards DR, Edwards TL (2017) Population stratification in genetic association studies. Curr Protoc Hum Genet 95:1.22.1–1.22.23

Higasa K, Miyake N, Yoshimura J, Okamura K, Niihori T, Saitsu H, Doi K, Shimizu M, Nakabayashi K, Aoki Y, Tsurusaki Y, Morishita S, Kawaguchi T, Migita O, Nakayama K, Nakashima M, Mitsui J, Narahara M, Hayashi K, Funayama R, Yamaguchi D, Ishiura H, Ko W-Y, Hata K, Nagashima T, Yamada R, Matsubara Y, Umezawa A, Tsuji S, Matsumoto N, Matsuda F (2016) Human genetic variation database, a reference database of genetic variations in the Japanese population. J Hum Genet 61:547–553

Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM (2004) Design and analysis of admixture mapping studies. Am J Hum Genet 74:965–978

Ikegawa S (2012) A Short History of the Genome-wide association study: where we were and where we are going. Genom Inf 10:220–225

Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Jonasdottir A, Gylfason A, Kristinsson KT, Gudjonsson SA, Frigge ML, Helgason A, Thorsteinsdottir U, Stefansson K (2010) Fine-scale recombination rate differences between sexes, populations and individuals. Nature 467:1099–1103

Leabman MK, Huang CC, DeYoung J, Carlson EJ, Taylor TR, de la Cruz M, Johns SJ, Stryke D, Kawamoto M, Urban TJ, Kroetz DL, Ferrin TE, Clark AG, Risch N, Herskowitz I, Giacomini KM, Pharmacogenetics of membrane transporters investigators (2003) Natural variation in human membrane transporter genes reveals evolutionary and functional constraints. Proc Natl Acad Sci USA 100:5896–5901

Li WH, Sadler LA (1991) Low nucleotide diversity in man. Genetics 129:513–523

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, Skoglund P, Lazaridis I, Sankararaman S, Fu Q, Rohland N, Renaud G, Erlich Y, Willems T, Gallo C, Spence JP, Song YS, Poletti G, Balloux F, van Driem G, de Knijff P, Romero IG, Jha AR, Behar DM, Bravi CM, Capelli C, Hervig T, Moreno-Estrada A, Posukh OL, Balanovska E, Balanovsky O, Karachanak-Yankova S, Sahakyan H, Toncheva D, Yepiskoposyan L, Tyler-Smith C, Xue Y, Abdullah MS, Ruiz-Linares A, Beall CM, Di Rienzo A, Jeong C, Starikovskaya EB, Metspalu E, Parik J, Villems R, Henn BM, Hodoglugil U, Mahley R, Sajantila A, Stamatoyannopoulos G, Wee JTS, Khusainova R, Khusnutdinova E, Litvinov S, Ayodo G, Comas D, Hammer MF, Kivisild T, Klitz W, Winkler CA, Labuda D, Bamshad M, Jorde LB, Tishkoff SA, Watkins WS, Metspalu M, Dryomov S, Sukernik R, Singh L, Thangaraj K, Pääbo S, Kelso J, Patterson N, Reich D (2016) The Simons Genome diversity project: 300 genomes from 142 diverse populations. Nature 538:201–206

Manu S, Acharya KK, Thiyagarajan S (2018) Systematic analyses of autosomal recombination rates from the 1000 Genomes Project uncovers the global recombination landscape in humans. bioRxiv. https://doi.org/10.1101/246702

Martin AR, Teferra S, Möller M, Hoal EG, Daly MJ (2018) The critical needs and challenges for genetic architecture studies in Africa. Curr Opin Genet Dev 53:113–120

Mazandu GK, Geza E, Seuneu M, Chimusa ER (2019) Orienting future trends in local ancestry deconvolution models to optimally decipher admixed individual genome variations. Bioinform Tools Detect Clin Interpret Genom Var. https://doi.org/10.5772/intechopen.82764

