nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-019-0947-6

Natural selection contributed to immunological differences between hunter-gatherers and agriculturalists

Genelle F. Harrison[1,2], Joaquin Sanz[2,3], Jonathan Boulais[2,3], Michael J. Mina[4,5], Jean-Christophe Grenier[2], Yumei Leng[5], Anne Dumaine[2], Vania Yotova[2], Christina M. Bergey[6], Samuel L. Nsobya[7], Stephen J. Elledge[5], Erwin Schurr[1], Lluis Quintana-Murci[8,9,10], George H. Perry[6,11,14] and Luis B. Barreiro[2,12,13,14]*

The shift from a hunter-gatherer to an agricultural mode of subsistence is believed to have been associated with profound changes in the burden and diversity of pathogens across human populations. Yet, the extent to which the advent of agriculture affected the evolution of the human immune system remains unknown. Here we present a comparative study of variation in the transcriptional responses of peripheral blood mononuclear cells to bacterial and viral stimuli between Batwa rainforest hunter-gatherers and Bakiga agriculturalists from Uganda. We observed increased divergence between hunter-gatherers and agriculturalists in the early transcriptional response to viruses compared with that for bacterial stimuli. We demonstrate that a significant fraction of these transcriptional differences are under genetic control and we show that positive natural selection has helped to shape population differences in immune regulation. Across the set of genetic variants underlying inter-population immune-response differences, however, the signatures of positive selection were disproportionately observed in the rainforest hunter-gatherers. This result is counter to expectations on the basis of the popularized notion that shifts in pathogen exposure due to the advent of agriculture imposed radically heightened selective pressures in agriculturalist populations.

The agricultural transition, beginning 10,000–12,000 bp, was associated with profound changes in human ecology[1], which in turn are suggested to have precipitated major infectious disease burdens[2–4]. Specifically, the construction of permanent settlements and a subsequent increase in population density associated with the agricultural transition[5,6] may have facilitated the establishment and transmission of infectious agents such as smallpox, measles, rubella and other pathogens that require hundreds to thousands of host individuals to spread and persist[7,8]. Agriculturalists and pastoralists also lived alongside their domesticated animals, providing opportunity for new or expanded zoonotic transmission[4] of pathogens potentially including rotavirus, measles virus and influenza[9–11]. Finally, agriculturists performed extensive modifications to the landscape, including clearing fields and constructing irrigation systems, which may have led to an increase in the incidence of vector-borne diseases, such as *Plasmodium falciparum* malaria[12,13]. In several instances, higher intestinal parasite burdens in agricultural (AG) relative to hunter-gatherer (HG) populations have also been reported[14].

Consequentially, the transition to an agriculturalist lifestyle is suggested to have contributed to the strong genetic signatures of recent positive selection that are repeatedly observed in or nearby immune-related genes in worldwide AG populations[15,16]. However, the absence of comparative functional studies from pairs of populations that differ in their modes of subsistence, that is, HG versus AG, have thus far precluded the development of hypotheses concerning specifically how the agricultural transition may have affected evolution of human immune system diversity. To begin studying this topic, we used a combination of evolutionary genomic and functional immunological tools to study differences in immune responses between the Batwa, a rainforest HG population from southwest Uganda and their Bantu-speaking AG neighbours, the Bakiga.

Results

Significant Batwa–Bakiga immune response differences. Whole blood samples from 103 individuals (59 HG-Batwa and 44 AG-Bakiga; Supplementary Fig. 1) were collected and peripheral blood mononuclear cells (PBMCs) from these samples were isolated and cryopreserved. PBMCs were collected and processed for both populations simultaneously during the same field expedition to minimize technical variability. Each individual was genotyped for ~1 million genome-wide single nucleotide polymorphisms (SNPs)[17], with additional imputation to 10,530,212 SNP genotypes (see Methods). These data were used to estimate genome-wide levels of HG-Batwa and AG-Bakiga ancestry, using the program

