Journal of Human Genetics (2010) 55, 366–374
© 2010 The Japan Society of Human Genetics All rights reserved 1434-5161/10 $32.00
www.nature.com/jhg

ORIGINAL ARTICLE

Comparison of genome-wide variation between Malawians and African ancestry HapMap populations

Bonnie R Joubert[1], Kari E North[1,2], Yunfei Wang[3,4], Victor Mwapasa[5], Nora Franceschini[1], Steven R Meshnick[1] and Ethan M Lange[2,3,4], and the NIAID Center for HIV/AIDS Vaccine Immunology

Understanding genetic variation between populations is important because it affects the portability of human genome-wide analytical methods. We compared genetic variation and substructure between Malawians and other African and non-African HapMap populations. Allele frequencies and adjacent linkage disequilibrium (LD) were measured for 617 715 single nucleotide polymorphisms (SNPs) across subject genomes. Allele frequencies in the Malawian population (N=226) were highly correlated with allele frequencies in HapMap populations of African ancestry (AFA, N=376), namely Yoruban in Ibadan, Nigeria (Spearman's r^2=0.97), Luhya in Webuye, Kenya (r^2=0.97), African Americans in the southwest United States (r^2=0.94) and Maasai in Kinyawa, Kenya (r^2=0.91). This correlation was much lower between Malawians and other ancestry populations ($r^2<0.52$). LD correlations between Malawians and HapMap populations were strongest for the populations of AFA (AFA $r^2>0.82$, other ancestries $r^2<0.57$). Principal components analyses revealed little population substructure within our Malawi sample but provided clear distinction between Malawians, AFA populations and two European populations. Five SNPs within the lactase gene (*LCT*) had substantially different allele frequencies between the Malawi population and Maasai in Kenyawa, Kenya (rs3769013, rs730005, rs3769012, rs2304370; *P*-values $<1\times10^{-33}$).

Journal of Human Genetics (2010) **55,** 366–374; doi:10.1038/jhg.2010.41; published online 20 May 2010

Keywords: African ancestry; genome variation; HapMap; lactase; *LCT*; population genetics; population substructure

INTRODUCTION

The International HapMap Project has offered an extraordinary amount of information on common genetic variation across the human genome, leading to publicly available data including more than 1 million single nucleotide polymorphisms (SNPs) genotyped in populations across the world.[1] Currently, genetic data from 11 populations are included: Han Chinese in Bejing, China (CHB), Japanese in Tokyo, Japan (JPT), Utah residents with Northern and Western European ancestry from the CEPH collection (CEU), Yoruba in Ibadan, Nigeria (YRI), African ancestry in Southwest USA (ASW), Chinese in Metropolitan Denver, Colorado (CHD), Gujarati Indians in Houston, Texas (GIH), Luhya in Webuye, Kenya (LWK), Mexican ancestry in Los Angeles, California (MEX), Maasai in Kinyawa, Kenya (MKK), and Toscans in Italy (TSI). HapMap data allow researchers to characterize haplotype patterns and allele frequencies of SNPs in the HapMap populations and to compare such patterns to those observed in other populations. This exercise has helped researchers to more adequately understand global genetic diversity and has facilitated a greater understanding of the genetic etiology of disease.

A number of studies have described transferability or 'portability' of tagSNPs in the HapMap populations to tagSNPs in other populations. Most of these studies have focused on specific ENCODE regions,[2–4] candidate genes[5,6] or collections of SNPs genotyped across one to three chromosomes.[7–10] No such studies have included Malawian individuals. The goal of this work was to compare genetic variation among HapMap populations of African ancestry (AFA) with a population in Blantyre, Malawi. Information on 617 715 SNPs across 22 autosomal chromosomes of the human genome is described. To our knowledge, this study is foremost in incorporating Malawians into the population genetics forum, and adds an additional assessment of genetic variability within Malawi in relation to self-reported ethnicity. The findings from this study add to our understanding of genomic variation across the African continent as well as within one urban area of Malawi.

MATERIALS AND METHODS

Malawi study population

The participants involved in this work are a subset of a larger previously conducted cohort study of malaria and HIV in pregnancy.[11,12] The prospective cohort was conducted from 2000 to 2004 and included 3825 consenting

[1]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA; [2]Carolina Center for Genome Sciences, University of North Carolina, Chapel Hill, NC, USA; [3]Department of Genetics, School of Medicine, University of North Carolina, Chapel Hill, NC, USA; [4]Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA and [5]College of Medicine, University of Malawi, Blantyre, Malawi
Correspondence: Dr BR Joubert, Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, McGavran-Greenberg Hall, CB #7435, Chapel Hill, NC 27599, USA.
E-mail: joubert.bonnie@epa.gov

Received 20 November 2009; revised 24 February 2010; accepted 24 March 2010; published online 20 May 2010

pregnant women admitted to Queen Elizabeth Central Hospital.[12] HIV testing was performed at delivery and patients were followed up at 6 and 12 weeks post delivery. A total of 1157 women tested positive for HIV, 884 of which delivered at Queen Elizabeth Central Hospital resulting in 807 singleton live births. At delivery, 751 infants were tested for HIV, identifying 65 HIV positive infants at birth. Of the 668 HIV negative infants, 179 were lost to follow-up. A total of 507 infants were tested for HIV at 6 or 12 weeks, resulting in 89 additional HIV positive infants. A subset of the cohort (N=246) consisting of all HIV positive (at birth, 6 weeks or 12 weeks) infants and an equal proportion of HIV negative (at all visits) infants of HIV positive mothers were selected. The HIV negative infants were obtained from a random sample of the HIV-exposed negative infants and had a similar distribution across time of enrollment as the cases. Both positive and negative infants were required to have quality DNA samples available. This subset was originally used in a genome-wide association study to assess the association between SNPs across infant genomes and susceptibility to maternal HIV infection.[13] The genome-wide SNP data were applied to this work to evaluate the external generalizability of our findings.

Self-reported ethnicity for our final dataset of 226 subjects after quality control (see below) was available for the Malawi dataset, which included the following groups: Ngoni (N=62), Lomwe (N=58), Yao (N=33), Chewa (N=18), Tumbuka (N=15), Mang'anja/Nyanja (N=15), Sena (N=12), Khokhola (N=4), Chipeta (N=1), Likoma (N=1), Nkhonde (N=1), Ntcheu (N=1), Portugues (N=1), Tonga (N=1) and 3 missing. A categorical variable was created corresponding to the first 7 groups listed and groups with a frequency <12 were combined into one category (Other, N=15). The Nyanja and Mang'anja are different names for the same ethnic group.[14] The study population included from our data collection will be referred to as individuals of various self-reported ethnicity in Blantyre, Malawi (BMW).

Genotyping

Genotyping was completed for 114 HIV-exposed, infected infants and 132 exposed, uninfected infants. Additional cases were not available because DNA samples were of poor quality or had a very low concentration of DNA. The genotyping was performed at Duke University Genotyping Core Laboratories, using Illumina's HumanHap650Y Genotyping BeadChip version 3 (Illumina, San Diego, CA, USA). This BeadChip enabled whole-genome genotyping of over 655 000 tagSNPs derived from the International HapMap Project[15] and over 100 000 tag SNPs selected based on the Yoruban Nigerian HapMap Population. The 650Y BeadChip v3 used information in dbSNP up to version 126.

