

1           **Uganda Genome Resource: A rich research database for genomic**  
2           **studies of communicable and non-communicable diseases in Africa**

3  
4           Segun Fatumo<sup>1,2,3</sup>, Joseph Mugisha<sup>3</sup>, Opeyemi S Soremekun<sup>1,3</sup>, Allan Kalungi<sup>1,3</sup>, Richard Mayanja<sup>1,4</sup>,  
5           Christopher Kintu<sup>1,4</sup>, Ronald Makanga<sup>3</sup>, Ayoub Kakande<sup>3</sup>, Andrew Abaasa<sup>3</sup>, Gershim Asiki<sup>5</sup>, Robert  
6           Kalyesubula<sup>3,4</sup>, Robert Newton<sup>3</sup>, Moffat Nyirenda<sup>2,3</sup>, Manj S. Sandhu<sup>6</sup>, Pontiano Kaleebu<sup>2,3</sup>

7  
8           <sup>1</sup>*The African Computational Genomics (TACG) Research Group, MRC/UVRI and LSHTM,*  
9           Entebbe, Uganda

10          <sup>2</sup>London School of Hygiene and Tropical Medicine London, United Kingdom.

11          <sup>3</sup>Medical Research Council/ Uganda Virus Research Institute/London School of Hygiene and  
12          Tropical Medicine (MRC/UVRI/LSHTM) Uganda research unit, Entebbe, Uganda

13          <sup>4</sup>College of Health Sciences, Makerere University, Kampala Uganda

14          <sup>5</sup>Health and Systems for Health Research Unit, African Population and Health Research  
15          Center, Nairobi, Kenya.

16          <sup>6</sup>Department of Epidemiology and Biostatistics, School of Public Health, Imperial College  
17          London, United Kingdom.

18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30          **Correspondence**

31          Dr. Segun Fatumo  
32          *The African Computational Genomics (TACG) Research Group,*  
33          MRC/UVRI and LSHTM, Entebbe, Uganda  
34          [segun.fatumo@mrcuganda.org](mailto:segun.fatumo@mrcuganda.org)  
35          [segun.fatumo@lshtm.ac.uk](mailto:segun.fatumo@lshtm.ac.uk)  
36

37 **Abstract**

38 The Uganda Genome Resource (UGR) is a well characterised genomic database, with a range  
39 of phenotypic communicable and non-communicable diseases and risk factors generated  
40 from the Uganda General Population Cohort (GPC) - a population-based open cohort study  
41 established in 1989 by the Medical Research Council (MRC) UK in collaboration with the  
42 Uganda Virus Research Institute (UVRI).

43

44 In 2011, UGR was launched with genotype data on ~5000 and whole genome sequence data  
45 on ~2000 Ugandan individuals from 9 ethno-linguistic groups. Leveraging other available  
46 platforms at the MRC Uganda such as Biorepository centre for sample storage, Clinical  
47 Diagnostic Laboratory Service (CDLS) for sample diagnostic testing, sequencing platform  
48 for DNA extraction, Uganda Medical informatics Unit (UMIC) for large-scale data analysis,  
49 GPC for additional sample collection, UGR is strategically poised to expand and generate  
50 scientific discoveries.

51

52 Here, we describe UGR and highlight the important genetic findings thus far including how  
53 UGR is providing opportunities to: (1) discover novel disease susceptibility genetic loci; (2)  
54 refine association signals at new and existing loci; (3) develop and test Polygenic Risk Score  
55 (PRS) to determine individual's disease risk; 4) assess how some risk factors including  
56 infectious diseases are causally related to non-communicable diseases (NCDs) in Africa; (5)  
57 develop research capacity for genomics in Africa; and (6) enhance African participation in  
58 the global genomics research arena. Leveraging established research infrastructure, expertise,  
59 local genomic leadership, global collaboration and strategic funding, we anticipate that UGR  
60 can develop further to a comparable level of European and Asian large-scale genomic  
61 initiatives.

