Uganda Genome Resource enables insights into population history and genomic discovery in Africa

Gurdasani D.*[1], Carstensen T.* , Fatumo S.* [1,2], Chen G.*[2], Franklin CS.*[1], Prado-Martinez J.* [1], Bouman H.* [1], Abascal F. [1] , Haber M. [1], Tachmazidou I. [1], Mathieson I.[3], Ekoru K.[1,4], DeGorter MK. [5], Nsubuga RN.[6], Finan C. [1], Wheeler E. [1], Chen L.[1], Cooper DN.[6], Schiffels S.[7], Chen Y. [1], Ritchie GRS. [1], Pollard MO. [1], Fortune MD. [1], Mentzer AJ.[8], Garrison E. [1], Bergström A.[1], Hatzikotoulas K.[1], Adebowale A. [4], Doumatey A. [4], Elding H. [1], Wain LV.[9],[10], Ehret G.[11,12], Auer PL.[13], Kooperberg CL.[14] , Reiner AP.[15,16], Franceschini N.[17], Maher DP.[6], Montgomery SB.[7,18], Kadie C.[19], Widmer C.[20], Xue Y.[1], Seeley J. [6,21], Asiki G. [6] , Kamali A. [6,19], Young EH. [1], Pomilla C. [1], Soranzo N. [1,22,23], Zeggini E. [1], Pirie F.[24], Morris AP.[25,10], Heckerman D.[20], Tyler-Smith C[1‡], Motala A.[25‡], Rotimi C[4‡], Kaleebu P. [‡6,21], Barroso I.[‡ 1], Sandhu MS.[1,26‡]

*joint authors
‡ equal contribution

[1] William Harvey Research Institute, Queen Mary's University of London, London, UK
[2] Center for Research on Genomics and Global Health, National Institute of Health, USA
[3] Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA
[4] Medical Research Council/Uganda Virus Research Institute (MRC/UVRI) Uganda Research Unit on AIDS, Uganda
[5] Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA
[6] Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, UK
[7] Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany
[8] The Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK
[9] Department of Health Sciences, University of Leicester, Leicester, UK
[10] National Institute for Health Research, Leicester Biomedical Research Centre, Glenfield Hospital, Leicester, UK
[11] McKusick-Nathans Institute of Genetic Medicine Johns Hopkins University School of Medicine Baltimore MD, USA
[12] Geneva University Hospitals, Rue Gabrielle-Perret-Gentil,41211 Genève 14, Switzerland
[13] Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee WI, USA
[14] Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle WA, USA
[15] Department of Epidemiology, University of Washington, Seattle WA, USA
[16] Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle WA, USA
[17] Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA
[18] Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA
[19] Microsoft Research, Redmond, CA, USA
[20] Microsoft Research, Los Angeles, CA, USA
[21] London School of Hygiene and Tropical Medicine, London, UK
[22] Department of Haematology, University of Cambridge, Cambridge, UK
[23] The National Institute for Health Research Blood and Transplant Unit (NIHR BTRU) in Donor Health and Genomics, University of Cambridge, Cambridge, UK
[24] Department of Diabetes and Endocrinology, University of KwaZulu-Natal, Durban, South Africa
[25] Department of Biostatistics, University of Liverpool, Liverpool, UK
[26] Department of Medicine, University of Cambridge, Cambridge, UK

Corresponding authors:
Manjinder Sandhu ms23@sanger.ac.uk
Inês Barroso ib1@sanger.ac.uk
Charles Rotimi rotimic@mail.nih.gov
Ayesha Motala MOTALA@ukzn.ac.za
Chris Tyler-Smith cts@sanger.ac.uk
Pontiano Kaleebu pontiano.kaleebu@mrcuganda.org

Abstract

Genomic studies in African populations provide unique opportunities to understand disease aetiology, human genetic diversity and population history in a regional and a global context. In the largest study of its kind to date, comprising genome-wide data from 6,400 individuals from rural Uganda, and including whole-genome sequences from 1,978 individuals, we find evidence of geographically correlated fine-scale population substructure. Historically, we show that the ancestry of modern Ugandans is best represented by a mixture of ancient East African pastoralist populations. We demonstrate the value of the largest sequence panel from Africa to date as a global resource for population genetics, imputation and understanding the mutational spectrum and its clinical relevance in African populations. Examining 34 cardiometabolic traits, we show systematic differences in trait heritability between European and African populations, probably reflecting the differential impact of genetic and environmental factors. In the first multi-trait pan-African GWAS in up to 14,126 individuals, we identify 10 novel loci associated with anthropometric, haematological, lipid and glycemic traits. Our findings suggest that several functionally important signals at known and novel loci may be driven by differentiated variants within and specific to Africa, highlighting the value of including diverse study populations in African GWAS. We provide a rich new genomic and phenotypic resource for researchers in Africa and globally.

Introduction

Africa is central to our understanding of human origins, genetic diversity and disease susceptibility.[1] The marked genomic diversity and allelic differentiation among populations in Africa, in combination with the substantially lower linkage disequilibrium (correlation) among genetic variants, has the potential to provide new opportunities to understand disease aetiology relevant to African populations but also globally.[1,2] Consequently, there is a clear scientific and public health need to develop large-scale efforts that examine disease susceptibility across diverse populations within the African continent. Such efforts will need to be fully integrated with research-capacity-building initiatives across the region.[3]

Countries in Africa are undergoing epidemiological transitions—with a high burden of endemic infectious disease and growing prevalence of non-communicable diseases.[4] Importantly, because of varying environments, population history, and adaptive evolution, the distribution of risk factors for a broad range of cardiometabolic and infectious diseases, and their individual contributions, may differ among populations globally.[5] Differences in allele frequencies among populations, due to either selection or genetic drift provide unique opportunities to identify novel disease susceptibility loci; highlighting the value of conducting such studies in African populations. However, while there has been a recent increase in genetic studies of cardiometabolic traits including African-Americans,[6,7] there have been relatively few investigations of population diversity or the genetic determinants of cardiometabolic or infectious traits and diseases across the continent.

To conduct genetic studies in diverse populations across Africa, appropriate study designs that account for population structure, admixture and genetic relatedness (overt and cryptic), as well as the development of genetic tools to capture variation in African genomes, are needed.[2] To leverage the

relative benefits of different strategies, we undertook a combined approach of genotyping and low coverage whole-genome sequencing (WGS) in a population-based study of 6,400 individuals from a geographically defined rural community in South-West Uganda (**Figure 1a, STAR Methods, Figure 1, Figure S1** and **Table S1.1**). We present data from 4,778 individuals with genotypes for ~2.2 million SNPs from the Uganda genome-wide association study (UGWAS) resource (**STAR Methods**), and sequence data (**STAR Methods, Table S1.1**) on up to 1,978 individuals including 41.5M SNPs and 4.5M indels (UG2G) (**Figure 3, Figure S1, Table S1.8, Table S4.1** and **STAR Methods**). Collectively, these data represent the Uganda Genome Resource (UGR). To enhance trait-associated locus discovery, we also include collective data on up to 14,126 individuals from across the African continent for genome wide association analysis (**STAR Methods**).

Using these resources, we conducted a series of analyses to: 1) understand the population structure, admixture and demographic history in a geographically-defined population from Uganda (**STAR Methods**); 2) describe the spectrum of disease-causing mutations in the UG2G cohort (**STAR Methods**); and 3) highlight the value of the UG2G sequence panel as an imputation resource (**STAR Methods**). 4) refine estimates of heritability of 34 complex traits, accounting for environmental correlation among individuals (**STAR Methods**); and 5) assess the spectrum of genetic variants associated with cardiometabolic and other complex traits in populations from sub-Saharan Africa (**STAR Methods**)**.** Importantly, the UGR was designed to help develop local resources for public health and genomic research, including building research capacity, training and collaboration across the region. We envisage that data from these studies will provide a global resource for researchers, as well as facilitate genetic studies in African populations.