Quality control

For the Malawi study population, quality control for genotyping error was performed at Duke University Genomic Laboratories and as described earlier.[16] Briefly, all samples were brought into a BeadStudio data file using an Illuminum cluster file and clustering of samples was evaluated to determine random clustering of SNPs. Samples with very low call rates ($<95\%$) were excluded. Subsequent reclustering of undeleted SNPs and additional exclusion by call rate was performed.[16] SNPs with Het Excess values between -1.0 to -0.1 and 0.1 to 1.0 were evaluated to determine whether raw and normalized data indicated clean calls for the genotypes.[16]

Statistical quality control measures were performed at UNC Chapel Hill. Individuals missing more than 10% of marker data and SNPs with a genotyping rate less than or equal to 10% were excluded from analyses. Related individuals were identified by first estimating identity by descent. A small number of individuals with estimated genome-wide identity by descent values >0.05 were removed (N=5). All statistical quality control measures were performed in *PLINK* version 1.05.[17] After completing quality control, a total sample size of N=226 BMW subjects were included in subsequent data analyses.

Data management and integration of HapMap populations

All HapMap populations from Phase 3 were included in this study. HapMap data were downloaded in *PLINK* format from the International HapMap Project website (http://hapmap.ncbi.nlm.nih.gov/). The samples from the 11 HapMap populations were genotyped on approximately 1.5 million SNPs using the Illumina Human1M and Affymetrix Genome-Wide Human SNP Array 6.0 platforms. For this study, populations were processed through identical statistical quality control as completed for the Malawi population with the exception of identity by descent estimation. To ensure the populations included only unrelated individuals, offspring of populations with duos or trios (ASW, CEU, MEX, MKK, YRI) were removed. HapMap populations were then merged, either individually or jointly, with the Malawi population. Owing to strand differences between the Malawi population and the HapMap populations, some Malawi strands were flipped and the files were remerged. The genotype data from HapMap and our sample were based on the same dbSNP build 126.

Statistical analysis

Four analytic methods were conducted to compare genetic variation between populations: (1) the correlation of allele frequencies across the genome; (2) the correlation of adjacent SNP–SNP linkage disequilibrium (LD) across the genome; (3) allelic-based χ^2-tests to evaluate the association between population and SNPs; and (4) principal component (PC) analysis to evaluate population substructure. Each HapMap population was individually compared with the Malawi population and within Malawi; the self-reported ethnic groups were compared. SNPs included on Illumina's HumanHap650Y Genotyping BeadChip were selected based on allele frequencies and LD patterns in the CEU and YRI HapMap samples. To reduce the potential for bias because of the SNP selection strategy on the HumanHap650Y panel, when comparing different populations we excluded SNPs with relatively rare minor allele frequencies (MAF) in our calculations. Specifically, in all analyses, comparisons were only made between SNPs with MAF >0.05 in each included population sample.

Comparison of allele frequencies and adjacent SNP–SNP LD

Allele frequencies were computed for each population using *PLINK* version 1.05.[17] Each allele frequency file was formatted and imported into *STATA* version 10.[18] To compare allele frequencies across populations, a generic coded allele was set for each SNP, using alphabetical hierarchy (A<C<T<G). For example, if the alleles were A and G for a particular SNP, the coded allele would be designated as A and the non-coded allele as G, regardless of the observed allele frequency in a population. Spearman's correlation coefficient was computed for the allele frequencies of the coded alleles between each pair of populations (that is BMW vs ASW, BMW vs YRI, and so on) using all SNPs with MAF >0.05 in both populations. Adjacent LD was estimated by calculating the standard r^2 value for all pairs of adjacent SNPs using *PLINK* version 1.05.[17] For each pair of populations, Spearman's correlation coefficient between populations was calculated using all adjacent SNP pair r^2 estimates.

Allelic-based (1 d.f.) χ^2-tests were systematically computed to identify SNPs with significantly different allele frequencies between the Malawi population and each HapMap population containing subjects of AFA. Specifically, we defined a dichotomous outcome variable, with value of '1' assigned to the BMW population and value of '2' assigned to the comparison population (ASW, LWK, MKK or YRI). Two measures of the genomic inflation factor (GIF), based on the median and mean χ^2 statistic over all SNP comparisons, were computed for each genome-wide association analysis using *PLINK* version 1.05.[17] As χ^2 statistics are affected by sample size under the alternative hypothesis, analyses were also performed using a random selection of 42 individuals from each HapMap comparison group to facilitate more direct comparison of the results across the HapMap populations. The value of 42 was used as it represented the smallest group (ASW, N=42).

Population substructure

Population substructure was evaluated using PC analyses for (1) the Malawi population; (2) the Malawi population combined with the HapMap populations of African (ASW, LWK, MKK, YRI) ancestry; and (3) the Malawi population combined with the HapMap populations of African and European (CEU, TSI) ancestry. The PC analyses were conducted using *EIGENSOFT* version 2.0.[19] SNP inclusion in the PC analysis was restricted to autosomal SNPs that had MAF >0.05 and observed genotype frequencies consistent with Hardy–Weinberg equilibrium expected proportions ($P>0.001$) in each participating individual population. Strict SNP pruning based on pair-wise SNP–SNP LD was conducted to identify a subset of independent SNPs for inclusion in PC analysis. Specifically, we calculated pair-wise SNP–SNP LD, measured by r^2, between all SNP pairs within 500 kb in the BMW sample using *PLINK*. A custom computer program was used to select the largest number of SNPs from each chromosome such that each selected SNP had no other selected SNPs within 500 kb that were in LD with it (defined by $r^2>0.01$). On the basis of these selection rules, we identified (1) 23 612, (2) 18 481 and (3) 16 912 SNPs for use in the three PC analyses, respectively.

Finally, we performed global ancestry estimation using the software *ADMIXTURE* on our combined sample of seven African and European populations using the same 16 912 SNPs included in the PC analyses.[20] *ADMIXTURE* uses a maximum likelihood approach to model the probabilities of the observed genotype data using ancestry proportions and population allele frequencies. Similar to the program *STRUCTURE*, *ADMIXTURE* requires the user specification of the number of postulated ancestral populations that preceded the observed populations included in the study sample. For this study, we considered K=2, 3 and 5 ancestral populations.

RESULTS

Quality control

The HapMap data available at the time of this study contained approximately 1.5 million SNPs for 1115 individuals of 11 unique ethnic groups. The Malawi data included 112 males and 114 females, with a genotyping call rate of 99.975%. Following quality control, the combined HapMap and Malawi dataset included 1150 individuals, 602 of which were of AFA, and 633 763 SNPs, 617 715 of which were on autosomal chromosomes and incorporated in the analyses.