62

63

64

65

66

67 **Keywords:** Genomics, Uganda, NCDs, GPC

68

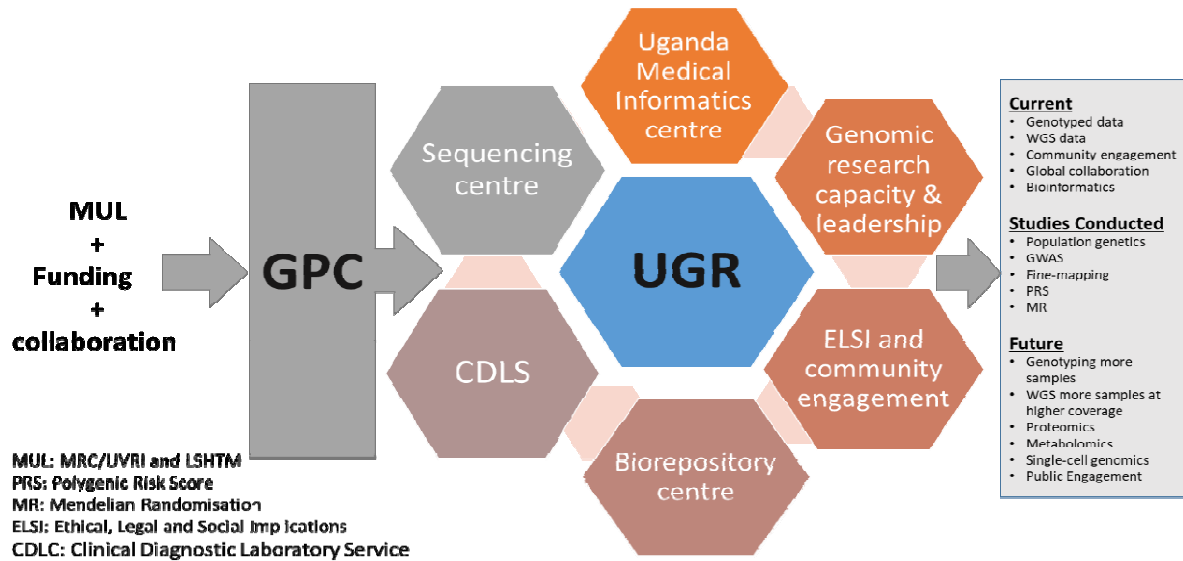
69 **Introduction**

70 The genetic diversity in Africa is far greater than in any other region across the globe but  
71 unfortunately, the vast majority of genomic studies have been performed in European  
72 ancestry populations (PMID: 35145307). Uganda is located in East Africa with four major  
73 ethnic groups and over 40 languages. The rich linguistic, ethnic, and cultural diversity of  
74 Uganda provides an unprecedented opportunity to understand the level of the genetic  
75 structure in Uganda populations. To advance genetic epidemiology of communicable and  
76 non-communicable diseases (NCDs) in Uganda, the Ugandan Genome Resource (UGR) was  
77 launched in 2011 to prospectively collect a wide range of NCDs, infectious disease risk  
78 factors including information on lifestyle, family history social determinant, demographics,  
79 sexual health & reproductive behaviour, past illness, mental health, treatment and  
80 immunisation and environmental risk factors (Asiki *et al.*, 2013).

81

82 Here, we provide a detailed description of UGR which is different from previous publications  
 83 on GPC that focused on specific aspects (Asiki *et al.*, 2013) or population genetics and  
 84 genome-wide association analyses of cardiometabolic traits in UGR data (Gurdasani *et al.*,  
 85 2019), we aim to give an overview of UGR as a resource including detailed phenotype  
 86 availability, genomic data generation, sample characteristics, genetic discoveries to date, and  
 87 finally to its data access and sharing policy.

### UGR's Strategic Design for Continuous Expansion and Maintenance



88 Figure 1: UGR strategic vision for continuous expansion and maintenance. UGR continues to  
 89 offer as a valuable platform for investigating the genetics of NCDs and relationship with  
 90 infectious diseases in Africa.

#### 91 Study Population – The General Population Cohort (GPC)

92 The GPC is a population-based study of approximately 22,000 individuals residing in 25  
 93 neighbouring villages in the Kyamulibwa sub-county, Kalungu district in rural south-western  
 94 Uganda. The study was founded in 1989 by the Medical Research Council UK (MRC UK) in  
 95 collaboration with the Uganda Virus Research Institute (UVRI) to study the epidemiology of  
 96 HIV in a general population. The GPC population was initially recruited and assessed  
 97 through annual house-to-house census and survey rounds until 2012, when biannual surveys  
 98 commenced. Since its establishment, 26 rounds of survey and 29 rounds of census have been  
 99 undertaken. Before any survey procedures are carried out, written informed consent is  
 100 obtained from participants on the use of their clinical records for research purposes and  
 101 sample storage for future use (Asiki *et al.*, 2013). Data collected includes serological,

102 demographic, and medical information from participants. Information regarding mortality,  
103 fertility, sexual behaviour migration, HIV infection perception are routinely collated.

104

105 The GPC Round 22 study of 2011 focused on the genetics and epidemiology of  
106 communicable and non-communicable disease. The survey round which was used to establish  
107 the Ugandan Genome Resource (UGR) consisted of five main stages, including mobilization  
108 (recruitment and consenting), mapping, census, survey, and results feedback and clinical  
109 follow-up. The specific objectives of this survey then were:

- 110 • To create a one-of-a-kind study for expanding on a large-scale prospective cohort  
111 research in an African population to evaluate a wide range of health indices—and to  
112 lay the platform for longer-term investigations.
- 113 • Using population, genetic and epidemiological techniques to provide aetiological  
114 insights into variance in cardio-metabolic and infectious risk factors.
- 115 • To help develop public health policies in other African countries by informing health  
116 policy and public health programs aimed at addressing the rise in NCDs in Uganda.