Results

Population structure and demographic history in a rural Ugandan community

Uganda has a diverse and complex history of extensive historical migration from surrounding regions over several hundred years. Migration has included economic migration for labour, as well as migration due to conflict in surrounding regions. Uganda is home to several diverse ethno-linguistic groups. The Ganda ('Baganda') are most common ethno-linguistic group in central Uganda (previously the Kingdom of Buganda). This central region has also seen extensive migration from the surrounding regions of Rwanda, Burundi (formerly Ruanda-Urundi) and Tanzania (formerly the district of Tanganyika) (**Figure 1**)[8] identifying as the 'Banyarwanda', 'Barundi' and 'Batanzania', respectively.[8] More recent migration has occurred from Rwanda, due to displacement following conflict (identified as 'Rwandese Ugandans', distinct from the 'Banyarwanda'). In addition to migration from surrounding regions, there have been large movements of people within Uganda relating to economic incentives during the colonial era. These include the Bakiga from Kigezi (Kiga), the Banyankole (Nkole) and Bafumbira from Kisoro from south-western Uganda, and the Batooro (Toro), Basoga (Soga) from regions adjacent to central Uganda (**Figure 1**).[8] There are a number of other ethnic groups that have migrated to Buganda from adjoining areas of South Sudan (the Madi and Acholi), the Democratic Republic of Congo on the north-western Ugandan border, as well as the from the West Nile region of Uganda (the Lugbara and Alur), and are referred to as "West Nile" migrants.[8] These groups often speak Nilotic languages. In our cohort, these ethno-linguistic groups are collectively classified as 'Others', as their fine-scale ethno-linguistic group

information was not available for these individuals. In this study, ethnolinguistic groups are based on self-identification and should be considered as representing a broad construct that encompasses shared cultural heritage, ancestry, history, homeland, language or ideology.

We characterised genetic diversity and fine-scale structure among nine ethno-linguistic population groups from a geographically-defined rural community from the Kalungu district in South-West Uganda (**Figure 1**, **STAR Methods**). Principal components (PCs) 1 and 2 explained 0.3% and 0.1% of the genetic variation observed, respectively, with the cline along PC1 (**Figure S2**) being strongly correlated with with Eurasian admixture ($r=-0.98$, $p<2x10^{-16}$) as inferred from ADMIXTURE, K=4 (**Figure 2**). This was corroborated in principal component analysis of Ugandan ethno-linguistic groups in the context of global populations (**Figure S2**) and our fineSTRUCTURE[9] analysis (**Figure 1**). FineSTRUCTURE analysis of the co-ancestry matrix inferred from linked genetic variants showed evidence suggestive of population substructure (**Figure 1**, **Figure S2**, **Table S1.4** and **STAR Methods**) with PCs 1 and 2 explaining 11.9% and 3.5% of observed variation, respectively. Clines along fineSTRUCTURE PC1 and PC2 were highly correlated with Eurasian ($r=-0.90$) and East African Nilo-Saharan ancestry ($r=-0.98$) as delineated by ADMIXTURE, K=4, respectively (**Figure 1, STAR Methods**). Here, Nilo-Saharan ancestry is defined as the ancestral component in ADMIXTURE analysis that was most prominent among the Dinka. The PC2 cline representing Nilo-Saharan ancestry was seen predominantly among the ethnolinguistic group classified as 'Others', consistent with these representing ethno-linguistic groups that have migrated into Uganda from the North Western region along the Nile. This suggests that the largest proportion of variation among the cohort was possibly driven by Eurasian and East African Nilo-Saharan gene flow.

Using Procrustes analyses, we find that substructure among ethno-linguisitic groups in this rural Ugandan community is correlated with the historical geographical origins of these migrant populations (**Figure 1, Tables S2.2-2.4 and STAR Methods**). This suggests that in spite of extensive migration and mixture, substructure does exist among individuals in Uganda, and this substructure shows statistically significant correlation with the historical distribution of population groups across the region. We find no clear association with current geographical coordinates, consistent with extensive movement and mixing following migration within this region (**Table S2.1**). These findings are corroborated by fineSTRUCTURE tree inference from the co-ancestry matrix which also shows clade structure reflecting historical geographical regions from which these populations have migrated. Ethno-linguistic groups from the central region of Uganda (the Baganda, Basoga and Batooro), migrant populations from Rwanda, Burundi, Tanzania (Banyarwanda, Rwandese Ugandans, Barundi and Batanzania, respectively) and those from South-western Uganda (Bakiga, Banyankole and Bafumbira) form separate clades (**Figure 1 and STAR Methods**). This clade structure may potentially also reflect the different amounts of Eurasian admixture observed among these populations, as we discuss subsequently.

On unsupervised fineSTRUCTURE analysis, we identify 52 population clusters (**Figure 1e**). These clusters appear represent a combination of factors, including ethno-linguistic group, historical geographical context (**Figure 1d and 1e**), as well as proportion of Eurasian and Nilo-Saharan ancestry, as estimated by ADMIXTURE, K=4. No clear pattern was observed by current GPS coordinate (**Figure 1c**), consistent with Procrustes analysis.

Using QpAdm, we identify evidence for at least three distinct streams of ancestry across the Ugandan populations relative to outgroups through qpWave analysis (rank 2, p=0.02) (**Table S3.8 and STAR Methods**). On examining change in rank on removing populations one at a time, we find that the distinct streams of ancestry correspond well with the clade structure inferred in fineSTRUCTURE, and historical geographic origins of these groups. Specifically, we find that the rank of the matrix drops by one on excluding Rwandese_Ugandan, Banyarwanda, Bakiga, Banyankole, suggesting that these include a distinct source of ancestry potentially not present in other populations (**Table S3.8**). Another stream of ancestry appears to be contributed by Barundi, and Mutanzania, consistent with the tree structure inferred by fineSTRUCTURE (**Figure 1d**). Baganda, Basoga and Mutooro appear to be relatively homogeneous, with only a single source of ancestry inferred across these populations (**Table S3.8**).

Formal tests for admixture (f_3, f_4 tests, MALDER and GLOBETROTTER analyses);[10 6] consistently supported evidence for Eurasian-like gene flow in Uganda (**Figure S4**, **Tables S3.1-S3.12, STAR Methods**). Eurasian-like gene flow may be inferred by these tests if the source population has allele frequency spectra correlated with modern Eurasians. This does not in itself provide evidence for Eurasian back migration into East Africa. We evaluate the source of this ancestry further. The presence of Eurasian MT (K1a, R0a1a, N1a1a3, HV1b1a, I, J1d1a1, and W8) and Y chromosome (R1b and H) haplogroups within Uganda provide support for back-migration, as these haplotypes are thought to have arisen from out-of-Africa (**Figure S6**, **Table S3.7, and STAR Methods**).[11-13] In order to distinguish Eurasian gene flow from ancient structure within East Africa, we also assessed the the double conditioned site-frequency spectrum among Ugandans, with the sfs being conditioned on alleles being derived in a French sample, and ancestral in Yoruba (YRI). A non-linear L shaped sfs, enriched for rare derived alleles would be consistent with recent admixture, and not ancient substructure. Our results confirm an observed dcsfs enriched for rare derived alleles and consistent with Eurasian gene flow. On assessing the fit of simulated data under different parameters, and observed data, we find that gene flow from Eurasian populations into Ugandans is necessary to explain the observed frequency spectra (**STAR Methods, Figure S5, and Tables S3.3-S3.4**). Overall a the dual model of admixture (~7% admixture) and ancient structure outperformed other models, including a model of ancient structure alone ($p<0.005$) (**Table S3.4**). We note, however that it is possible that fine-scale geographical spatial structure among populations could also explain these findings.[14]

Using the Conditional Random Field model (CRF), we assessed the presence of Neanderthal haplotypes among Ugandans. As Neanderthal ancestry is restricted to populations outside Africa, any evidence of Neanderthal ancestry among Africans is likely to be due to Eurasian back migration. We show evidence of detectable Neanderthal ancestry in Uganda, providing support for Eurasian admixture resulting from back-to-Africa migration. (**STAR Methods**, **Table S3.5-3.6**). We first validated our approach by confirming enrichment of inferred Neanderthal sites within Eurasian segments in simulated data ($p<0.001$) (**Table S3.6**). We find that segments of inferred Neanderthal ancestry among Ugandans show high (95%) overlap with inferred Eurasian haplotype segments in the same individuals (as inferred by ChromoPainter[9]). On assessing the overlap of segments of Neanderthal ancestry with the known map of Neanderthal ancestry among Europeans and Asians in the 1000 Genomes project,[15] we find that 90% of segments identified as be Neanderthal in origin (permutation $p<0.001$), overlapped with known maps of Neanderthal introgression in European and Asian genomes, as defined in the 1000 Genomes

Project (**Table S3.6**)[15] Furthermore, in line with expectations, we also find evidence of significantly lower background selection in identified regions of Neanderthal ancestry relative to other regions (mean B scores 920 and 799, respectively, permutation p<0.003). Collectively, our analyses support the Eurasian back-migration into Uganda, consistent with previous work [16] [17-19](**Figure S4, STAR Methods, Tables S3.1-3.13**).