Merker JD, Wenger AM, Sneddon T, Grove M, Zappala Z, Fresard L, Waggott D, Utiramerur S, Hou Y, Smith KS, Montgomery SB, Wheeler M, Buchan JG, Lambert CC, Eng KS, Hickey L, Korlach J, Ford J, Ashley EA (2018) Long-read genome sequencing identifies causal structural variation in a Mendelian disease. Genet Med Off J Am Coll Med Genet 20:159–163

Mulder N, Abimiku A, Adebamowo SN, de Vries J, Matimba A, Olowoyo P, Ramsay M, Skelton M, Stein DJ (2018) H3Africa: current perspectives. Pharmacogenom Pers Med 11:59–66

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310:321–324

Narang A, Jha P, Rawat V, Mukhopadhayay A, Dash D, Basu A, Mukerji M (2011) Recent admixture in an Indian Population of African Ancestry. Am J Hum Genet 89:111–120

Paşaniuc B, Kennedy J, Măndoiu I (2009) Imputation-Based Local Ancestry Inference in Admixed Populations. In: Măndoiu I, Narasimhan G, Zhang Y (eds) Bioinformatics research and applications. Springer, Berlin, pp 221–233

Patin E, Lopez M, Grollemund R, Verdu P, Harmant C, Quach H, Laval G, Perry GH, Barreiro LB, Froment A, Heyer E, Massougbodji A, Fortes-Lima C, Migot-Nabias F, Bellis G, Dugoujon J-M, Pereira JB, Fernandes V, Pereira L, der Veen LV, Mouguiama-Daouda P, Bustamante CD, Hombert J-M, Quintana-Murci L (2017) Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. Science 356:543–546

Pickrell JK, Patterson N, Barbieri C, Berthold F, Gerlach L, Güldemann T, Kure B, Mpoloka SW, Nakagawa H, Naumann C, Lipson M, Loh P-R, Lachance J, Mountain J, Bustamante CD, Berger B, Tishkoff SA, Henn BM, Stoneking M, Reich D, Pakendorf B (2012) The genetic prehistory of southern Africa. Nat Commun 3:1–6

Retshabile G, Mlotshwa BC, Williams L, Mwesigwa S, Mboowa G, Huang Z, Rustagi N, Swaminathan S, Katagirya E, Kyobe S, Wayengera M, Kisitu GP, Kateete DP, Wampande EM, Maplanka K, Kasvosve I, Pettitt ED, Matshaba M, Nsangi B, Marape M, Tsimako-Johnstone M, Brown CW, Yu F, Kekitiinwa A, Joloba M, Mpoloka SW, Mardon G, Anabwani G, Hanchard NA, Collaborative African Genomics Network (CAfGEN) of the H3Africa Consortium (2018) Whole-exome sequencing reveals uncaptured variation and distinct ancestry in the Southern African Population of Botswana. Am J Hum Genet 102:731–743

Rosenfeld JA, Mason CE, Smith TM (2012) Limitations of the Human Reference Genome for Personalized Genomics. PLoS One 7:e40294. https://doi.org/10.1371/journal.pone.0040294

Salter-Townshend M, Myers S (2018) Fine-scale inference of ancestry segments without prior knowledge of admixing groups. bioRxiv. https://doi.org/10.1101/376137

Schurz H, Müller SJ, van Helden PD, Tromp G, Hoal EG, Kinnear CJ, Möller M (2019) Evaluating the accuracy of imputation methods in a five-way admixed population. Front Genet 10:34. https://doi.org/10.3389/fgene.2019.00034

Secolin R, Mas-Sandoval A, Arauna LR, Torres FR, de Araujo TK, Santos ML, Rocha CS, Carvalho BS, Cendes F, Lopes-Cendes I, Comas D (2019) Distribution of local ancestry and evidence of adaptation in admixed populations. Sci Rep 9:1–12

Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, Boorgula MP, Chavan S, Vergara C, Ortega VE, Levin AM, Eng C, Yazdanbakhsh M, Wilson JG, Marrugo J, Lange LA, Williams LK, Watson H, Ware LB, Olopade CO, Olopade O, Oliveira RR, Ober C, Nicolae DL, Meyers DA, Mayorga A, Knight-Madden J, Hartert T, Hansel NN, Foreman MG, Ford JG, Faruque MU, Dunston GM, Caraballo L, Burchard EG, Bleecker ER, Araujo

MI, Herrera-Paz EF, Campbell M, Foster C, Taub MA, Beaty TH, Ruczinski I, Mathias RA, Barnes KC, Salzberg SL (2019) Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat Genet 51:30

Shriner D (2013) Overview of admixture mapping. Curr Protoc Hum Genet Chapter 1:Unit 1.23. https://doi.org/10.1002/0471142905.hg0123s76

Skotte L, Jørsboe E, Korneliussen TS, Moltke I, Albrechtsen A (2019) Ancestry-specific association mapping in admixed populations. Genet Epidemiol 43:506–521

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, Konkel MK, Malhotra A, Stütz AM, Shi X, Paolo Casale F, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJP, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HYK, Jasmine MuX, Alkan C, Antaki D, Bae T, Cerveira E, Chines P, Chong Z, Clarke L, Dal E, Ding L, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Y, Lameijer E-W, McCarthy S, Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R, Shabalin AA, Untergasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T, Sebat J, Batzer MA, McCarroll SA, Mills RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korbel JO, The 1000 Genomes Project Consortium (2015) An integrated map of structural variation in 2,504 human genomes. Nature 526:75–81

Suwinski P, Ong C, Ling MHT, Poh YM, Khan AM, Ong HS (2019) Advancing personalized medicine through the application of whole exome sequencing and big data analytics. Front Genet 10:49. https://doi.org/10.3389/fgene.2019.00049

Thornton TA, Bermejo JL (2014) Local and global ancestry inference, and applications to genetic association analysis for admixed populations. Genet Epidemiol 38:S5–S12. https://doi.org/10.1002/gepi.21819

Tucci S, Akey JM (2019) The long walk to African genomics. Genome Biol 20:130. https://doi.org/10.1186/s13059-019-1740-1

Uren C, Kim M, Martin AR, Bobo D, Gignoux CR, van Helden PD, Möller M, Hoal EG, Henn BM (2016) Fine-scale human population structure in Southern Africa reflects ecogeographic boundaries. Genetics 204:303–314

Uren C, Möller M, van Helden PD, Henn BM, Hoal EG (2017b) Population structure and infectious disease risk in southern Africa. Mol Genet Genom MGG 292:499–509

Uren C, Henn BM, Franke A, Wittig M, van Helden PD, Hoal EG, Möller M (2017a) A post-GWAS analysis of predicted regulatory variants and tuberculosis susceptibility. PLoS One 12:e0174738. https://doi.org/10.1371/journal.pone.0174738

Vergara C, Parker MM, Franco L, Cho MH, Valencia-Duarte AV, Beaty TH, Duggal P (2018) Genotype imputation performance of three reference panels using African ancestry individuals. Hum Genet 137:431–436

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J (2017) 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet 10:5–22