117 The cohort continues to offer as a valuable platform for investigating the relationship  
118 between communicable illnesses and NCDs in a regular annual survey of GPC.

119

## 120 **Genotype generation, quality control and Imputation**

121 The 2.5M Illumina chip array was used to genotype nearly 5,000 Ugandans at the Wellcome  
122 Trust Sanger Institute. Gurdasani and Fatumo et al 2019 have presented the quality control  
123 steps (Gurdasani *et al.*, 2019). In summary, we used a strict quality control process to  
124 perform a series of steps in a logical order to eliminate a total of 39,368 autosomal markers  
125 that failed to meet the quality metrics for SNP call rate (>97 percent, 25,037 SNPs) and HWE  
126 ( $p < 1 \times 10^{-8}$ , 14,331 SNPs). During sample QC, a total of 91 samples were eliminated because  
127 they failed the quality standards for sample call rate (>97%) or heterozygosity  
128 ( $H_0 = 0.209333 \pm 0.007416$  matching to the mean  $\pm 3SD$ ), or the sex extrapolated from the X-  
129 chromosome did not correspond to the reported sex. Three further samples were eliminated  
130 due to high relatedness ( $IBD > 0.90$ ). There were no samples that were classified as outliers in  
131 terms of population or ancestry. 2,230,258 autosomal markers and 4778 samples that met the  
132 stated threshold were subjected to further analysis. We carried out SNP Phasing with the aid  
133 of SHAPEIT2 (Delaneau, Coulonges and Zagury, 2008) using default settings, then  
134 imputation was done with IMPUTE2 (Howie, Donnelly and Marchini, 2009). All samples

135 were imputed with a combined reference that was created by combining the UG2G sequence  
136 resource (n = 2,000, whole genome sequence data from the African Genome Variation  
137 Project (n = 320),), and the 1000 Genomes phase 3 project (n = 2,504). The principal  
138 components analysis plot for the GPC participants (n=4,778) was published (Gurdasani et al.,  
139 2019) and is shown here in supplementary material figure S1.

140

#### 141 **Uganda 2000 Genomes (UG2G)**

142 The entire genomes of over 2,000 Ugandans from nine ethnolinguistic groups were  
143 sequenced using the Illumina HiSeq 2000 with 75bp paired end reads at low coverage, with  
144 an average coverage of 4x for each sample. 343 of these samples overlapped with people who  
145 had already been genotyped. An automated quality control process was used to bring down  
146 the data files that needs manual processing to ascertain the quality of BAM files produced.  
147 This method was based on the one developed for the UK10K project (Walter *et al.*, 2015)  
148 which used a set of algorithmically derived standards to determine summary data computed  
149 from the input BAMs. Any line that fell below the "fail" standard for any of the metrics was  
150 deleted; while lines falling below the warn standard for any of the scores were manually  
151 investigated; and any line that passes any of these scores was given a status of "pass".  
152 Overall, we deleted fourteen samples from the study. Full detailed on the quality control and  
153 how we computed the summary data has be described in Gurdasani *et al.*, 2019.

154

155

#### 156 **Merging of Sequenced and Genotyped Data**

157 We integrated sequenced and imputed genotyped data to produce an aggregated dataset to  
158 boost power for discovery in a genome-wide association studies. Because cryptic and family  
159 relatedness persisted across sequenced and genotyped data, we produced an aggregated  
160 dataset for analysis instead of separately meta-analysing the data, because data would be  
161 correlated rather than independent. As a result, conclusions from mixed model analysis that  
162 explicitly model this relationship are more likely to be true. We examined and deleted any  
163 consistent discrepancies between sequences and imputed genotype data after merging the  
164 two datasets. This was done by performing principal component analysis on the dataset to see  
165 if there was any distinction by data modality (imputed genotype data vs. sequenced data)  
166 among the 343 people who had their genotypes and sequences done in duplicate. On PCA,

167 we noticed a strong separation of genotype imputed and sequence data points. For these 343  
168 samples, we tested alternative concordance criteria between sequencing and imputed  
169 genotype data, screening out SNPs with a concordance of 0.80 and 0.90 the dataset. We  
170 discovered that to eliminate systematic effects detected between genotyping array and  
171 sequence data on PCA, a minimum concordance criterion of 0.90 was necessary.

172

173 There were no systematic changes between sequenced and genotyped data in PCAs after  
174 excluding 904,283 SNPs that exhibited 90 percent consonance in genotypes between the  
175 sequence and imputed genotype data . We examined the top ten PCs to confirm that  
176 systematic variations in the genomic data did not constitute an important axis of variation.