[1]Department of Human Genetics, Faculty of Medicine, McGill University, Montreal, Quebec, Canada. [2]Department of Genetics, CHU Sainte-Justine Research Center, Montreal, Quebec, Canada. [3]Department of Biochemistry, Faculty of Medicine, Université de Montréal, Montréal, Quebec, Canada. [4]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [5]Department of Genetics, Harvard Medical School and Division of Genetics, Brigham and Women's Hospital, Howard Hughes Medical Institute, Boston, MA, USA. [6]Departments of Anthropology and Biology, Pennsylvania State University, University Park, PA, USA. [7]Department of Pathology, School Biomedical, Makerere University, Kampala, Uganda. [8]Unit of Human Evolutionary Genetics, Institut Pasteur, Paris, France. [9]Centre National de la Recherche Scientifique, UMR2000, Paris, France. [10]Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, Paris, France. [11]Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA, USA. [12]Department of Pediatrics, University of Montreal, Montreal, Quebec, Canada. [13]Department of Medicine, Section of Genetic Medicine, University of Chicago, Chicago, IL, USA. [14]These authors contributed equally: George H. Perry, Luis B. Barreiro. *e-mail: lbarreiro@uchicago.edu

ADMIXTURE (ref. [18]). We observed variable but considerable levels of AG-Bakiga ancestry among self-identified HG-Batwa individuals (mean 21.0%; range 0–93.3%). However, estimated levels of HG-Batwa ancestry among self-identified AG-Bakiga individuals were typically lower (mean 4.3%; range 0–9.7%; Fig. 1a). In what follows, we used these continuous estimates of genetic ancestry (as opposed to a binary classification of individuals into HG-Batwa versus AG-Bakiga ancestry) to identify ancestry-associated variation in gene expression and other immune-related traits.

To characterize variation in the immune response between HG-Batwa and AG-Bakiga populations we exposed PBMCs to: gardiquimod (GARD, TLR7 agonist), which mimics an infection with a single-stranded RNA virus; and to lipopolysaccharide (LPS, TLR4 agonist), which simulates an infection with Gram-negative bacteria. We also maintained an unexposed control in the same experimental conditions (CTL). Following 4 h of stimulation, we collected RNA-sequencing data from matched non-stimulated and stimulated PBMCs (Fig. 1a). Following quality control filtering, we analysed high-quality RNA-sequencing profiles ($n = 229$ RNA-sequencing profiles across treatment combinations) from 99 individuals (57 HG-Batwa and 42 AG-Bakiga; see Methods, Supplementary Fig. 1 and Supplementary Table 1). To confirm successful ligand stimulation, we performed a principal component analysis on the correlation matrix of normalized gene-expression levels for all conditions. The first principal component (PC) explained 51.1% of the variance in the expression values and effectively separated the LPS condition from an unstimulated control (CTL). The combination of the second and third principal components further separated the GARD-stimulated PBMCs from the CTL cells (Fig. 1c). As expected, the set of genes upregulated in response to both stimuli were significantly enriched (false discovery rate (FDR) $< 1 \times 10^{-15}$) for genes known to be involved in immune defence and inflammatory responses, with a particularly strong enrichment for antiviral response genes in the GARD condition (Supplementary Table 3).

Because PBMCs are a composite of various innate and adaptive immunity cell types, we first determined whether there were differences in the cellular compositions of PBMCs between the HG-Batwa and AG-Bakiga. Using fluorescence-activated cell sorting we estimate the proportion of each of the main cell types comprising PBMCs for every individual (Supplementary Fig. 2). We found that the proportion of $CD14^+$ monocytes was higher in individuals with greater HG-Batwa ancestry ($P = 4.9 \times 10^{-08}$), while the proportion of $CD3^+/CD4^+$ helper T cells was higher in individuals with greater AG-Bakiga ancestry ($P = 8.2 \times 10^{-06}$; Fig. 1b). Using linear models that account for variation in cell composition, sex and additional technical covariates, we next identified genes whose expression levels were linearly correlated with ancestry in each of the experimental conditions: population differentially expressed (PopDE) genes. Of the 10,885 expressed genes tested, 1,836 genes (16.9% of the total) were found to be PopDE (FDR < 0.05) in at least one condition (Fig. 1d with Fig. 1e for an example). Among PopDE genes, genetic ancestry explains, on average, 14.4% (quantile 5%–95% confidence interval 6.8–25.1) of the overall variance in gene expression observed among individuals, an amount comparable to the proportion of variation that can be attributed to differences in cell composition (mean 16.8%; quantile 5%–95% confidence interval 2.9–39.0) and much higher than the proportion explained by sex (mean 3.4%; quantile 5%–95% confidence interval 0.2–9.8; Supplementary Fig. 3).