Consistent with the extensive history of migration into this region, unsupervised ADMIXTURE,[20] and GLOBETROTTER[21] analyses suggest that Ugandans are best represented by a mosaic of East African (Bantu, Nilo-Saharan, Afro-Asiatic and rf-HG) and Eurasian-like ancestral components among modern global human populations (**Figures 1 and 2, Figures S2-S4, Tables S2.1-2.4, Tables S3.1-3.2, STAR Methods).** These findings are in keeping with other recent studies among East African populations that have suggested modern East African populations have been subject to complex admixture events over the past 5000 years.[19,22] The proportion of Eurasian admixture appears to be lower in Baganda, Basoga and Batooro, (**Figure 1d**), suggesting that waves of admixture may have occurred with regional specificity within Uganda.

Analysis with MALDER also detects multiple complex admixture events, with the older events inferred as best represented by modern rf-HG-like and Eurasian-like ancestral components having occurred ~2000-4500 years ago, (**Figure S4**), and more recent Eurasian-like admixture ~7-11 generations ago, consistent with previous reports.[2] [23]. Given the relatively low proportion of rf-HG admixture inferred within Ugandans by ADMIXTURE, GLOBETROTTER, and fineSTRUCTURE analysis, we evaluated this further. ALDER suggests low levels of rf-like admixture in Baganda (lower bound 4.4%), consistent with previous reports,[23] and our results from ADMIXTURE and GLOBETROTTER analysis. Inference of rf-HG-like and Eurasian ancestry as primary sources of admixture by MALDER here is likely to reflect the known bias of the algorithm towards identifying source ancestral components that are more drifted, even if they contribute proportionately little to ancestry.[24]

Asymmetrical gene flow has previously been noted between rf-HGs and East Africans, with predominantly Bantu admixture inferred within regional rf-HGs. We recapitulate these findings, [23,25] confirming substantial Bantu admixture in rf-HG (Mbuti) dating to ~760 years ago in ALDER analysis (lower bound admixture 18%). Collectively, our findings suggests that assimilation of eastern rf-HG like ancestry into East African Bantu populatons may have occurred during early migrations as part of the Bantu expansion, as these populations expanded into this region.[2] The route through which this ancestry entered these populations is unclear, and may have involved gene flow between Bantu and possibly other regional pastoralist or HG populations. We explore this further by examining ancient East African populations as possible representative sources of ancestry among modern Ugandans.

QpAdm analysis examining possible sources of admixture in modern Ugandans suggests that among global modern and ancient populations, modern Ugandan populations are best represented by ancestral components relating to ancient East African pastoralist populations (Tanzania_Pemba_700BP, and Tanzania_Luxmanda_3000BP) (**STAR Methods, Tables S3.12-S3.13**). These ancient pastoralists have been shown to be represented by multiple ancestral components, including ancient hunter-gatherer (Mota) and Eurasian (Levant-like) ancestry,[26] suggesting that these ancestral components may have entered modern Ugandans proximately through ancient East African

pastoralists in the region (**STAR Methods**). Our primary results identify a single source of ancestry represented by Tanzania_Pemba_700BP in Baganda and Basoga, consistent with previous qpWave analyses (**Table S3.12**). Other populations can be modelled either as a mixture of Tanzania_Pemba_700BP and Tanzania_Luxmanda_3000BP, or as a mixture of Tanzania_Pemba_700BP and modern or ancient Eurasians. Eurasian admixture in Ugandans varies from 5.8-10.9% (**Table S3.12**). Consistent with qpWave results suggesting multiple streams of admixture within Uganda, we find that Banyarwanda and Rwandese Ugandans cannot be modelled by any combination of two or three-source populations, reflecting complex ancestry in these ethno-linguistic groups.

We also note that although Tanzania_Pemba_700BP has been shown to be represented well by Mende previously [26] (a finding we were able to recapitulate in our analyses), replacing Tanzania_Pemba_700BP with Mende as a source population for admixture into Uganda in our models (**Table S3.12**) results in a poor model fit ($p<0.01$ in all cases). Our findings also suggest that West African populations may not reliably represent Bantu ancestry in East African Bantu populations. In order to assess this, we examine the f4 statistic *f4*(chimp, Ancient South African; YRI/Mende, Uganda); we find asymmetry of Ugandan and West African populations relative to ancient South African Khoe-San, evidenced by statistically significantly positive f4 statistics. Recent evidence has suggested that West Africans may carry a differential contribution of ancestry from an ancient population basal to ancient South Africans, leading to different West African populations (e.g. YRI and Mende) being asymmetrically related to ancient South Africans.[26] In this context, the asymmetry observed between West and East Afrians relative to ancient Khoe-San may be due to lower or absent basal ancestry in East African bantu populations relative to West Africans (**STAR Methods, Tables S3.9-3.11**). Alternatively, this may also be explained by Hadza-like or Khoe-San related ancestry in modern Ugandans. Further evaluation and interpretation of these findings will require a wider sampling of ancient DNA samples from across Africa.

To investigate ancient population size changes and split events, we examined a Ugandan trio sequenced at high depth (30x) using MSMC2[27] (**STAR Methods, Tables S1.6-1.7 and Figure S7**). We find that the demographic history of Ugandans is broadly comparable to other Africans such as Yoruba and Luhya (LWK), with an estimated effective population size of ~20,000 over the past 10,000 years (**Table S1.7a and S1.7b**). However, recent changes in population size of Ugandans seem more similar to LWK, as compared with YRI, and are consistent with patterns described by Schiffels et al. for LWK in the recent past (<10,000 years).[27] Schiffels et al. observed a long 'hump' in ancestral population size extending back from 6,000 years ago to beyond 50,000 years ago; we see a similar pattern in Uganda, likely reflecting complex admixture in Uganda, with modern Ugandans being a mosaic of multiple structured popilations that were separated for several thousands of year, unril recent admixture due to the extensive migration into this region.

On examining cross-coalescence between Uganda, YRI and LWK, we find that Ugandan populations split from Yoruba, Nigeria (YRI) ~11,500 ya, with subsequent gene flow between Uganda and LWK in recent times (**Figure S7 and STAR Methods**). The Uganda-YRI divergence is older than the Bantu expansion, [28] and may reflect varying patterns of Eurasian, basal and regional admixture in East and West African populations. It also should be noted that these divergence times are lower bounds, and are likely to be affected by gene flow between these populations following divergence, as previously

documented. [27] We note that while our cross-coalescence results for Uganda-YRI from 1000 Genomes Project reference based phasing are more in line with trio based phasing, reference based phasing using Complete Genomics data is suggestive of more recent split times (Figure S7g), suggesting that results from reference based phasing with the 1000 Genomes Project dataset are likely to have been more reliable. This is also in line with previous reports that inaccuracies in statistical phasing can impact inferences of split times. (ref) Our results support the sequencing of trios in diverse population sets to maximise phasing accuracy, or alternatively using strategies that can greatly improve phasing accuracy, such as linked read sequencing,[29] optical nano-technology, or SMRT sequencing, as implemented with the PacBio platform.