177

## 178 **Phenotype and laboratory measurement**

179 During survey round 22 which was conducted in 2011, several phenotypes based on clinical  
180 and physical examinations, laboratory tests, and self-reported questionnaires were collected  
181 from the respondents (Table 1) and these respondents who are still known to be alive and  
182 have not moved out of the GPC have been followed every year since then. A blood specimen  
183 was analysed for non-fasting blood lipids, blood cell traits (mean cell haemoglobin, red cell  
184 count, white cell count, mean cell haemoglobin concentration, haemoglobin, packed cell  
185 volume, mean cell volume and platelet), glycaemic characteristics, renal function, infectious  
186 biomarkers (HIV, hepatitis B and C). Basic demographics such as age, sex, marital status,  
187 and education level; anthropometrics such as BMI, weight, waist-to-hip ratio, height; blood  
188 pressure measurements; and lifestyle information such as smoking status, physical activities,  
189 and diet; sexual health & reproductive behaviour; sex education, condom use, pregnancy &  
190 outcome, and number of offspring were also collected (Table 1).

191

## 192 **Genetic discoveries and polygenic prediction in UGR**

193 A case in point in the use of our rich African genomics and phenotypic data, we undertook  
194 GWAS in 34 cardiometabolic traits including lipid, anthropometry traits, blood cell indices,  
195 HbA1c and reported novel loci associated with anthropometric, haematological, lipid, and  
196 glycaemic traits (Gurdasani *et al.*, 2019). In another study (Fatumo *et al* 2020 -eGFR), we  
197 reported the first ever GWAS of kidney function in continental Africa. Leveraging clinical  
198 relatedness and correlations among phenotypes, we explored the power of multivariate

199 GWAS to identify genetic risk factors implicating pleiotropic effects in blood cell traits  
200 (Fatumo *et al.*, 2019; Soremekun *et al.*, 2021), body shape (Nakabuye *et al.*, 2022) and liver  
201 function. Recently, we showed that genetic risk score derived from data of African American  
202 individuals enhance polygenic prediction of lipid traits and T2DM in Sub-Sahara African, but  
203 prediction varied greatly between another dataset from South Africa and our East African  
204 genomic data (Chikowore *et al.*, 2022, Kamiza *et al.*, 2022). We have also demonstrated the  
205 Mendelian randomisation evidence of relation between lipid trait and T2DM (Soremekun *et*  
206 *al.*, 2022), metabolic traits and stroke (Fatumo *et al.*, 2021). Collectively, our studies show a  
207 need for improved representation of Africans in genomic studies and ensuring the  
208 generalisation of findings for genomic medicine. This is further supported by findings from  
209 another study as well (Martin *et al.*, 2017). The UGR data has also been used to create a  
210 genotype imputation reference panel using UG2G available from the Sanger Imputation  
211 Service ([imputation.sanger.ac.uk](http://imputation.sanger.ac.uk)).

212

213

#### 214 **Contribution to collaborative studies**

215 We contribute to global genetic studies through partnerships and consortia, such as the  
216 African Partnership for Chronic Disease Research (APCDR), an international network of  
217 research groups that collaborate to support and promote collaborative chronic disease  
218 research across Africa. An initiative created in response to the changing distribution of  
219 communicable diseases and the rising burden of noncommunicable diseases, as well as the  
220 recognition that low- and middle-income countries (LMICs), including those in Sub-Saharan  
221 Africa, will need to expand their health-care capacities to effectively respond to these  
222 epidemiological transitions.

223

224 We combine research expertise with three other MRC Units (MRC Integrative Epidemiology  
225 Unit, MRC Population Health Research Unit, and MRC Unit for Lifelong Health and  
226 Ageing) to investigate the potential to use Mendelian randomization (MR) to assess the  
227 generalisability of existing drugs (e.g., statins, anti-diabetics, and anti-hypertensives) and  
228 identify the potential to tailor drugs with pilot studies focusing on established  
229 pharmacological targets to specific subpopulations (e.g. *CETP*, *HMGCR*) and to see how  
230 changes in genetic architecture affect efficacy estimates in different groups.

231



232 We are part of the CARDINAL (CARDiometabolic Disorders IN African-ancestry  
233 PopuLations) which is study site of an NIH-funded Polygenic Risk Methods in Diverse  
234 Populations (PRIMED) Consortium <https://primedconsortium.org/>. CARDINAL  
235 (Adebamowo, C.A. *et al.*, 2022) aims to integrate phenotype and genomic datasets from  
236 50,000 African individuals from seven cohort studies and evaluate PRSs to develop a novel  
237 method that considers ancestry-specific genomic regions to improve PRS prediction in  
238 populations with genetic substructure.