Comparison of allele frequencies across populations

The allele frequencies of the Malawi population were highly correlated with the allele frequencies of the HapMap AFA populations. Among the AFA subgroups, the allele frequencies of the YRI and the Luhya in Webuye, Kenya were most strongly correlated to those of the Malawians (Table 1; Supplementary Figure S1). Interestingly, allele frequencies of the Malawi population were more closely correlated with allele frequencies of individuals of AFA in the Southwest USA than they were with that of the Maasai in Kinyawa, Kenya (Table 1; Supplementary Figure S1). Much lower correlations in allele frequencies were observed between the Malawi population and HapMap populations of other ancestry ($r^2<0.52$).

A total of 14 self-reported ethnic groups comprised the BMW group. The genotyping was completed for infants of the mother–infant pairs, so the ethnic groups reflect self-reported maternal ethnicity. Data on paternal ethnic group were not available. Allele frequencies were highly correlated across all ethnic groups in the Malawian study population (Table 2). The greatest correlation in allele frequency was observed for Ngoni and Lomwe, which represented the majority of infants in the dataset (27 and 26%, respectively). The smallest correlation was observed between the Sena and Mang'anja/Nyanja ethnic groups (Table 2). All SNPs summarized were restricted to having an MAF >0.05. This resulted in approximately 569 373 SNPs compared by population. This number was slightly different for each comparison, as the number of SNPs with an MAF >0.05 varied by population.

Similar to allele frequencies, adjacent SNP–SNP LD was highly correlated across populations of AFA (Table 3). A lower correlation in adjacent LD was observed between the Malawi population and other ancestry HapMap populations ($r^2<0.57$; Table 3; Supplementary Figure S2). The average LD between pairs of adjacent SNPs in the Malawi population was similar to that observed in the other AFA HapMap populations and substantially lower than the average LD observed between adjacent SNPs in the other ancestry HapMap populations (Supplementary Table S1).

Table 2 Correlation of allele frequency across Malawi ethnic groups[a]

Population	*Ngoni*	*Lomwe*	*Yao*	*Chewa*	*Tumbuka*	*Nyanja/ Mang'anja*	*Sena*	*Other*
Ngoni	1							
Lomwe	0.969	1						
Yao	0.958	0.956	1					
Chewa	0.937	0.935	0.925	1				
Tumbuka	0.928	0.927	0.916	0.897	1			
Nyanja/ Mang'anja	0.919	0.918	0.907	0.888	0.880	1		
Sena	0.914	0.913	0.903	0.883	0.875	0.867	1	
Other	0.928	0.927	0.916	0.897	0.889	0.879	0.876	1

[a]Spearman's correlation coefficients for allele frequencies, minor allele frequencies >0.05.

Table 1 Correlation of allele frequency across populations[a]

Ancestry[b]	*Population*	*BMW*	*YRI*	*LWK*	*MKK*	*ASW*	*CEU*	*TSI*	*CHB*	*CHD*	*GIH*	*JPT*	*MEX*
AFA	BMW	1											
AFA	YRI	0.972	1										
AFA	LWK	0.971	0.960	1									
AFA	MKK	0.913	0.906	0.932	1								
AFA	ASW	0.942	0.947	0.937	0.917	1							
EUA	CEU	0.475	0.474	0.498	0.618	0.622	1						
EUA	TSI	0.486	0.485	0.510	0.634	0.628	0.967	1					
ASA	CHB	0.418	0.417	0.433	0.500	0.498	0.607	0.602	1				
ASA	CHD	0.415	0.415	0.430	0.496	0.495	0.602	0.597	0.976	1			
ASA	GIH	0.511	0.510	0.532	0.632	0.628	0.850	0.848	0.712	0.709	1		
ASA	JPT	0.415	0.415	0.430	0.497	0.495	0.603	0.598	0.959	0.952	0.709	1	
MXA	MEX	0.490	0.490	0.508	0.603	0.609	0.834	0.826	0.735	0.727	0.811	0.733	1

Abbreviations: ASW, African ancestry in Southwest USA; BMW, Individuals of various self-reported ancestry in Blantyre, Malawi; CEU, Utah residents with Northern and Western European ancestry from the CEPH collection; CHB, Han Chinese in Bejing, China; CHD, Chinese in Metropolitan Denver, Colorado; GIH, Gujarati Indians in Houston, Texas; JPT, Japanese in Tokyo, Japan; LWK, Luhya in Webuye, Kenya; MEX, Mexican ancestry in Los Angeles, California; MKK, Maasai in Kinyawa, Kenya; TSI, Toscans in Italy; YRI, Yoruba in Ibadan, Nigeria.

[a]Spearman's correlation coefficients for allele frequencies, minor allele frequencies >0.05.

[b]Ancestry abbreviations recommended by HapMap: AFA, African ancestry; ASA, Asian ancestry; EUA, European ancestry; MXA, Mexican ancestry.

Table 3 Correlation of adjacent linkage disequilibrium across populations[a]

Ancestry[b]	*Population*	*BMW*	*YRI*	*LWK*	*MKK*	*ASW*	*CEU*	*TSI*	*CHB*	*CHD*	*GIH*	*JPT*	*MEX*
AFA	BMW	1											
AFA	YRI	0.896	1										
AFA	LWK	0.887	0.857	1									
AFA	MKK	0.826	0.805	0.836	1								
AFA	ASW	0.822	0.819	0.802	0.805	1							
EUA	CEU	0.543	0.533	0.555	0.666	0.634	1						
EUA	TSI	0.550	0.540	0.562	0.676	0.636	0.937	1					
ASA	CHB	0.523	0.514	0.529	0.604	0.569	0.700	0.695	1				
ASA	CHD	0.520	0.511	0.526	0.600	0.564	0.694	0.689	0.945	1			
ASA	GIH	0.564	0.553	0.574	0.673	0.634	0.852	0.848	0.768	0.761	1		
ASA	JPT	0.517	0.508	0.523	0.597	0.562	0.692	0.687	0.932	0.925	0.761	1	
MXA	MEX	0.551	0.542	0.560	0.652	0.623	0.829	0.820	0.760	0.753	0.810	0.755	1

Abbreviations: ASW, African ancestry in Southwest USA; BMW, Individuals of various self-reported ancestry in Blantyre, Malawi; CEU, Utah residents with Northern and Western European ancestry from the CEPH collection; CHB, Han Chinese in Bejing, China; CHD, Chinese in Metropolitan Denver, Colorado; GIH, Gujarati Indians in Houston, Texas; JPT, Japanese in Tokyo, Japan; LWK, Luhya in Webuye, Kenya; MEX, Mexican ancestry in Los Angeles, California; MKK, Maasai in Kinyawa, Kenya; TSI, Toscans in Italy; YRI, Yoruba in Ibadan, Nigeria.

[a] Spearman's correlation coefficients, minor allele frequencies >0.05. Linkage disequilibrium measured in each population using adjacent marker r^2.

[b] Ancestry abbreviations recommended by HapMap: AFA, African ancestry; ASA, Asian ancestry; EUA, European ancestry; MXA, Mexican ancestry.