We explored more recent population history by examining rare variant sharing between the Baganda and other populations; we examined variants occurring only twice in the entire dataset (designated f_2) (**Figure S3 and STAR Methods**). On assessing average f2 sharing on repeatedly subsampled random haplotypes (n=40) from each population, we see extensive sharing of f2 variants between Ugandan populations and other Niger-Congo language speaking populations in the 1000 Genomes Project from East and West Africa. We also see extensive sharing with European and Asian populations consistent with Eurasian gene flow into these populations (**Figure S3a**). Paradoxically, we see little sharing among Ugandan populations; however, it must be noted that this is likely to be a consequence of our ascertainment scheme, with f2 variants being rarer among the Ugandan populations, and therefore, less likely to be sampled in a random set of 40 haplotypes (**Figure S3a, STAR Methods**).

Dating haplotypes surrounding f_2 variants can provide important information about the interrelation among populations, including ancient and recent population divergence.[30] Using this approach, we observe a total of 12,477,686 f2 variants in our dataset belonging to 9,875,361 f2 haplotypes. Given our ascertainment of f2 variants in a sample size comprising largely Ugandans, we expect f2 variation within Ugandans to be more recent than within other populations; therefore, we decided only to focus on the relationship of f2 variation between Ugandan and other populations, as this is likely to be relatively unbiased. We find that f_2 variants shared between European and Ugandan populations are more recent than those shared between European and West African populations (median f_2 dates were ~19,500 ya for Baganda compared with ~51,000 ya for YRI). This finding is consistent with back migration[17] and Eurasian admixture in the Uganda populations;[2,18] however, this may also reflect bias due to ascertainment of f2 variants in a larger population of Ugandans, thereby resulting in f2 variation representing rarer, and therefore more recent variation. Examining Ugandan populations in the context of other African populations, we find that f_2 sharing between Ugandan populations and Ethiopians tend to be older (median f_2 dating was ~23,000 ya) than Ugandan-West African splits, probably reflecting a combination of deeper population splits between Bantu- and Afro-Asiatic-speaking groups, and relatively high Eurasian admixture in the Ethiopian populations. We also find evidence of very ancient divergence (with a median f_2 dating of ~29,000 ya) between Baganda and Zulu (**Figure S3**); this could reflect old f_2 sharing with highly divergent Khoe-San haplotypes present among Zulu and other Southern African populations.[2] Our large African sequence resource allows the first such examination of shared rare variation among populations, and highlights the complex demographic histories of populations in this region.

A whole genome resource for population and medical genetics

With the largest whole genome sequence dataset from Africa to date (**Figure 3** and **STAR Methods**), we present a unique resource representing the spectrum of human genetic diversity in East Africa, as well as a resource to facilitate medical genetics studies in the region.

As expected, and consistent with the out-of-Africa model, Africans carry higher levels of variation relative to other continental populations, the overwhelming majority being rare (**Figure 3, Table S4.1, and STAR Methods**). In line with these observations, African populations provide greater opportunities for variant discovery as a function of sample size (**Table S4.1** and **STAR Methods**). We find that despite higher sequencing coverage within UK10K, discovery of genetic variation with increases in sample sizes among the Ugandans is greater than with European individuals from UK10K, at least up to a sample size of 500, after which gains plateaued (**Table X**). We identify 9.5 M novel variants in the UG2G resource that are not present in the 1000 Genomes Project (1000G) Phase 3, African Genome Variation Project (AGVP) and UK10K reference panels (**Figure 3**), highlighting the importance of assessing diverse populations on a larger scale. Multi-allelic variants represented 0.87% of called SNPs.

The average number of variants/individual in UG2G was greater than variation per individual observed in the UK10K cohorts dataset (4,298,968 and 3,412,214 in UG2G and UK10K cohorts, respectively), consistent with African populations having greater genetic diversity (**Table S4.1**). We also note a much greater proportion of rare variants among Ugandans, when comparing with an equal number of European individuals from the 1000 Genomes Project Phase 3, which has comparable depth of coverage. The differences in site frequency spectrum observed are consistent with a historical population bottleneck in Europeans, and greater genetic diversity with enrichment of rare variation among African populations. Heterozygosity rates among Ugandans were comparable to other African populations, except Ethiopia, where heterozygosity was lower, consistent with high levels of Eurasian admixture in these popultaions.

We also explored the predicted functional consequences of variation in the UG2G population (**Figure 3, Table S4.1-4.3, Figure S8 and STAR Methods**). Consistent with overall diversity, UG2G participants carried more missense variants per individual compared with the UK10K population (12,198 and 10,153 variants/individual respectively) (**STAR Methods**). As with previous studies, we find that in spite of the lower absolute number of missense mutations (149,251 in UG2G, and 69,761 in UK10K ALSPAC) in Europeans, these form a higher relative proportion of total variation (0.4% and 0.5% in UG2G and UK10K, respectively, $p<2e-16$) among Europeans. (**STAR Methods**). For disease-causing mutations (DMs), as annotated by the HGMD (**Figure 3 and STAR Methods**), we identified a median of 29 DMs/individual in our cohort compared to 25 DMs/individual in UK10K, despite more extensive studies in European populations, and potentially biased ascertainment.[31] By contrast, in UG2G, we observed a median of 3 homozygous DMs/individual compared with to 4 homozygous DMs/individual in UK10K (**STAR Methods**) ($p<2x10^{-16}$). In contrast to the GoNL study, where more than half of the DM variants were common (>5% AF), the Ugandan population shows the opposite pattern, with DM variants predominantly being rare (MAF <0.5%) in our cohort (**Figure 3**). A total of 650 out of the 998 DM variants had a frequency lower than 0.5%, whereas only 47 were common (>5% AF) in the UG2G. These findings are consistent with previous reports that suggest a shift towards the higher frequency spectrum for deleterious variants in out-of-Africa populations. However, these differences to some extent may also represent ascertainment of DMs primarily in Europeans.

On examining the number of ClinVar mutations per individual (2015 Clinvar database) in UG2G compared with the UK10K ALSPAC, and 1000 Genomes Phase III African and European populations, we observed greater number of median alleles/individual in the African individuals (UG2G and 1000 Genomes Project Phase III African populations) compared to Europeans (UK10K ALSPAC and 1000 Genomes Project phase III) in spite of the higher coverage of the ALSPAC dataset compared to UG2G (**Table S4.2**). Our results do not support substantial ascertainment bias in either the HGMD or ClinVar database, in contrast with previous reports of ascertainment. [31] [32] On comparing results using an older version of the ClinVar database (2014 version), we find clear evidence of ascertainment bias in the older database, with a greater number of clinically significant disease alleles/individual among Europeans compared with Africans, as have been reported before (**Table S4.2**).[32] Our findings suggest that generation of larger scale sequence data in more diverse panels have contributed to reduction in ascertainment bias among mutation databases over time.

The distribution of the mutational spectrum in African and European populations is consistent with previous reports, [33,34] and the impact of differences in demographic history among these populations. [33] The higher burden of homozygous deleterious variation in Europeans is consistent with previous literature[35,36] ; resulting from a loss of rare alleles following a population bottleneck thereby leading to greater co-occurrence of these mutations in recessive form. [33] The differences observed are unikely to represent differences in efficiency of selection in European and African populations since the split, but rather non-selective demographic forces of drift and mutation in an expanding population after a bottleneck, as has been suggested previously.[33] The higher frequency of deleterious variation in European populations may also be related to ascertainment bias, with more common recessive variation in European populations more likely to be identified and catalogued.[37]

Allele frequency differences between populations along with clinical phenotype data may provide insights into the functional relevance of putative DMs. On assessing 38 DMs that were common in our cohort (MAF>5%) but rare or absent in the UK10K data (MAF<1%) (**Table S4.3**),[38] we identify established causal loci associated with haematological traits, such as the *G6PD* and sickle cell (HBB) variants, which are common in UG2G, but absent from the UK10K data, consistent with these loci being under positive or balancing selection and protective against malaria (**Table S4.3**).[39] However, we also demonstrate that several putative DMs associated that are common in UG2G but rare in UK10K, do not show strong evidence for association with relevant cardiometabolic or hematological traits (**Figure S8**). These include rs41264848 in the *LPA* region (p=0.40 for association with total cholesterol); rs36220239 in the *ADAMTS13 region*, (p=0.90 for association with platelet count); and rs115080759 in the *HNF1A* gene associated with MODY3 showing no association with HbA1C (p=0.20 in entire cohort, and p=0.29 when only including individuals >40 yrs age). Our results for rs115080759 are consistent with reports that suggest this variant is benign.[40] This emphasises the need to carefully and comprehensively evaluate the impact of putative functional or disease-causing mutations across global populations, because they may not have any clinical or biological relevance, or be readily transferable across populations.[31,41] The lack of strong associations between these DMs and phenotypes in our cohort indicate that they are unlikely to be causal for the associated traits or may have different or lower penetrance within African populations due to complex factors, including epistasis, or gene-environment interplay.