239

240 Furthermore, we recently provided GWAS data to the Meta-Analyses of Glucose and Insulin-  
241 related Variables Consortium (MAGIC) in order to find additional loci that influence  
242 glycaemic and metabolic traits (Chen, J. *et al* 2021). We are aiming for opportunities to  
243 contribute key phenotypes such as lipids, blood cell traits, kidney, etc to other consortia. For  
244 GBMI we will contribute all phenotype in Table 1 when require, including opportunity to  
245 measure not previously collected phenotype using resources in our organisation. We believe  
246 that team science allows scientists to make the most progress toward breakthrough  
247 discoveries that benefit human health.

248

#### 249 **Future directions**

250 The GPC is an active cohort of more than 22,000 participants. Genotype and sequence data is  
251 available for 6,657 respondents (N=5,000, 2,000 and 343 for genotype, sequence and  
252 overlapping samples respectively). We hope to genotype more samples to add on this  
253 resource. We also hope to sequence more samples at higher coverage in order to provide a  
254 reference panel with increased genome coverage. We also hope to extend our research into  
255 proteomics, metabolomics and single cell genomics in order to gain insights into the different  
256 mechanisms and pathways that could be implicated in different disease processes.

257

#### 258 **Data access and sharing of the UGR data**

259 Request for resources and information should be directed to UGR's Data Access Committee  
260 (DAC) via the Lead Contact, Dr. Segun Fatumo ([segun.fatumo@mrcuganda.org](mailto:segun.fatumo@mrcuganda.org);  
261 [segun.fatumo@lshtm.ac.uk](mailto:segun.fatumo@lshtm.ac.uk)). UGR's individual level data, genotype and sequence data are  
262 available under managed access to researchers. Requests for access will be granted for all  
263 research consistent with the consent provided by participants. This would include any

264 research in the context of health and disease, that does not involve identifying the participants  
265 in any way.

266

267 The array and low and high depth sequence data have been deposited at the European  
268 Genome-phenome Archive (EGA, <https://www.ebi.ac.uk/ega/>, accession numbers  
269 EGAS00001001558/EGAD00010000965, EGAS00001000545/EGAD00001001639 and  
270 EGAS00001000545/EGAD00001005346 respectively. Requests for access to data may be  
271 directed to [segun.fatumo@mrcuganda.org](mailto:segun.fatumo@mrcuganda.org). Applications are reviewed by data access  
272 committee (DAC) and access is granted if the request is consistent with the consent provided  
273 by participants. The data producers may be consulted by the DAC to evaluate potential  
274 ethical conflicts. Requestors also sign an agreement which governs the terms on which access  
275 to data is granted.

276

277 However, full GWAS summary statistics of UGR is freely available on GWAS catalog  
278 <https://www.ebi.ac.uk/gwas/> with study accession numbers: GCST009041 (Eosinophil  
279 counts), GCST009042 (Total cholesterol levels), GCST009043 (LDL cholesterol levels),  
280 GCST0090414 (HDL cholesterol levels), GCST009045 (Triglyceride levels),  
281 GCST009046 (Aspartate aminotransferase levels), GCST009047 (Alanine aminotransferase  
282 levels), GCST009048 (Serum albumin levels), GCST009049 (Serum alkaline phosphatase  
283 levels), GCST009050 (Gamma glutamyl transferase levels), GCST009051 (Bilirubin levels)  
284 , GCST009052 (Diastolic blood pressure) GCST009053 (Systolic blood pressure),  
285 GCST009054 (Hemoglobin A1c levels), GCST009055 (Height), GCST009056 (Weight),  
286 GCST009057 (Body mass index), GCST009058 (Waist circumference), GCST009059 (Hip  
287 circumference), GCST009060 (Waist-hip ratio), GCST009061 (White blood cell count),  
288 GCST009062 (Red blood cell count), GCST009063 (mean corpuscular hemoglobin),  
289 GCST009064 (mean corpuscular hemoglobin concentration), GCST009065 (mean  
290 corpuscular volume), GCST009032 (red blood cell distribution width), GCST009033  
291 (hematocrit), GCST009034 (hemoglobin measurement), GCST009035 (mean platelet  
292 volume), GCST009036 (platelet count), GCST009037 (lymphocyte count), GCST009038  
293 (monocyte count), GCST009039 (basophil count), GCST009040 (neutrophil count)

294

## 295 **Conclusions**

296 The Uganda Genome Resource is designed to make direct impact in biomedical and genetic  
297 research of health and disease in Uganda, Africa and globally. UGR has become one of the

298 model genomic resources in Africa and offers training opportunities to researchers from  
299 Uganda and the world at large. Here we present an overview of the UGR, showcase its broad  
300 range of phenotypic data, and highlights the genetic discoveries from UGR till date. In the  
301 next few years, UGR will continue to grow in sample size, and include proteomics,  
302 metabolomics, and single-cell genomic studies.