SNPs associated with population membership

Allelic-based (1 d.f.) χ^2-tests revealed that the allele frequencies of many SNPs were significantly different between the Malawians and all four HapMap populations comprised of individuals of AFA. The greatest differences in observed allele frequencies existed between the Malawians and the Maasai in Kinyawa, Kenya (BMW vs MKK GIF or median χ^2=9.21, mean χ^2=9.16; random sample of 42 GIF=4.31, mean χ^2=4.49). The Malawians and the individuals of AFA in the Southwest United States (BMW vs ASW) also showed strong differences in allele frequencies (GIF=2.54, mean χ^2=2.72, for 42 total individuals in the ASW group). Smaller, though still highly significant, differences in allele frequencies were observed between the Malawi population and the Luhya in Webuye, Kenya (BMW vs LWK) (GIF=2.22, mean χ^2=2.25; sample of 42 GIF=1.77, mean χ^2=1.78) and between the Malawi population and the Yorubans in Ibadan, Nigeria (BMW vs YRI) (GIF=2.59, mean χ^2=2.62; sample of 42 GIF=1.78, mean χ^2=1.81).

The top SNPs associated with population membership were investigated for functional significance, for each BMW vs population comparison. Depending on the comparison of interest, many SNPs reached genome-wide statistical significance, having Bonferroni corrected *P*-values <0.05 (Figure 1; Table 4). For the comparison of allele frequencies between BMW and MKK, over 100 SNPs had a Bonferroni corrected *P*-value $<1\times10^{-23}$. The most significant SNPs (the top 16) were found on chromosome 2. Of the 30 most significant SNPs within genes, 4 were located within the lactase gene (*LCT*) (frequencies of the 3 most significant SNPs are shown in Figure 2 and Table 4). This gene is involved in production of the lactase enzyme, essential for the digestion of lactose, and has clinical implications for lactose intolerance.[21]

Fewer SNPs were statistically significantly different between BMW and LWK. Three SNPs had a Bonferroni corrected $P\leqslant1\times10^{-7}$, two of which were located within genes (Table 4). For BMW vs YRI, eight SNPs had a Bonferroni corrected $P\leqslant1\times10^{-7}$. Three SNPs with Bonferroni corrected $P\leqslant1\times10^{-4}$ (one shown in Table 4 with Bonferroni corrected $P\leqslant1\times10^{-7}$) were within the major histocompatibility complex, class II, DP α 1 (*HLA-DPA1*) gene. This gene is involved in many immunological functions, including interaction with HIV-1.[21] The comparison between BMW and ASW resulted in 69 SNPs with a Bonferroni corrected $P\leqslant1\times10^{-7}$, most of which were not located within genes (66.7%).

Population substructure

Three separate PC analyses were performed to evaluate population substructure: (1) for the Malawi population by itself; (2) for the HapMap populations of AFA (ASW, LWK, MKK, YRI) combined with the Malawi population; and (3) for the HapMap populations of AFA, the Malawi population and two HapMap populations of European ancestry (ASW, LWK, MKK, YRI, BMW, CEU, TSI). The eigenvalues (EVs) for the first 10 PCs for each analysis is reported in the Supplementary Material (Supplementary Table S2). PC analysis for the Malawi population revealed little evidence for population substructure (Figure 3) and we were unable to detect any genetic variation across self-reported ethnicity. The largest EV, corresponding to PC1, was only 1.21 and nine next largest EVs decreased very slowly (Supplementary Table S2).

PC analyses using the HapMap AFA populations and the Malawi population revealed clear distinction among the different populations of AFA (Figure 4). Overall, the BMW, LWK and YRI samples showed tight within population clustering of PC values whereas the ASW and MKK populations showed relatively strong dispersion. PC1 (corresponding EV=11.62) values were generally ordered MKK>ASW>LWK>YRI>BMW, with considerable overlap between ASW and both MKK and LWK (Figure 4). PC1 provided distinction between BMW and all HapMap populations of AFA with the exception of YRI. Two BMW subjects had overlap with LWK. PC2 (corresponding EV=2.96) provided two very distinct clusters, YRI and ASW in one cluster and BMW, LWK and MKK in the other cluster (Figure 4). The next eight PCs (corresponding EVs ranging from 2.75 to 2.13) were all driven by diversity within the MKK sample and provided little distinction between the other African samples (data not shown).

PC analyses including the Malawi population and the HapMap African and European ancestry populations revealed clear separation across all seven populations (Supplementary Figures S3 and S4). PC1 (corresponding EV=74.78) presented three distinct clusters, CEU and TSI in one cluster, ASW and MKK in the intermediate cluster and LWK, YRI and BMW in the third cluster. PC2 (corresponding EV=6.42) showed separation of LWK and MKK from ASW, YRI and BMW as well as separation of MKK from CEU and TSI (Supplementary Figure S3). The next eight PCs (EVs ranging from 2.48 to 1.95) were largely driven by the variability in MKK samples