Finally, we assess the impact of the addition of the UG2G panel to existing reference panels on imputation accuracy among populations from sub-Saharan Africa (**Figure 4**). We show that addition of the UG2G panel to existing sequence panels with African haplotypes, such as the 1000G Phase 3, and AGVP (combined n=3,895), markedly improved imputation accuracy (r^2 increase by 0.08 (MAF<=0.01) and 0.04 (all MAF)) for rare and common variants in Ugandan populations (**Figure 4 and STAR methods**). Additionally, we observe a substantial increase in imputation accuracy across the allele frequency spectrum generally in East African populations, including Nilo-Saharan linguistic groups such as the Kalenjin (**Figure 4**), probably reflecting haplotype sharing across the region. The number of variants "successfully" imputed (info≥0.3) substantially increased using the UG2G panel in comparison with the 1000G Phase III and AGVP panels combined, with an additional 8M variants being successfully imputed in Baganda, and 1.5M additional variants successfully imputed among other East African populations (**Figure 4**). These analyses emphasise the importance of building regional sequence based resources to facilitating genetic studies in Africa, including alongside current initiatives such as the Haplotype Consortium[42].

Heritability of cardiometabolic traits in a rural Ugandan community

Narrow-sense heritability represents the fraction of phenotypic variation in a population that is due to additive genetic variation. As such, it represents an important metric determining the genetic basis of complex traits and diseases. There have been no comprehensive evaluations of heritability and the interrelation with environment among African populations. We, therefore, assessed heritability for 34 complex cardiometabolic traits using a mixed model approach that also models environmental correlation[43] (**Figure 5** and **STAR Methods**)**.**

Estimates of heritability corrected for environmental correlation varied from relatively modest (e.g. 10% for GGT, a liver biomarker) to 55% for traits such as mean platelet volume (MPV) (**Table S5.1**). (**Figure 5**,**STAR Methods, Table S5.1**) We find clear statistical differences in heritability estimates for several traits, compared to European populations (**Figure 5** and **Tables S5.2-5.4**). For example, the narrow-sense heritability for height was 49% in Ugandans, compared with estimates of 70-80% in European populations (p<0.0001); by contrast, the heritability estimates for LDL were statistically significantly higher in the Ugandan population (54% vs 20-43% in European studies, p<0.002) (**Figure 3, Tables S5.2-5.4 and STAR Methods**). We speculate that these differences may be due to varying patterns of genetic loci influencing these traits in European and African populations, or perhaps more plausibly due to a larger proportion of environmental variation explaining phenotypic variance. For example, malnutrition or nutritional deficits in rural African populations may attenuate the effects of genetic variance on height, whereas dietary consumption and obesogenic environments in European populations may reduce the impact of genetic factors on the variation in LDL levels.[44] We note, however, that lower estimates of heritability (e.g. for height) in the Ugandan cohort may also arise from differences in LD (lower LD with causal variants), lack of adjustment for shared environment in previous studies, or gene-evironment interactions. While we do not find statistically significant gene-environment interactions for height, we find evidence for statistical gene-environment interaction for waist-hip ratio, red blood cell distribution width (RDW) and haematocrit (permutation p=<0.0001). These statistical interactions may represent interplay between genetic factors and dietary factors, iron stores and nutritional status (**Table S5.1).** Reliable assessment of the interrelation between genetic and

environmental variation, including specific environmental indices, will require application of these methods in much larger-scale studies with relevant phenotypic information. Examining locus-specific heritability would complement direct assessments of population differences in heritability of population traits.

GWAS of cardiometabolic traits in African populations

To assess the spectrum of genetic variants associated with cardiometabolic traits in African populations, we performed a GWAS of 34 cardiometabolic traits in up to 14,126 individuals from across the African continent, including populations from Ghana, Kenya, Nigeria, South Africa and Uganda (**STAR Methods, Table 1,** and **Table S6.1-S6.12**). To maximise opportunities for genomic discovery, we meta-analysed GWAS data from all study populations imputed with the UG2G-1000GP3-AGV combined panel, using the Han-Eskin random-effect meta-analytic approach implemented in METASOFT[45] to allow for potential heterogeneity in allelic effects (**STAR Methods**). We first re-assessed thresholds for genome-wide statistical significance in African populations using several approaches[46-49] and found that a statistical threshold of $5.0x10^{-9}$ is more relevant in populations with high genetic diversity and relatively lower levels of LD (**Table S6.1 and STAR Methods**).

In our meta-analysis, we identified 43 distinct signals statistically significantly associated with at least one trait (**Table S6.2**). Following visual inspection of locusview plots, two association signals were excluded (**Figure S9g and h**) as likely to be artefactual. More than half of all remaining signals (23/41) were attributable to genetic variants specific to African populations or extremely rare in other populations (**Table S6.2, Table S6.3 and STAR Methods**). Among these, we identified ten distinct or secondary signals at previously identified loci (**Table 1**), of which nine were driven by genetic variants that were specific to Africa or extremely rare in other populations (**Table S6.2**).
We also identified ten association signals within novel loci. These novel signals included associations with anthropometric indices, lipid, heamatological and blood cell traits (**Table 1, Figure 6, Figure S9** and **Table S6.2**). Among these novel signals, three were noted to have been previously identified as associated with biologically related traits.

Our novel association signals included a functionally relevant association between a 3.8Kb deletion (-α3.7), known to cause alpha thalassemia, and total bilirubin levels at $p = 2x10^{-12}$ (**STAR Methods, Table 1** and **Figure 6**). The -α3.7 variant is thought to have risen to high frequencies in African populations in regions endemic for malaria by virtue of providing resistance to severe malaria.[50]

We also identified a novel association with BMI on chromosome 1 (p=2.8x-10) in the intergenic region between *PLD5* and *SDCCAG8* (**Table S6.2**). The *SDCCAG8* locus has been previously associated with extreme childhood obesity in Europeans.[51] Recent unpublished summary data from GIANT and UKBiobank suggests that this locus may be associated with BMI (peak SNP rs11807000, p=5.7e-11). Our peak SNP is not present in these data or in the GIANT summary data. However, the presence of a comparably statistically significant association at this locus in a relatively small study (with respect to the UK Biobank and GIANT meta-analysis which examined ~700K individuals) is interesting, and needs further exploration. We also identified a novel association signal for the SNP rs7798566 (RE2 p=3e-15) with BMI on chr 7 in the intergenic region within the *TAS2R* gene family (**Table S6.2**). The TAS2R family of genes are expressed within the GI tract, are involved in taste sensitivity bitter-tasting compounds,[52]

and regulation of thyroid activity. Both these loci showed significant statistical heterogeneity of effect, with the association being seen only within the AADM cohort. The heterogeneity of effect for the *SDCCAG8* locus among African cohorts, and European cohorts may point to differential effects in different environments or genetic backgrounds (epistasis), or differences in demographic make up of these studies. The significance of these novel discoveries will require further evaluation across diverse population groups.

Among haematological traits, we identified a novel association on chr 11 between the *PDHX* and *CD44* region with WBC count. *CD44* encodes a cell-surface protein that regulates neutrophil adhesion, migration and apoptosis,[53,54] among other functions (**Table S6.2** and **Figure 6**). We also identified a novel association between rs1347767, an Africa-specific (MAF=10%) variant, downstream to *R3HDM1* associated with neutrophil count (**Table S6.2**). While this locus has not been previously associated with neutrophil count, this region lies near the *LCT* locus, known to be associated with WBC count in an exome association study of African-Americans.[55] The association at this locus was noted to be dependent on ancestry at the *LCT* locus in this study, suggesting the association may be population specific.[55] We also observed an association of the SNP causing sickle cell anemia (rs334) with RDW within our analysis. Notably, this SNP has not been identified as associated with RDW in the UK Biobank analysis of ~171K individuals (p=0.006) highlighting the utility of examining diverse cohorts in identifying functionally important associations with disease.