303

304

### 305 **Ethics**

306 The study was approved by the Science and Ethics Committee of the Uganda Virus Research  
307 Institute Research (UVRI) and Ethics Committee (UVRI-REC #HS 1978) and the Uganda  
308 National Council for Science and Technology (UNCST #SS 4283) and the East of England-  
309 Cambridge South (formerly Cambridgeshire 4) NHS Research Ethics Committee UK

310

311

312

313

314

315

316 **Table 1:** Sample characteristics of UGR participants at baseline

<b>Characteristic</b>	<b>All (n=7833)</b>	<b>Males (n=3425)</b>	<b>Female (n=4404)</b>
Number of individuals interviewed, N	7833	3425	4404
Age (years), median (IQR)	30 (17-46)	27 (17-44)	31 (19-47)
<b>Age group (years), N(%)</b>			
13-19	2398 (30.6)	1218 (35.6)	1180 (26.8)
20-29	1475 (18.8)	612 (17.9)	863 (19.6)
30-39	1311 (16.8)	495 (14.5)	816 (18.5)
40-49	1047 (13.4)	439 (12.8)	608 (13.8)
50-59	711 (9.1)	297 (8.7)	414 (9.4)
60+	887 (11.3)	364 (10.6)	523 (11.9)
BMI (kg/m <sup>2</sup> ), mean +/- SD	21.2 ± 3.83	20.1 ± 3.11	22.0 ± 4.13
<b>BMI Classification, N(%)</b>			
Underweight	1712 (22.6)	1031 (30.4)	681 (16.3)
Normal	4919 (65.0)	2188 (64.4)	2731 (65.4)
Overweight	739 (9.8)	156 (4.6)	583 (14.0)
Obese	201 (2.6)	21 (0.6)	180 (4.3)
<b>Smoking Status, N(%)</b>			
Current Smoker	641 (8.2)	553 (16.2)	88 (2.0)
Ex-Smoker	194 (2.5)	169 (4.9)	25 (0.6)
Never smoked	6990 (89.3)	2700 (78.9)	4290 (97.4)
<b>Alcohol Consumption, N(%)</b>			
Never or no alcohol use	5040 (70.1)	2052 (64.6)	2988 (74.4)
Infrequent current drinker	537 (7.5)	165 (5.2)	372 (9.3)
Frequent current drinker	1618 (22.5)	961 (30.2)	657 (16.4)
<b>Cardio metabolic Quantitative measurements (mean ± SD)</b>			
TC (mmol/L), mean ± SD	3.5 ± 0.98	3.3 ± 0.91	3.7 ± 1.00
HDL (mmol/L), mean ± SD	1.0 ± 0.41	0.9 ± 0.42	1.0 ± 0.40
LDL (mmol/L), mean ± SD	2.0 ± 0.77	1.8 ± 0.63	2.1 ± 0.80
Albumin, mean ± SD	41.3 ± 4.10	41.8 ± 4.15	40.9 ± 4.02
HbA1c, mean ± SD	3.3 ± 0.68	3.3 ± 0.69	3.3 ± 0.73
TG (mmol/L), mean ± SD	1.2 ± 0.61	1.1 ± 0.61	1.2 ± 0.62
ALT, median (IQR)	14.0 (17.8-22.9)	19.4 (15.6-25.1)	13.0 (16.4-21.3)
ALK, median (IQR)	71.3 (92.5-144.1)	74.3 (96.8-208.0)	68.5 (89.5-123.1)
AST, median (IQR)	21.2 (25.1-30.4)	23.8 (28.0-33.0)	19.8 (23.1-27.4)
GGT, median (IQR)	13.5 (18.7-28.0)	15.6 (21.6-33.0)	12.2 (17.0-24.2)
Bilirubin, median (IQR)	5.2 (7.7-12.0)	5.92 (8.9-14.2)	4.8 (6.9-10.5)
<b>Anthropometric Measurements</b>			
Weight (kg), mean ± SD	52.6 ± 11.35	52.4 ± 11.37	52.7 ± 11.33
Height (cm) , mean ± SD	157.2 ± 9.19	160.7 ± 10.54	154.5 ± 6.83
SBP (mmHg) , mean ± SD	122.4 ± 17.0	123.5 ± 16.2	121.6 ± 17.51
DBP (mmHg) , mean ± SD	74.2 ± 10.26	73.5 ± 10.39	74.7 ± 10.12
<b>Anaemia</b>			
WBC	5.1 ± 1.51	5.1 ± 1.58	5.1 ± 1.58