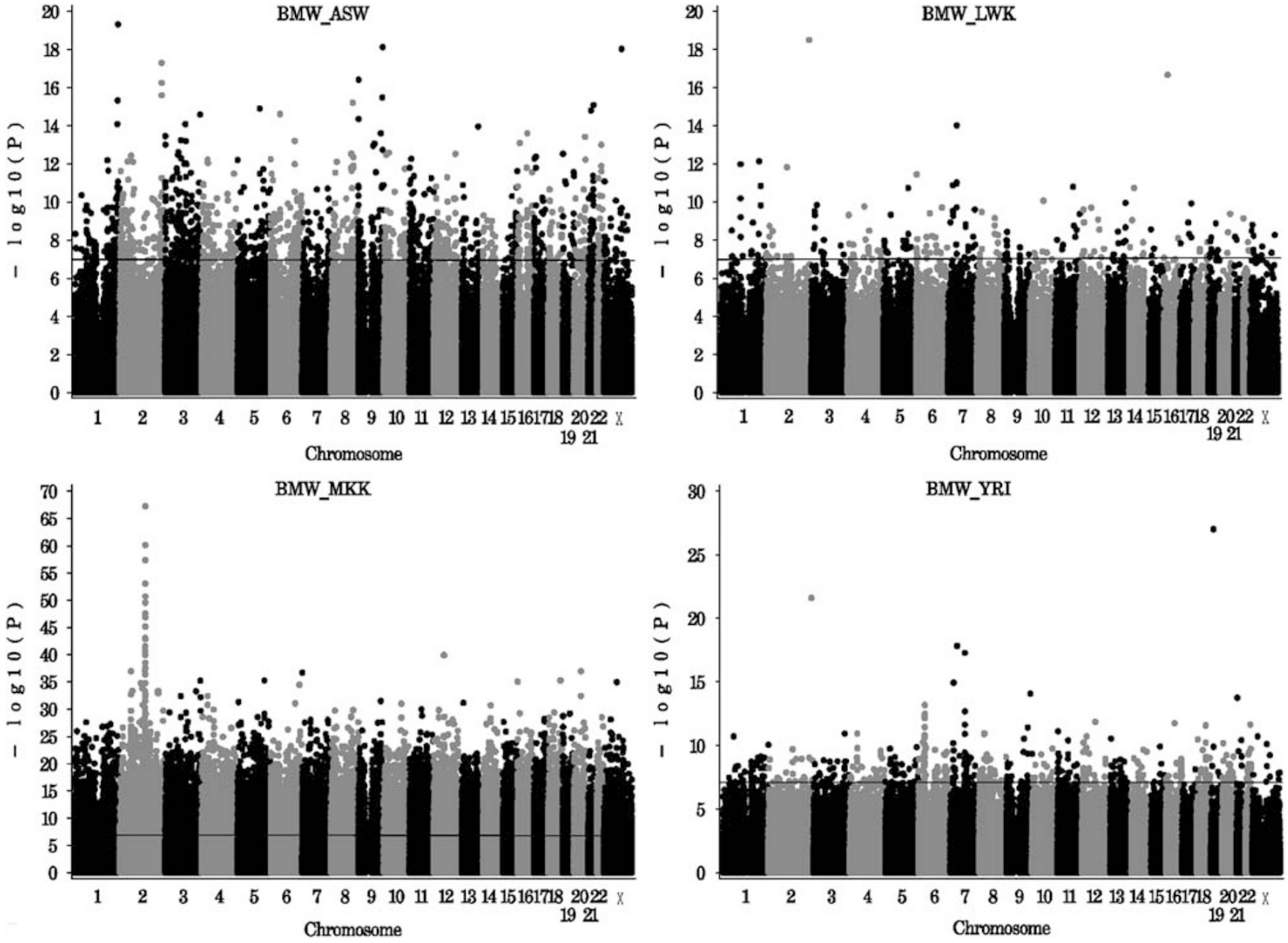

Figure 1 SNP associations with population membership. Individuals in Blantyre, Malawi (BMW) were compared with each African ancestry HapMap population: individuals of African ancestry in the Southwest USA (ASW), Luhya in Webuye, Kenya (LWK), Maasai in Kinyawa, Kenya (MKK) and Yoruban of Ibadan, Nigeria (YRI).

(data not shown); however, PC4 showed clear separation between BMW and YRI samples (Supplementary Figure S4).

Finally, model-based ancestry estimation results using the program *ADMIXTURE* for our sample that included all seven African and European populations are summarized in Table 5 (for K=2, 3 and 5 postulated ancestral populations) and graphically presented in a triangle plot for K=3 postulated ancestral populations in Supplementary Figure S5. For K=2, individual (data not shown) and summary measures for each population suggest a strong clustering of the European populations in one ancestry population and three African populations (LWK, BMW and YRI) in the other ancestry population. ASW and MKK were largely indistinguishable and had proportions for both putative ancestry populations that were between the European and other African populations. When expanding the number of ancestral populations to K=3, we noted complete separation between ASW and MKK, with MKK dominating one ancestral population, the Europeans dominating another ancestral population and two African populations (BMW and YRI) defining the extremes of the third ancestral population. ASW is positioned between the European and two African (BMW and YRI) populations whereas LWK is clearly defined by its own cluster positioned between MKK and the two African populations (BMW and YRI). The resulting triangle plot (Supplementary Figure S5) looks similar to the PC plot of PC1 vs PC2 (Supplementary Figure S3). Expanding the number of postulated ancestral populations to K=5 provided clear separation between all populations and evidence for separation between individual members of MKK. Interestingly, ASW subjects were estimated to be a mixture of the three postulated ancestries most closely linked to YRI, BMW and CEU/TSI, respectively, even after effectively separating YRI from BMW.

DISCUSSION

The aim of this work was to compare genetic variation of HapMap populations of AFA to a population from BMW. We also present some results contrasting Malawians and HapMap populations of other ancestry. HapMap populations were first compared to Malawians based on observed allele frequencies and adjacent SNP–SNP LD r^2 values across the autosomal genome. Allele frequency correlations were strong between the Malawi population (BMW) and the HapMap populations of AFA, with Spearman's $r^2 > 90\%$. However, it was observed that the Malawi population appeared to be more closely related to the individuals of AFA in the Southwest United States (ASW) than they were with the Maasai in Kinyawa, Kenya (MKK). This was interesting, as we expected the European ancestry in ASW subjects would dilute the correlation of allele frequencies between the Malawi and ASW samples. The stronger contrast in allele frequencies between BMW and MKK compared with BMW and other AFA populations is likely a reflection of the variable regional ancestry

Table 4 Top 10 outlier SNPs for each comparison of population from BMW vs HapMap populations of African ancestry[a]

Comparison	*CHR*	*SNP*	*POS*	*MAF1*	*MAF2*	*UNADJ*	*BONF*	*Gene*
BMW vs MKK	2	rs6430594	136435643	0.69	0.08	5.31E−68	2.64E−62	Aspartyl-tRNA synthetase (*DARS*)
	2	rs12472293	136364547	0.11	0.70	7.30E−61	3.63E−55	NA
	2	rs309143	136430648	0.78	0.18	4.70E−58	2.34E−52	Aspartyl-tRNA synthetase (*DARS*)
	2	rs3769013	136272652	0.81	0.23	8.33E−54	4.14E−48	Lactase (*LCT*)
	2	rs730005	136299164	0.71	0.16	2.13E−51	1.06E−45	Lactase (*LCT*)
	2	rs3769012	136272950	0.71	0.16	2.73E−50	1.36E−44	Lactase (*LCT*)
	2	rs961360	136110128	0.70	0.17	2.81E−48	1.40E−42	R3H domain containing 1 (*R3HDM1*)
	2	rs6430585	136223397	0.15	0.70	1.35E−47	6.72E−42	UBX domain protein 4 (*UBXN4*)
	2	rs3806502	136004743	0.73	0.21	6.98E−46	3.47E−40	R3H domain containing 1 (*R3HDM1*)
	2	rs2305248	135644782	0.72	0.21	9.44E−44	4.69E−38	RAB3 GTPase activating protein subunit 1 (catalytic) (*RAB3GAP1*)
BMW vs YRI	19	rs2190687	14765415	0.50	0.06	9.26E−28	4.68E−22	NA
	2	rs6733349	231976556	0.44	0.06	2.48E−22	1.25E−16	NA
	7	rs1717725	38071558	0.05	0.30	1.50E−18	7.59E−13	NA
	7	rs6944302	79942827	0.49	0.17	5.20E−18	2.62E−12	Guanine nucleotide binding protein, α transducing 3 (*GNAT3*)
	7	rs12700014	18930601	0.54	0.23	1.19E−15	6.00E−10	Histone deacetylase 9 (*HDAC9*)
	9	rs3739821	129742298	0.32	0.08	8.61E−15	4.35E−09	Family with sequence similarity 102, member A (*FAM102A*)
	21	rs494619	18347077	0.35	0.07	1.84E−14	9.27E−09	NA
	6	rs2301220	33146744	0.26	0.57	6.68E−14	3.38E−08	Major histocompatibility complex, class II, DP α 1 (*HLA-DPA1*)
	7	rs10216027	79968467	0.30	0.08	2.12E−13	1.07E−07	Guanine nucleotide binding protein, α transducing 3 (*GNAT3*)
	6	rs6457713	33185754	0.33	0.63	3.33E−13	1.68E−07	NA
BMW vs LWK	2	rs6733349	231976556	0.44	0.05	3.22E−19	1.62E−13	NA
	16	rs1017228	21971218	0.31	0.06	2.16E−17	1.08E−11	Chromosome 16 open reading frame 52 (*C16orf52*)
	7	rs11772387	48019075	0.29	0.06	9.83E−15	4.95E−09	Sad1 and UNC84 domain containing 1 (*SUNC1*)
	1	rs2236906	208038108	0.58	0.27	7.48E−13	3.77E−07	Interferon regulatory factor 6 (*IRF6*)
	1	rs4304614	107256813	0.25	0.55	1.08E−12	5.45E−07	NA
	2	rs3789106	111437355	0.28	0.07	1.54E−12	7.77E−07	Acyl-coenzyme A oxidase-like (*ACOXL*)
	6	rs1572438	803970	0.16	0.42	3.71E−12	1.87E−06	NA
	7	rs1915960	48011003	0.07	0.27	9.55E−12	4.80E−06	Sad1 and UNC84 domain containing 1 (*SUNC1*)
	7	rs10248243	47478639	0.05	0.24	1.09E−11	5.49E−06	Tensin 3 (*TNS3*)
	7	rs983186	27155184	0.06	0.25	1.36E−11	6.84E−06	*LOC100133311*, similar to *hCG1644697*. No known function.
BMW vs ASW	1	rs12030126	234879762	0.05	0.39	4.84E−20	2.43E−14	NA
	9	rs7020021	132242790	0.48	0.10	7.63E−19	3.83E−13	Hemicentin 2 (*HMCN2*)
	23	rs226711	98339530	0.08	0.49	9.45E−19	4.74E−13	NA
	2	rs282268	224628420	0.06	0.39	5.16E−18	2.59E−12	NA
	9	rs7045276	191644	0.07	0.39	3.87E−17	1.94E−11	NA
	2	rs282273	224631266	0.05	0.35	5.75E−17	2.89E−11	NA
	2	rs2577284	224637656	0.07	0.39	2.56E−16	1.28E−10	NA
	9	rs3739821	129742298	0.42	0.08	3.32E−16	1.66E−10	Family with sequence similarity 102, member A (*FAM102A*)
	1	rs6586395	232715442	0.10	0.45	4.77E−16	2.39E−10	NA
	8	rs7003117	115759827	0.08	0.39	6.30E−16	3.16E−10	NA