Fine mapping with MANTRA resulted in narrow credible intervals for most traits with 16 of 41 distinct loci being mapped to a single SNP in the credible interval (**Table S6.4**).[56] We also resolved the previously identified association with HbA1c at the *ITFG3* locus to the $\alpha^{-3.7}$ thalassemia deletion, which explained 3% of variation in HbA1c levels. We note that both associations were driven primarily by the Ugandan cohort, and not observed within other cohorts, consistent with the higher allele frequency of the deletion observed in Ugandans and the endemicity of malaria within this region. Our findings recapitulate the need to more fully understand functional variation, including for heamoglobinopathies, that may explain a substantial proportion of variation in HbA1c in African populations. These factors may have a direct impact on the utility of using HbA1C as a clinical tool for detection and diagnosis of diabetes in Africa.[57]

Given the complex and regionally-specific genetic diversity within Africa, we assessed patterns of heterogeneity and transferability of association signals across the four cohorts; to inform the design of medical genetics studies as well as understand the utility of European-centric polygenic scores for risk prediction in African populations. While most known associations with data available in >1 cohort were transferable (had nominally statistically significant p values in two or more cohorts), we identified several known and functionally important loci – the *LIPC* locus associated with HDL, the *DARC* locus encoding the Duffy antigen associated with monocyte count, and the the $\alpha^{-3.7}$ thalassemia variant at the *HBA1/A2* locus associated with RBC count and HbA1c that only had statistical support from a single cohort. Limited transferability at some of these loci appears to reflect allele frequency differences among cohorts potentially related to positive selection relating to the endemicity of malaria in some geographical regions and not others (e.g. the *DARC* and *HBA1/A2 loci*).[58-60] However, lack of transferability for other loci (e.g. *LIPC*) where the candidate SNP is common across all cohorts may reflect several factors, including allelic heterogeneity (multiple distinct variants at loci) or gene-

environment interactions, and will need further investigation in large-scale studies of diverse African populations. Additionally, there were four associations at known loci where the association signal was driven by a single cohort due to population-specificity of the variant examined, or rarity of the variant in other cohorts (MAF<0.5%) (**Table S6.3**). These included the *GPT* locus associated with ALT, with variants driving the association specific to Uganda (no association was observed at this locus in other cohorts), and *TIMD4* locus associated with LDL and total cholesterol levels.

Expectedly, transferability was observed to be lower among novel association signals. Among nine novel associations with data in >1 cohort identified, 5 were noted to have support only from a single cohort (**Table S6.3**); among these was the functionally relevant the sickle cell locus associated with RDW, and the *SDCCAG8* previously associated with childhood obesity[51], associated with BMI in our data. While the reasons for specificity of some of the novel loci to a single cohort relate to allele frequency differences of variants among cohorts (e.g. for the sickle cell locus), reasons for specificity at other loci are less clear, and require further exploration.

To systematically examine differences in effect sizes across cohorts we examined statistical heterogeneity of effect at associated loci among studies. While most peak associated SNPs did not show evidence of statistically significant heterogeneity, we found strong evidence of statistical heterogeneity in regions around several peak SNPs within known and biologically important regions associated with total cholesterol, LDL(e.g. the *PCSK9* and the *APOE* regions), bilirubin (UGT1A3-9 genes), GGT (*GGT1* locus), MCHC (*HBA1/A2* locus), ALT levels (*GPT*) and neutrophil count (*DARC* locus) (**Table S6.2**). This heterogeneity was partly attributable to differences in LD structure around causal or peak variants across populations, or the presence of multiple distinct variants at loci—allelic heterogeneity (**Figure S9 and STAR Methods**). For example, joint and conditional analysis at the *UG1TA3-9* locus associated with bilirubin in UGR showed evidence for three distinct SNPs associated with total bilirubin in joint and conditional analysis in the UGR (**Figure S6c and Table S6.9**), suggesting that statistical heterogeneity at a locus can provide important information about the genetic architecture of traits. Using the same approach, we also identified three distinct association signals at the *GGT1* locus in UGR, (**Figure S6d and Table S6.10**), with differences in LD around these distinct signals potentially explaining the statistical heterogeneity observed within this locus between cohorts.

In addition to allelic heterogeneity representing multiple distinct associations at a given locus, we also identified loci where distinct associations were identified as driving the association signal with a given trait among different populations. One example of this is the GPT locus associated with ALT levels, where distinct population-specific variants drive the association in Africans and Europeans. [61] We also identified a distinct association with ALP levels at the known ALPL locus. Peak associated SNPs at this locus have been previously noted to be different across large studies of European, [62] Chinese[63] and Japanese[64] cohorts (Table S6.5); these peak SNPs were not in LD with the peak SNP in Uganda, suggesting that multiple signals may be driving these associations at the locus in different populations. An alternate explanation is that all these SNPs may be differentially tagging an as yet unidentified causal variant.

Collectively, our findings highlight the utility of genetic resources from diverse populations in novel discovery, especially for population-specific and low frequency association signals. In this context,

differences in frequencies of functional alleles, allelic heterogeneity and differences in LD structure provide unique opportunities for discovery and resolution of causal loci, and a better understanding of the genetic architecture of disease.

Discussion

Here we present, the largest whole-genome sequence dataset from an East African population to date, as well as a large genome-wide genotyped and pehnotyped dataset from the same population. We provide rich genomic resources for studies of human population history and GWAS, and a mechanism to evaluate the clinical relevance of genetic diversity both in African populations and globally.

We present evidence for fine-scale structure and admixture in this Ugandan population, reflecting complex ancient and recent population migrations and expansions in East Africa. Our findings highlight the need for larger-scale deep sequencing, including a systematic assessment of hunter-gatherer populations across Africa, to more fully understand the genetic history and diversity of Africa. Sequencing of DNA from ancient skeletal material across Africa will greatly facilitate such efforts[65]—allowing stronger inferences into the source of genetic diversity and population history in Africa and globally.

Accounting for environmental correlation, we describe statistical differences in heritability for traits between African and European populations; these may be suggestive of differences in the interplay between genetic and environmental effects on heritable traits, as well as the impact of differences in genetic architecture as a result of selection, drift and historical demographic events. Our findings re-iterate the dynamic and context-specific nature of heritability, potentially varying among populations, demographic factors and environmental exposures.[66]

Lastly, in a combined meta-analyses of pan-African cohorts from five different countries across African totaling 14,126 individuals, we present results from trait-association discovery efforts. Our identification of several novel susceptibility loci across a range of complex traits argues for scaling efforts in the region. The continental and population-specificity of a large proportion of these association signals suggests that inclusion of diverse populations across Africa in GWAS may have the greatest potential for discovery and refinement of novel loci. Collectively, these findings provide the first empirical evidence to support theoretical models that suggest that power for discovery increases in meta-analyses of ethnically diverse populations, specifically driven by increased detection of low frequency and population-specific novel associations.[67]

Given high genetic diversity, and regionally specific patterns of admixture, we highlight the need to design GWAS studies to leverage these differences in allele frequency spectrum, and LD patterns across the African cohorts, including the creation of more diverse African whole genomic resources. The differences in LD structure observed around peak association signals across African populations will facilitate the refinement of association signals, and help identify causal variants. With caveats for rare variant discovery in some scenarios (**STAR Methods**), our analyses emphasize the value of utilizing diverse populations across the region—to maximise opportunities for genomic discovery,[68] and replication particularly in the context of rare and population-specific associations. Furthermore,

understanding differences in heritability, and identifying the full spectrum of genetic variation associated with complex traits and diseases across Africa, will require much larger-scale prospective studies that should include rich genomic and phenotypic data for complex traits and diseases, as well as information on environmental factors. In these contexts, our results provide a framework for undertaking more extensive GWAS in populations from Africa. Our findings also emphasise the need to develop methods to understand and compare heritability across populations. Recently, methods have been developed to assess heritabilities from summary statistics from GWAS, accounting for LD structure[69]; however, these methods will need to be extended to studies of diverse admixed populations with significant tracts of admixture LD, and within populations with high levels of relatedness.