Wall JD, Stawiski EW, Ratan A, Kim HL, Kim C, Gupta R, Suryamohan K, Gusareva ES, Purbojati RW, Bhangale T, Stepanov V, Kharkov V, Schröder MS, Ramprasad V, Tom J, Durinck S, Bei Q, Li J, Guillory J, Phalke S, Basu A, Stinson J, Nair S, Malaichamy S, Biswas NK, Chambers JC, Cheng KC, George JT, Khor SS, Kim J-I, Cho B, Menon R, Sattibabu T, Bassi A, Deshmukh M, Verma A, Gopalan V, Shin J-Y, Pratapneni M, Santhosh S, Tokunaga K, Md-Zain BM, Chan KG, Parani M, Natarajan P, Hauser M, Allingham RR, Santiago-Turla C, Ghosh A, Gadde SGK, Fuchsberger C, Forer L, Schoenherr S, Sudoyo H, Lansing JS, Friedlaender J, Koki G, Cox MP, Hammer M, Karafet T, Ang KC, Mehdi SQ, Radha V, Mohan V, Majumder PP, Seshagiri S, Seo J-S, Schuster SC, Peterson AS, GenomeAsia100K Consortium (2019) The GenomeAsia 100K Project enables genetic discoveries across Asia. Nature 576:106–111

Wang H, Cade BE, Sofer T, Sands SA, Chen H, Browning SR, Stilp AM, Louie TL, Thornton TA, Johnson WC, Below JE, Conomos MP, Evans DS, Gharib SA, Guo X, Wood AC, Mei H, Yaffe K, Loredo JS, Ramos AR, Barrett-Connor E, Ancoli-Israel S, Zee PC, Arens R, Shah NA, Taylor KD, Tranah GJ, Stone KL, Hanis CL, Wilson JG, Gottlieb DJ, Patel SR, Rice K, Post WS, Rotter JI, Sunyaev SR, Cai J, Lin X, Purcell SM, Laurie CC, Saxena R, Redline S, Zhu X (2019) Admixture mapping identifies novel loci for obstructive sleep apnea in Hispanic/Latino Americans. Hum Mol Genet 28:675–687

Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, Highland HM, Patel YM, Sorokin EP, Avery CL, Belbin GM, Bien SA, Cheng I, Cullina S, Hodonsky CJ, Hu Y, Huckins LM, Jeff J, Justice AE, Kocarnik JM, Lim U, Lin BM, Lu Y, Nelson SC, Park S-SL, Poisner H, Preuss MH, Richard MA, Schurmann C, Setiawan VW, Sockell A, Vahi K, Verbanck M, Vishnu A, Walker RW, Young KL, Zubair N, Acuña-Alonso V, Ambite JL, Barnes KC, Boerwinkle E, Bottinger EP, Bustamante CD, Caberto C, Canizales-Quinteros S, Conomos MP, Deelman E, Do R, Doheny K, Fernández-Rhodes L, Fornage M, Hailu B, Heiss G, Henn BM, Hindorff LA, Jackson RD, Laurie CA, Laurie CC, Li Y, Lin D-Y, Moreno-Estrada A, Nadkarni G, Norman PJ, Pooler LC, Reiner AP, Romm J, Sabatti C, Sandoval K, Sheng X, Stahl EA, Stram DO, Thornton TA, Wassel CL, Wilkens LR, Winkler CA, Yoneyama S, Buyske S, Haiman CA, Kooperberg C, Marchand LL, Loos RJF, Matise TC, North KE, Peters U, Kenny EE, Carlson CS (2019) Genetic analyses of diverse populations improves discovery for complex traits. Nature 570:514–518

Yang X, Lee W-P, Ye K, Lee C (2019) One reference genome is not enough. Genome Biol 20:104. https://doi.org/10.1186/s13059-019-1717-0

Zhang J, Stram DO (2014) The role of local ancestry adjustment in association studies using admixed populations. Genet Epidemiol 38:502–515

Zheng-Bradley X, Flicek P (2017) Applications of the 1000 Genomes Project resources. Brief Funct Genom 16:163–170

Zhu X, Wang H (2017) The analysis of ethnic mixtures. In: Elston RC (ed) Statistical human genetics: methods and protocols. Springer, New York, pp 505–525

Zhu X, Luke A, Cooper RS, Quertermous T, Hanis C, Mosley T, Gu CC, Tang H, Rao DC, Risch N, Weder A (2005) Admixture mapping for hypertension loci with genome-scan markers. Nat Genet 37:177–181

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