Abbreviations: ASW, African ancestry in Southwest USA; BMW, Individuals of various self-reported ancestry in Blantyre, Malawi; LWK, Luhya in Webuye, Kenya; MKK, Maasai in Kinyawa, Kenya; YRI, Yoruba in Ibadan, Nigeria.

[a] BONF, Bonferroni adjusted *P*-value; CHR, chromosome; Gene, official gene name (symbol); MAF1, minor allele frequency in the Malawi population; MAF2, minor allele frequency in the comparison population; NA, if SNP not located within a gene; POS, base pair position; UNADJ, unadjusted *P*-value.

within the African continent. The Maasai are classified as a Nilotic population and speak Maa, a Nilo-Saharan language.[22] Thus, they may have stronger ancestral roots from North-Eastern Africa whereas the BMW, YRI and LWK claim origins closer to West-Central Africa. Results from the association analyses were consistent with the findings of the allele frequency comparisons by population, showing that the Malawi population is most similar to the YRI and LWK populations, less similar to the ASW and least similar to the MKK. Local LD patterns, as measured by r^2 for all adjacent SNP pairs, showed a similar pattern as observed when evaluating allele frequencies, with BMW having the most similar SNP–SNP LD values to YRI and LWK, and less similar LD values to ASW and MKK. Still, it is noted that the differences in the correlation of allele frequencies and pair-wise SNP–SNP LD between BMI and MKK or ASW are considerably smaller than the differences between any sample of AFA and any HapMap sample of non-AFA. The differences in allele frequencies and SNP–SNP LD values between the Malawi population and the HapMap populations of other ancestry (that is CEU, TSI, JPT, and so on) were striking. These findings illustrate inter-continental and cross-continental genetic diversity and suggest that care must be taken when assessing generalization of a genetic study; for example, the results of a drug clinical trial.

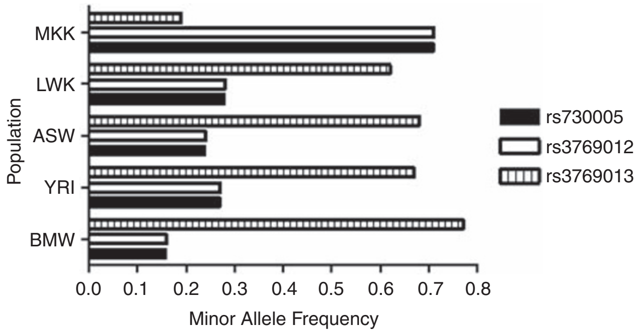

Figure 2 Lactase gene SNP frequencies by African ancestry population. Abbreviations: ASW, African ancestry in Southwest USA; BMW, Individuals of various self-reported ancestry in Blantyre, Malawi; LWK, Luhya in Webuye, Kenya; MKK, Maasai in Kinyawa, Kenya; YRI, Yoruba in Ibadan, Nigeria.

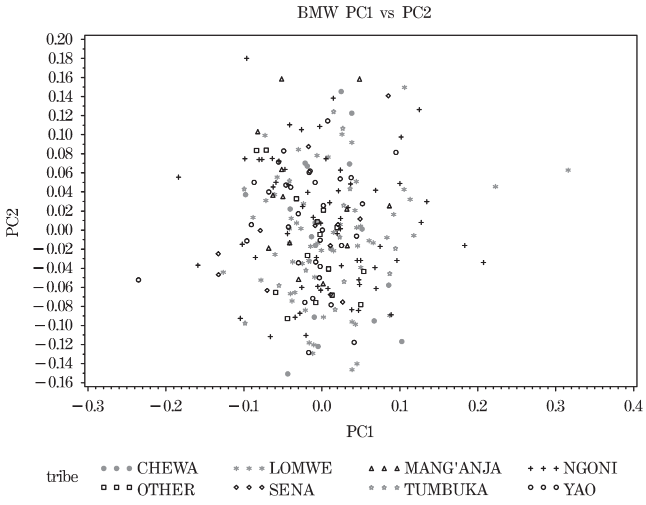

Figure 3 No evidence of population substructure in Malawi population: component 1 vs 2. Analyses performed in *EIGENSOFT* software using 23 612 SNPs.