RBC	4.7 ± 0.62	4.9 ± 0.65	4.6 ± 0.56
HGB	13.6 ± 1.62	14.2 ± 1.74	13.1 ± 1.33
WHR	0.85 ± 0.16	0.86 ± 0.17	0.8 ± 0.16
MCH	28.9 ± 2.91	29.1 ± 2.92	28.8 ± 2.90
MCHC	33.7 ± 1.19	33.7 ± 1.24	33.7 ± 1.15
RDW	13.1 ± 1.34	13.1 ± 1.40	13.1 ± 1.29
MPV	8.7 ± 0.83	8.7 ± 0.83	8.7 ± 0.82
Platelet count (PLT)	216.9 ± 77.7	207.9 ± 77.3	223.9 ± 77.30
Lymphocytes	2.4 ± 0.83	2.5 ± 0.92	2.4 ± 0.76
Monocytes	0.3 ± 0.12	0.29 ± 0.14	0.3 ± 0.11
Basophils	0.05 ± 0.04	0.05 ± 0.42	0.05 ± 1.58
Neutrophils	1.9 ± 0.86	1.9 ± 0.84	2.0 ± 0.88
Eosinophils	0.35 ± 0.39	0.4 ± 0.40	0.3 ± 0.39
Vaccination			
<b>Received BCG vaccine, N(%)</b>			
Yes	1421 (18.2)	631 (18.4)	790 (18.0)
No	553 (7.1)	221 (6.5)	332 (7.5)
Don't know	5850 (74.8)	2570 (75.1)	3280 (74.5)
<b>Received Oral Polio vaccine, N(%)</b>			
Yes	1398 (17.9)	612 (17.9)	786 (17.9)
No	578 (7.4)	237 (6.9)	341 (7.8)
Don't know	5848 (74.7)	2573 (75.2)	3275 (74.7)
<b>Received DPT vaccine, N(%)</b>			
Yes	1415 (18.1)	626 (18.3)	789 (17.9)
No	541 (6.9)	212 (6.2)	329 (7.5)
Don't know	5868 (75.0)	2584 (75.5)	3284 (74.6)
<b>Received Measles vaccine, N(%)</b>			
Yes	1561 (20.0)	685 (20.0)	876 (19.9)
No	561 (7.2)	226 (6.6)	335 (7.6)
Don't know	5702 (72.9)	2511 (73.4)	3191 (72.5)
<b>Received TB vaccine, N(%)</b>			
Yes	54 (0.7)	22 (0.6)	32 (0.7)
No	7326 (92.5)	3131 (91.5)	4105 (93.2)
Don't know	535 (6.8)	269 (7.9)	266 (6.1)
<b>Received Hepatitis B vaccine, N(%)</b>			
Yes	54 (0.69)	22 (0.6)	32 (0.7)
No	7258 (92.8)	3141 (91.8)	4117 (93.5)
Don't know	513 (6.56)	259 (7.6)	254 (5.8)
<b>Received Tetanus vaccine, N(%)</b>			
Yes	1881 (24.0)	9 (0.3)	1872 (42.5)
No	5462 (69.8)	3154 (92.2)	2308 (52.4)
Don't know	482 (6.2)	259 (7.6)	223 (5.1)
<b>Received Tetanus booster vaccine, N(%)</b>			
Yes	1285 (16.4)	298 (8.7)	987 (22.4)
No	6044 (77.3)	2863 (83.7)	3181 (72.3)
Don't know	495 (6.3)	260 (7.6)	235 (5.3)

<b>Received Rabies vaccine, N(%)</b>			
Yes	46 (0.6)	17 (0.5)	29 (0.7)
No	7268 (92.9)	3138 (91.7)	4130 (93.8)
Don't know	511 (6.5)	257 (7.8)	244 (5.5)
<b>Self-report of diseases</b>			
<b>Hypertension, N(%)</b>			
Hypertensive	487 (6.2)	130 (3.8)	357 (8.1)
Normal	7338 (93.8)	3292 (96.2)	4046 (91.9)
<b>Diabetes, N (%)</b>			
Diabetic	102 (1.3)	44 (1.3)	58 (1.3)
Normal	7723 (98.7)	3378 (98.7)	4345 (98.7)
<b>Blood test results</b>			
<b>HIV status, N (%)</b>			
Negative	7185 (92.4)	3197 (94.0)	3988 (91.1)
Positive	593 (7.6)	204 (6.0)	389 (8.9)
<b>Hepatitis B, N (%)</b>			
Negative	7536 (97.3)	3268 (96.5)	4268 (97.9)
Positive	210 (2.7)	117 (3.5)	93 (2.1)
<b>Hepatitis C, N (%)</b>			
Negative	7536 (97.3)	3268 (96.5)	4268 (97.9)
Positive	210 (2.7)	117 (3.5)	93 (2.1)