Since genetic diversity is greatest in African populations, including a substantial proportion of genetic variation that is continentally and regionally distinct, it will be critical to understand the functional and biological relevance of this diversity. Understanding the biological basis for population-specific association signals, as well as the impact and tranferability of putatively functional and disease causing mutations at the individual and population level, will require representative genomic resources. We emphasise the need for the parallel development of transcriptomic and cellular biological resources at the population level to better reflect global human diversity.[70]

STAR Methods

        - Diversity of the UGR
        - Variant discovery
        - UGR: Assessment as an imputation panel
    - Heritability of traits in the General Population Cohort (GPC)
        - Statistical Model
        - Heritability estimates for traits in the GPC
    - Genome-wide association study of 34 traits
        - Meta-analysis across cohorts to maximise discovery
        - Mixed model analysis of data
        - Meta-analysis: statistical methods
        - Genome-wide threshold for significance and defining significant loci
        - Derivation of a genome-wide significance threshold for GWAS in African populations
        - Conditional analysis and meta-analyses
        - Analyses of transferability
        - Results for discovery GWAS across 34 traits
        - Fine mapping with MANTRA
- **DATA AVAILABILITY**

Acknowledgements
This work was funded by the Wellcome Trust, The Wellcome Sanger Institute (WT098051), the UK Medical Research Council (G0901213-92157, G0801566, and MR/K013491/1), and the Medical Research Council/Uganda Virus Research Institute Uganda Research Unit on AIDS core funding. This work was funded in part by IAVI with the generous support of the United States Agency for International Development (USAID) and other donors. The full list of IAVI donors is available at http://www.iavi.org. The contents of this manuscript are the responsibility of IAVI and co-authors and do not necessarily reflect the views of USAID or the US Government. DG is funded by a UKRI HDR-UK Innovation Fellowship (reference number MR/S003711/1). We thank the African Partnership for Chronic Disease Research (APCDR) for providing a network to support this study as well as a repository for deposition of curated data. We thank all study participants who contributed to this study. We also acknowledge the National Institute for Health Research Cambridge Biomedical Research Centre. The authors wish to acknowledge the use of The Uganda Medical Informatics Centre (UMIC) compute cluster. Computational support from UMIC was made possible through funding from the Medical Research Council (MC_EX_MR/L016273/1). We acknowledge the Sanger core pipeline teams for their help with sequencing and mapping the whole genome sequence data. The authors acknowledge with thanks the participants in the AADM project, their families and their physicians. The study was supported in part by the Intramural Research Program of the National Institutes of Health in the Center for Research on Genomics and Global Health (CRGGH). The CRGGH is supported by the National Human Genome Research Institute (NHGRI), the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), the Center for Information Technology, and the Office of the Director at the

National Institutes of Health (1ZIAHG200362). NS's research is supported by the Wellcome Trust (Grant Codes WT098051 and WT091310), the EU FP7 (EPIGENESYS Grant Code 257082 and BLUEPRINT Grant Code HEALTH-F5-2011-282510) and the National Institute for Health Research Blood and Transplant Research Unit (NIHR BTRU) in Donor Health and Genomics at the University of Cambridge in partnership with NHS Blood and Transplant (NHSBT). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health or NHSBT. DG was funded by the MRC (MR/S003711/1). AJM was funded by the Wellcome Trust (WT106289). We acknowledge use of summary data from the Global Lipids Genomics Consortium (GLGC). [71]

We acknowledge the H3Africa Bioinformatics Network (H3ABioNet) Node, National Biotechnology Development Agency (NABDA), Federal Ministry of Science and Technology (FMST) Abuja, Nigeria for funding SF for his post-doctoral research. DNC wishes to acknowledge the financial support of Qiagen Inc through a License Agreement with Cardiff University. We also acknowledge the 1000 Genomes Project, UK10K, Simon's Foundation Genome Diversity Project and African Genome Variation Project (AGVP) for providing data resources that were used to contextualise the UG2G data. The GATK3 program was made available through the generosity of Medical and Population Genetics program at the Broad Institute, Inc. The research was partially supported by the NIHR Leicester Biomedical Research Centre; the views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. L.V. Wain holds a GSK/British Lung Foundation Chair in Respiratory Research.

Bibliography

1 Tishkoff, S. A. *et al.* The genetic structure and history of Africans and African Americans. *Science* **324**, 1035-1044, doi:10.1126/science.1172257 (2009).

2 Gurdasani D., C. T., Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, Karthikeyan S, Iles L, Pollard M O, Choudhury A, Ritchie G R S, Xue Y, Asimit J, Nsubuga R N, Young E H, Pomilla C, Kivinen K, Rockett K , Kamali A, Doumatey A P, 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 S A, Pirie F, Kaleebu P, Kwiatkowski D, Tyler-Smith C, Rotimi C, Zeggini E and Sandhu M S. The African Genome Variation Project shapes medical genetics in Africa. *Nature* (2014).

3 Consortium, H. A. Research capacity. Enabling the genomic revolution in Africa. *Science* **344**, 1346-1348, doi:10.1126/science.1251546 (2014).

4 Organisation, W. H. Health Transition. (2015).

5 Campbell, M. C. & Tishkoff, S. A. The evolution of human genetic and phenotypic variation in Africa. *Curr Biol* **20**, R166-173, doi:10.1016/j.cub.2009.11.050 (2010).

6 Peprah, E., Xu, H., Tekola-Ayele, F. & Royal, C. D. Genome-wide association studies in Africans and African Americans: expanding the framework of the genomics of human traits and disease. *Public Health Genomics* **18**, 40-51, doi:10.1159/000367962 (2015).
7 Lanktree, M. B. *et al.* Genetic meta-analysis of 15,901 African Americans identifies variation in EXOC3L1 is associated with HDL concentration. *J Lipid Res* **56**, 1781-1786, doi:10.1194/jlr.P059477 (2015).
8 Richards, A. I. *Economic development and tribal change: a study of immigrant labour in Buganda*. (W. Heffer and Sons, 1954).
9 Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet* **8**, e1002453, doi:10.1371/journal.pgen.1002453 (2012).
10 Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065-1093, doi:10.1534/genetics.112.145037 (2012).
11 Richards, M. *et al.* Tracing European founder lineages in the Near Eastern mtDNA pool. *American journal of human genetics* **67**, 1251-1276 (2000).
12 Soares, P. *et al.* The archaeogenetics of Europe. *Current biology : CB* **20**, R174-183, doi:10.1016/j.cub.2009.11.054 (2010).
13 Mishmar, D. *et al.* Natural selection shaped regional mtDNA variation in humans. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 171-176, doi:10.1073/pnas.0136972100 (2003).
14 Eriksson, A. & Manica, A. The doubly conditioned frequency spectrum does not distinguish between ancient population structure and hybridization. *Mol Biol Evol* **31**, 1618-1621, doi:10.1093/molbev/msu103 (2014).
15 Sankararaman, S. *et al.* The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354-357, doi:10.1038/nature12961 (2014).
16 Gallego Llorente, M. *et al.* Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. *Science* **350**, 820-822, doi:10.1126/science.aad2879 (2015).
17 Henn, B. M. *et al.* Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet* **8**, e1002397, doi:10.1371/journal.pgen.1002397 (2012).
18 Pickrell, J. K. *et al.* Ancient west Eurasian ancestry in southern and eastern Africa. *Proc Natl Acad Sci U S A* **111**, 2632-2637, doi:10.1073/pnas.1313787111 (2014).
19 Fan, S. *et al.* African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. *Genome Biol* **20**, 82, doi:10.1186/s13059-019-1679-2 (2019).
20 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-1664, doi:10.1101/gr.094052.109 (2009).
21 Hellenthal, G. *et al.* A genetic atlas of human admixture history. *Science* **343**, 747-751, doi:10.1126/science.1243518 (2014).
22 Scheinfeldt, L. B. *et al.* Genomic evidence for shared common ancestry of East African hunting-gathering populations and insights into local adaptation. *Proc Natl Acad Sci U S A*, doi:10.1073/pnas.1817678116 (2019).
23 Patin, E. *et al.* Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science* **356**, 543-546, doi:10.1126/science.aal1988 (2017).
24 Pickrell, J., Patterson N, Loh P, Lipson M, Berger B, Stoneking M, Pakendorf B, Reich D. Ancient west Eurasian ancestry in southern and eastern Africa. *unpublished* (2013).