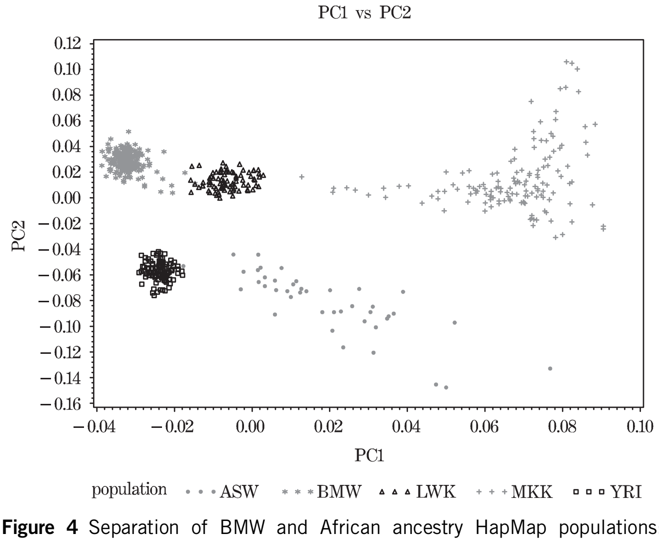

Figure 4 Separation of BMW and African ancestry HapMap populations: component 1 vs 2. Analyses performed in *EIGENSOFT* software using 18 481 SNPs.

A number of relative outlier SNPs identified by the comparison of allele frequencies across AFA populations were of interest with regards to function. This included three SNPs in the comparison of BMW to MKK located in the lactase (*LCT*) gene. This gene encodes a protein that is integral to plasma membrane and has both phlorizin hydrolase activity and lactase activity and has clinical relevance to lactose intolerance.[21] We speculate that the differences between BMW and MKK for the *LCT* gene may be evidence of recent selective pressure. The Maasai are a pastoralist community, relying heavily on the consumption of milk as part of a daily diet. The SNPs discussed here were located within an intron of *LCT*, and functional value was undetermined.

Our final analyses assessed population substructure by using PC and model-based admixture analyses. The Malawi sample itself was found to be reasonably homogenous, exemplified by the high correlation in allele frequencies by self-reported ethnicity and by the lack of apparent population substructure. The generalizability of these findings to the rest of Malawi is unclear, as this population was ascertained in Blantyre, the second largest city. Intermarriage or multiple ethnicities per household may be more common in urban areas compared with rural areas of Malawi. It is also important to note that ethnicity was reported by mothers in the study, and genotyping was conducted in the infants. Thus, the ethnicity was known to be a surrogate for infant ethnicity in this study. As no separation was observed, we would not expect information on ethnicity from the father or any other family members to alter the findings that this group was very genetically homogenous. It should also be noted that accounting for infant HIV status did not alter our conclusions (data not shown).

The PC and model-based admixture analyses showed that the Malawi population was genetically distinct from the other AFA populations. In fact, all five AFA populations (BMW, ASW, LWK, MKK and YRI) could be distinguished by three PCs. Interestingly, our results suggest that the Malawi population from Blantyre are possibly more similar genetically to Yorubans in Ibadan, Nigeria than to the Luhya in Webuye, Kenya despite the substantially larger geographical distance between Malawi and Nigeria. The subjects from Blantyre are clearly genetically more similar to both the Yorubans and the Luhya than to the Maasai in Kenyawa, Kenya. These findings would appear to violate Malicot's isolation-by-distance model, which predicts genetic similarity between populations will decrease exponentially as the geographic distance between them increases.[23] However, consideration of migratory history and corresponding ancestral populations may explain our observations. It is believed that most tribes in Malawi are descendants of the mass Bantu migration from West-Central Africa between the 10th and 15th centuries, and have predominantly Niger-Kordofanian ancestry.[24] In contrast, the Masaai are thought to have migrated down from the Nile region in Egypt in the middle of the 15th century, and considered of Nilo-Saharan origin.[24] The ancestral origin(s) of the Luhya tribes are disputed, with their own oral history suggesting they migrated from Egypt though historians generally believe the Luhya tribes migrated from West-Central Africa alongside other Bantu tribes.

Although considerable efforts were made to include stringent quality control, there was a potential for batch effects in genotype calling between the Malawi and HapMap samples and between the HapMap samples themselves. It is conceivable that such batch effects contributed to the differences in allele frequencies observed between the populations. To determine whether or not this was the case, additional genotyping of the Malawi population simultaneously with HapMap samples and/or replication of these findings in other Malawi populations are necessary.

Table 5 Admixture analyses for clusters of size *K*=2, 3 and 5 with reported means (standard deviations) and [ranges] by ancestral population

	CEU (N=*109)*	*TSI* (N=*77)*	*ASW* (N=*42)*	*YRI* (N=*108)*	*BMW* (N=*226)*	*LWK* (N=*83)*	*MKK* (N=*143)*
K=2							
1	0.014 (0.010)	0.028 (0.008)	0.733 (0.088)	0.930 (0.008)	0.959 (0.014)	0.900 (0.011)	0.714 (0.034)
	[0.000,0.057]	[0.010,0.049]	[0.457,0.906]	[0.909,0.948]	[0.865,0.984]	[0.869,0.926]	[0.617,0.834]
2	0.986 (0.010)	0.972 (0.008)	0.267 (0.088)	0.070 (0.008)	0.041 (0.014)	0.100 (0.011)	0.286 (0.034)
	[0.943,1.000]	[0.951,0.990]	[0.094,0.543]	[0.052,0.091]	[0.016,0.135]	[0.074,0.131]	[0.166,0.383]
K=3							
1	0.012 (0.013)	0.050 (0.013)	0.098 (0.024)	0.095 (0.023)	0.070 (0.021)	0.249 (0.037)	0.678 (0.124)
	[0.000,0.056]	[0.007,0.075]	[0.047,0.143]	[0.037,0.151]	[0.019,0.142]	[0.177,0.321]	[0.323,0.953]
2	0.017 (0.010)	0.006 (0.008)	0.663 (0.081)	0.858 (0.017)	0.905 (0.019)	0.721 (0.033)	0.242 (0.107)
	[0.000,0.039]	[0.000,0.032]	[0.411,0.831]	[0.814,0.896]	[0.834,0.947]	[0.655,0.785]	[0.035,0.605]
3	0.971 (0.013)	0.945 (0.010)	0.239 (0.089)	0.047 (0.011)	0.025 (0.016)	0.030 (0.010)	0.080 (0.034)
	[0.919,0.997]	[0.924,0.967]	[0.068,0.518]	[0.022,0.075]	[0.000,0.131]	[0.010,0.057]	[0.000,0.206]
K=5							
1	0.010 (0.011)	0.006 (0.010)	0.442 (0.062)	0.667 (0.045)	0.132 (0.044)	0.224 (0.044)	0.089 (0.050)
	[0.000,0.047]	[0.000,0.038]	[0.301,0.600]	[0.549,0.776]	[0.000,0.278]	[0.112,0.329]	[0.000,0.194]
2	0.007 (0.011)	0.036 (0.016)	0.055 (0.029)	0.042 (0.023)	0.043 (0.021)	0.202 (0.037)	0.601 (0.181)
	[0.000,0.046]	[0.003,0.075]	[0.000,0.127]	[0.000,0.106]	[0.000,0.010]	[0.133,0.290]	[0.000,1.000]
3	0.012 (0.013)	0.003 (0.007)	0.256 (0.048)	0.248 (0.042)	0.774 (0.045)	0.495 (0.050)	0.106 (0.097)
	[0.000,0.045]	[0.000,0.030]	[0.136,0.382]	[0.157,0.357]	[0.626,0.918]	[0.395,0.631]	[0.000,0.499]
4	0.963 (0.012)	0.938 (0.10)	0.219 (0.091)	0.019 (0.010)	0.021 (0.015)	0.019 (0.010)	0.066 (0.035)
	[0.912,0.989]	[0.917,0.957]	[0.042,0.503]	[0.000,0.044]	[0.000,0.125]	[0.001,0.047]	[0.000,0.189]
5	0.008 (0.009)	0.017 (0.012)	0.028 (0.016)	0.025 (0.014)	0.031 (0.014)	0.060 (0.013)	0.138 (0.160)
	[0.000,0.035]	[0.000,0.043]	[0.000,0.063]	[0.000,0.066]	[0.000,0.080]	[0.027,0.096]	[0.000,1.000]