317

318 SD, standard deviation; TC, Total Cholesterol; LDL, Low-density lipoprotein; TG,  
 319 Triglycerides; HDL, High-density lipoprotein; ALT, Alanine aminotransferase ; AST,  
 320 Aspartate aminotransferase ; ALP, Alkaline phosphatase; GGT, Gamma glutamyltransferase;  
 321 DBP, diastolic blood pressure; SBP, systolic blood pressure; BMI, Body mass index; WHR,  
 322 Waist-Hip Ratio; WBC, White blood cell, RBC, Red blood cell, MCV, mean corpuscular  
 323 volume; MCH, mean corpuscular haemoglobin; MCHC, mean corpuscular hemoglobin  
 324 concentration; RDW, red blood cell distribution width; DPT, Diphtheria, pertussis and  
 325 tetanus vaccine

## Reference

- Asiki, G. *et al.* (2013) 'The general population cohort in rural south-western Uganda: a platform for communicable and non-communicable disease studies.', *International journal of epidemiology*, 42(1), pp. 129–141. doi: 10.1093/ije/dys234.
- Chikowore, T. *et al.* (2022) 'Polygenic Prediction of Type 2 Diabetes in Africa.', *Diabetes care*, 45(3), pp. 717–723. doi: 10.2337/dc21-0365.
- Delaneau, O., Coulonges, C. and Zagury, J.-F. (2008) 'Shape-IT: new rapid and accurate algorithm for haplotype inference', *BMC Bioinformatics*, 9(1), p. 540. doi: 10.1186/1471-2105-9-540.
- Fatumo, S. *et al.* (2019) 'Complimentary methods for multivariate genome-wide association study identify new susceptibility genes for blood cell traits', *Frontiers in Genetics*, 10(APR), pp. 1–13. doi: 10.3389/fgene.2019.00334.
- Fatumo, S. *et al.* (2021) 'Metabolic Traits and Stroke Risk in Individuals of African Ancestry: Mendelian Randomization Analysis', *Stroke*. American Heart Association, 0(0), p. STROKEAHA.121.034747. doi: 10.1161/STROKEAHA.121.034747.
- Fatumo, S. *et al.* (2022) 'A roadmap to increase diversity in genomic studies', *Nature Medicine*, 28(2), pp. 243–250. doi: 10.1038/s41591-021-01672-4.
- Fatumo, S., Chikowore, T. and Kuchenbaecker, K. (2022) 'Editorial: Genetics of Complex Traits and Diseases From Under-Represented Populations', *Frontiers in Genetics*. Available at: <https://www.frontiersin.org/article/10.3389/fgene.2021.817683>.
- Gurdasani, D. *et al.* (2019) 'Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa', *Cell*, 179(4), pp. 984-1002.e36. doi: 10.1016/j.cell.2019.10.004.
- Howie, B. N., Donnelly, P. and Marchini, J. (2009) 'A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies', *PLoS Genetics*. Public Library of Science, 5(6), p. e1000529. Available at: <https://doi.org/10.1371/journal.pgen.1000529>.
- Nakabuye, M. *et al.* (2022) 'Genetic loci implicated in meta-analysis of body shape in Africans', *Nutrition, Metabolism and Cardiovascular Diseases*. doi: <https://doi.org/10.1016/j.numecd.2022.03.010>.
- Soremekun, O. *et al.* (2021) 'Genome-Wide Association and Mendelian Randomization Analysis Reveal the Causal Relationship Between White Blood Cell Subtypes and Asthma in Africans', *Frontiers in Genetics*. Available at: <https://www.frontiersin.org/article/10.3389/fgene.2021.749415>.
- Soremekun, O. *et al.* (2022) 'Lipid traits and type 2 diabetes risk in African ancestry individuals: A Mendelian Randomization study', *eBioMedicine*, 78, p. 103953. doi: <https://doi.org/10.1016/j.ebiom.2022.103953>.
- Walter, K. *et al.* (2015) 'The UK10K project identifies rare variants in health and disease', *Nature*, 526(7571), pp. 82–90. doi: 10.1038/nature14962.
- Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D. and Kenny, E.E., 2017. Human demographic history impacts genetic

risk prediction across diverse populations. *The American Journal of Human Genetics*, 100(4), pp.635-649.

Chen, J., Spracklen, C.N., Marenne, G. *et al.* The trans-ancestral genomic architecture of glyceic traits. *Nat Genet* **53**, 840–860 (2021). <https://doi.org/10.1038/s41588-021-00852-9>

Adebamowo, C.A., Adeyemo, A., Ashaye, A. *et al.* Polygenic risk scores for CARDINAL study. *Nat Genet* (2022). <https://doi.org/10.1038/s41588-022-01074-3>

Kamiza, A., Toure, S., Vujkovic, M., Machipisa, T., Soremekun, O., Chikowore, T & Fatumo, S., (2022). Polygenic prediction of lipid traits in sub-Saharan Africans. <https://doi.org/10.21203/rs.3.rs-824992/v1>

### Supplementary material

Figure S1: The principal components analysis plot for the general population cohort participants (Source: Gurdasani D *et al.*, 2019)