25 Patin, E. *et al.* The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. *Nat Commun* **5**, 3163, doi:10.1038/ncomms4163 (2014).
26 Skoglund, P. *et al.* Reconstructing Prehistoric African Population Structure. *Cell* **171**, 59-71 e21, doi:10.1016/j.cell.2017.08.049 (2017).
27 Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**, 919-925, doi:10.1038/ng.3015 (2014).
28 de Filippo, C., Bostoen, K., Stoneking, M. & Pakendorf, B. Bringing together linguistic and genetic evidence to test the Bantu expansion. *Proc Biol Sci* **279**, 3256-3263, doi:10.1098/rspb.2012.0318 (2012).
29 Zheng, G. X. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* **34**, 303-311, doi:10.1038/nbt.3432 (2016).
30 Mathieson, I. & McVean, G. Demography and the age of rare variants. *PLoS Genet* **10**, e1004528, doi:10.1371/journal.pgen.1004528 (2014).
31 Xue, Y. *et al.* Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *Am J Hum Genet* **91**, 1022-1032, doi:10.1016/j.ajhg.2012.10.015 (2012).
32 Consortium, G. P. A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).
33 Do, R. *et al.* No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat Genet* **47**, 126-131, doi:10.1038/ng.3186 (2015).
34 Lohmueller, K. E. *et al.* Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**, 994-997, doi:10.1038/nature06611 (2008).
35 Lohmueller, K. E. The impact of population demography and selection on the genetic architecture of complex traits. *PLoS Genet* **10**, e1004379, doi:10.1371/journal.pgen.1004379 (2014).
36 Henn, B. M. *et al.* Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proc Natl Acad Sci U S A* **113**, E440-449, doi:10.1073/pnas.1510805112 (2016).
37 Amorim, C. E. G. *et al.* The population genetics of human disease: The case of recessive, lethal mutations. *PLoS Genet* **13**, e1006915, doi:10.1371/journal.pgen.1006915 (2017).
38 Consortium, U. K. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90, doi:10.1038/nature14962 (2015).
39 Karlsson, E. K., Kwiatkowski, D. P. & Sabeti, P. C. Natural selection and infectious disease in human populations. *Nat Rev Genet* **15**, 379-393, doi:10.1038/nrg3734 (2014).
40 .
41 Saraf, S. L. *et al.* Differences in the clinical and genotypic presentation of sickle cell disease around the world. *Paediatr Respir Rev* **15**, 4-12, doi:10.1016/j.prrv.2013.11.003 (2014).
42 McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*, doi:10.1038/ng.3643 (2016).
43 Heckerman, D. *et al.* Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proc Natl Acad Sci U S A* **113**, 7377-7382, doi:10.1073/pnas.1510497113 (2016).

44 Nalwoga, A. *et al.* Nutritional status of children living in a community with high HIV prevalence in rural Uganda: a cross-sectional population-based survey. *Trop Med Int Health* **15**, 414-422, doi:10.1111/j.1365-3156.2010.02476.x (2010).
45 Han, B. & Eskin, E. Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. *American Journal of Human Genetics* **88**, 586-598, doi:10.1016/j.ajhg.2011.04.014 (2011).
46 Gao, X., Starmer, J. & Martin, E. R. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet Epidemiol* **32**, 361-369, doi:10.1002/gepi.20310 (2008).
47 Chen, Z. & Liu, Q. A new approach to account for the correlations among single nucleotide polymorphisms in genome: wide association studies. *Hum Hered* **72**, 1-9, doi:10.1159/000330135 (2011).
48 Moskvina, V. & Schmidt, K. M. On multiple-testing correction in genome-wide association studies. *Genet Epidemiol* **32**, 567-573, doi:10.1002/gepi.20331 (2008).
49 Nyholt, D. R. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* **74**, 765-769, doi:10.1086/383251 (2004).
50 Mockenhaupt, F. P. *et al.* Alpha(+)-thalassemia protects African children from severe malaria. *Blood* **104**, 2003-2006, doi:10.1182/blood-2003-11-4090 (2004).
51 Scherag, A. *et al.* Two new Loci for body-weight regulation identified in a joint analysis of genome-wide association studies for early-onset extreme obesity in French and german study groups. *PLoS Genet* **6**, e1000916, doi:10.1371/journal.pgen.1000916 (2010).
52 Bachmanov, A. A. & Beauchamp, G. K. Taste receptor genes. *Annu Rev Nutr* **27**, 389-414, doi:10.1146/annurev.nutr.26.061505.111329 (2007).
53 Wang, Q., Teder, P., Judd, N. P., Noble, P. W. & Doerschuk, C. M. CD44 deficiency leads to enhanced neutrophil migration and lung injury in Escherichia coli pneumonia in mice. *Am J Pathol* **161**, 2219-2228, doi:10.1016/S0002-9440(10)64498-7 (2002).
54 Khan, A. I. *et al.* Role of CD44 and hyaluronan in neutrophil recruitment. *J Immunol* **173**, 7594-7601 (2004).
55 Auer, P. L. *et al.* 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, doi:10.1016/j.ajhg.2012.08.031 (2012).
56 Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714-719, doi:10.1038/nature09266 (2010).
57 Herman, W. H. & Cohen, R. M. Racial and ethnic differences in the relationship between HbA1c and blood glucose: implications for the diagnosis of diabetes. *J Clin Endocrinol Metab* **97**, 1067-1072, doi:10.1210/jc.2011-1894 (2012).
58 Liu, X. *et al.* Detecting and characterizing genomic signatures of positive selection in global populations. *Am J Hum Genet* **92**, 866-881, doi:10.1016/j.ajhg.2013.04.021 (2013).
59 Hedrick, P. W. Resistance to malaria in humans: the impact of strong, recent selection. *Malar J* **11**, 349, doi:10.1186/1475-2875-11-349 (2012).
60 Hamblin, M. T., Thompson, E. E. & Di Rienzo, A. Complex signatures of natural selection at the Duffy blood group locus. *Am J Hum Genet* **70**, 369-383, doi:10.1086/338628 (2002).
61 Abul-Husn, N. S. *et al.* A Protein-Truncating HSD17B13 Variant and Protection from Chronic Liver Disease. *N Engl J Med* **378**, 1096-1106, doi:10.1056/NEJMoa1712191 (2018).

62 Chambers, J. C. *et al.* Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nat Genet* **43**, 1131-1138, doi:10.1038/ng.970 (2011).
63 Yuan, X. *et al.* Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes. *Am J Hum Genet* **83**, 520-528, doi:10.1016/j.ajhg.2008.09.012 (2008).
64 Kamatani, Y. *et al.* Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat Genet* **42**, 210-215, doi:10.1038/ng.531 (2010).
65 Pickrell, J. K. & Reich, D. Toward a new history and geography of human genes informed by ancient DNA. *Trends Genet* **30**, 377-389, doi:10.1016/j.tig.2014.07.007 (2014).
66 Haworth, C. M. & Davis, O. S. From observational to dynamic genetics. *Front Genet* **5**, 6, doi:10.3389/fgene.2014.00006 (2014).
67 Pulit, S. L., Voight, B. F. & de Bakker, P. I. Multiethnic genetic association studies improve power for locus discovery. *PLoS One* **5**, e12600, doi:10.1371/journal.pone.0012600 (2010).
68 Cook, J. P. & Morris, A. P. Multi-ethnic genome-wide association study identifies novel locus for type 2 diabetes susceptibility. *Eur J Hum Genet* **24**, 1175-1180, doi:10.1038/ejhg.2016.17 (2016).
69 Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47**, 1228-+, doi:10.1038/ng.3404 (2015).
70 Chang, E. A. *et al.* Derivation of Ethnically Diverse Human Induced Pluripotent Stem Cell Lines. *Sci Rep* **5**, 15234, doi:10.1038/srep15234 (2015).
71 Consortium, G. L. G. Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**, 1274-1283, doi:10.1038/ng.2797 (2013).