Abbreviations: ASW, African ancestry in Southwest USA; BMW, Individuals of various self-reported ancestry in Blantyre, Malawi; CEU, Utah residents with Northern and Western European ancestry from the CEPH collection; LWK, Luhya in Webuye, Kenya; MKK, Maasai in Kinyawa, Kenya; TSI, Toscans in Italy; YRI, Yoruba in Ibadan, Nigeria.

This study showed that the Malawi population in Blantyre does not exhibit strong genetic variability by self-reported ethnicity. We also showed that although highly correlated with regards to allele frequencies and adjacent SNP–SNP LD, the Malawi population and populations of AFA in the International HapMap Project are genetically distinct. Furthermore, we determined that the allele frequencies and LD are not strongly correlated between Malawi and the HapMap populations of non-AFA. The discordance in genetic variation observed both within and across continental lines highlights the necessity for researchers to consider ancestry and always account for population stratification. It also suggests that such differences should be taken into account when predicting drug and vaccine efficacy for patients across the African continent. Future work may involve more fine-tuning of these results, including projects to sequence specific regions of interest in BMW subjects followed by a comparison of sequence variation by population.

ACKNOWLEDGEMENTS

The genotyping in the Malawi population completed at Duke University was in collaboration with David Goldstein and Kevin Shianna. This work was supported by the NIAID Center for HIV/AIDS Vaccine Initiative (CHAVI) Grant AI067854. The lead author also received support from the Centers for Disease Control and Prevention Dissertation Award (PAR 231-07, 2008).

1 Consortium, I. H. The International HapMap Project. *Nature* **426,** 789–796 (2003).
2 Montpetit, A., Nelis, M., Laflamme, P., Magi, R., Ke, X., Remm, M. *et al.* An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population. *PLoS Genet.* **2,** e27 (2006).
3 Marvelle, A. F., Lange, L. A., Qin, L., Wang, Y., Lange, E. M., Adair, L. S. *et al.* Comparison of ENCODE region SNPs between Cebu Filipino and Asian HapMap samples. *J. Hum. Genet.* **52,** 729–737 (2007).
4 Arnold, J. C., Singh, K. K., Spector, S. A. & Sawyer, M. H. Undiagnosed respiratory viruses in children. *Pediatrics* **121,** e631–e637 (2008).
5 Ribas, G., Gonzalez-Neira, A., Salas, A., Milne, R. L., Vega, A., Carracedo, B. *et al.* Evaluating HapMap SNP data transferability in a large-scale genotyping project involving 175 cancer-associated genes. *Hum. Genet.* **118,** 669–679 (2006).
6 Mueller, J. C., Lohmussaar, E., Magi, R., Remm, M., Bettecken, T., Lichtner, P. *et al.* Linkage disequilibrium patterns and tagSNP transferability among European populations. *Am. J. Hum. Genet.* **76,** 387–398 (2005).
7 Willer, C. J., Scott, L. J., Bonnycastle, L. L., Jackson, A. U., Chines, P., Pruim, R. *et al.* Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. *Genet. Epidemiol.* **30,** 180–190 (2006).
8 Smith, E. M., Wang, X., Littrell, J., Eckert, J., Cole, R., Kissebah, A. H. *et al.* Comparison of linkage disequilibrium patterns between the HapMap CEPH samples and a family-based cohort of Northern European descent. *Genomics* **88,** 407–414 (2006).
9 Hu, C., Jia, W., Zhang, W., Wang, C., Zhang, R., Wang, J. *et al.* An evaluation of the performance of HapMap SNP data in a Shanghai Chinese population: analyses of allele frequency, linkage disequilibrium pattern and tagging SNPs transferability on chromosome 1q21–q25. *BMC Genet.* **9,** 19 (2008).
10 Evans, D. M. & Cardon, L. R. A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. *Am. J. Hum. Genet.* **76,** 681–687 (2005).
11 Mwapasa, V., Rogerson, S. J., Kwiek, J. J., Wilson, P. E., Milner, D., Molyneux, M. E. *et al.* Maternal syphilis infection is associated with increased risk of mother-to-child transmission of HIV in Malawi. *AIDS* **20,** 1869–1877 (2006).
12 Mwapasa, V., Rogerson, S. J., Molyneux, M. E., Abrams, E. T., Kamwendo, D. D., Lema, V. M. *et al.* The effect of Plasmodium falciparum malaria on peripheral and placental HIV-1 RNA concentrations in pregnant Malawian women. *AIDS* **18,** 1051–1059 (2004).
13 Joubert, B. R., Lange, E. M., Franceschini, N., Mwapasa, V., North, K. E., Meshnick, S. R. *et al.* A whole genome association study of mother-to-child transmission of HIV in Malawi. *Genome Med.* **2** (2010), http://genomemedicine.com/content/2/3/17.
14 Kaspin, D. The politics of ethnicity in Malawi's democratic transition. *J. Mod. Afr. Stud.* **33,** 595–620 (1995).
15 Thorisson, G. A., Smith, A. V., Krishnan, L. & Stein, L. D. The International HapMap Project Web site. *Genome Res.* **15,** 1592–1593 (2005).

16 Fellay, J., Shianna, K. V., Ge, D., Colombo, S., Ledergerber, B., Weale, M. *et al.* A whole-genome association study of major determinants for host control of HIV-1. *Science* **317,** 944–947 (2007).

17 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

18 StataCorp. *Stata Statistical Software: Release 10* (StataCorp LP, College Station, TX, 2007).

19 Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** e190 (2006).

20 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19,** 1655–1664 (2009).

21 Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **37** (Database issue), D5–D15 (2009).

22 NHGRI. *NHGRI Sample Repository for Human Genetic Research* (The Coriell Institute for Medical Research, Camden, 2009); Available from: http://ccr.coriell.org/sections/collections/NHGRI/?SsId=11.

23 Harpending, H. C & Ward, R. H *Chemical Systematics and Human Populations. Biochemical Aspects of Evolutionary Biology* 213–246 (University of Chicago Press, Chicago, 1981).

24 Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A. *et al.* The genetic structure and history of Africans and African Americans. *Science* **324,** 1035–1044 (2009).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)