Gene set enrichment analyses (GSEA) revealed that genes with higher expression levels in HG-Batwa individuals in LPS- and GARD-stimulated PBMCs were markedly enriched in pathways related to interferon-γ and interferon-α responses (FDR $< 1 \times 10^{-4}$; Fig. 1f), the key pathways involved in immune responses to viruses. In contrast, genes with higher expression levels in AG-Bakiga individuals are enriched for inflammatory response genes, particularly in LPS-stimulated PBMCs (FDR $< 1 \times 10^{-4}$; Fig. 1f; see Supplementary Table 3 for a complete list of all enriched pathways). These results suggest that increased AG-Bakiga ancestry is associated with a stronger inflammatory response while individuals with greater HG-Batwa ancestry have gene-expression signatures compatible with increased activation of antiviral pathways.

Viruses were probably the main driver of Batwa–Bakiga immune response differences. Several lines of evidence indicate that the regulation of the immune response to viral stimuli between HG-Batwa and the AG-Bakiga individuals is more divergent compared to that for bacterial stimuli. Among 'stimuli-responsive genes' (the set of genes that exhibit expression changes on LPS- or GARD-stimulation), we identified almost twice as many PopDE genes in the GARD condition as compared to the LPS condition (10.1% of all genes that respond to GARD versus 5.9% of all genes that respond to LPS; chi-squared test, $P < 2.2 \times 10^{-16}$). We considered the set of genes for which the intensity of the response to LPS and GARD—defined as the fold-change in the stimulated condition relative to the unstimulated condition—varied as a function of genetic ancestry. We coined these population differentially responsive (PopDR) genes (see, for example, Fig. 2a). We again observed approximately twice as many PopDR genes (FDR < 0.1) in GARD-stimulated cells compared with LPS-stimulated cells (258 PopDR for GARD versus 140 PopDR for LPS, Fig. 2b). Performing GSEA for PopDR genes also revealed striking enrichments for interferon-related pathways (FDR $< 1 \times 10^{-4}$) among genes that respond stronger to both LPS and GARD in HG-Batwa individuals relative to AG-Bakiga individuals (Supplementary Table 3).

The relatively divergent viral stimuli regulatory response is in part explained by a stronger response to GARD for the HG-Batwa individuals compared to their AG-Bakiga agriculturalist neighbours. Among the PopDR genes, the absolute fold-response to the viral ligand GARD was significantly stronger in the HG than the AG individuals (Fig. 2c; Mann–Whitney–Wilcoxon test $P = 7.74 \times 10^{-32}$), while a similar difference was not observed for LPS (Mann–Whitney–Wilcoxon test $P = 0.34$). Our data thus suggest that differences in viral exposure may have been an important factor contributing to the immune response divergence between the HG-Batwa and the AG-Bakiga.

While we do not have historical records of the viruses encountered by these populations, we can measure antiviral antibodies in present-day populations to gather information about their viral exposure. We used VirScan (ref. [19])—a high-throughput method that allows comprehensive analysis of antiviral antibodies—to measure, in all our samples, serum antibodies to 130 viruses known to be present in Africa (see Methods). In measuring the relative variation of epitope burden found among the 130 viruses tested, we identified antibodies to 35 viruses (27%) whose levels were significantly different (FDR < 0.05) between HG- and AG-ancestry individuals (see Methods). Among these 35 viruses, 32 (91.4%) showed a higher burden (increased seropositivity) in individuals of HG-Batwa ancestry (Fig. 2d and Supplementary Table 4). We observed increased seropositivity for only three viruses, all of which were human-specific single-strand RNA viruses, in the AG individuals. Interestingly, viruses with higher burdens in the HG-Batwa population were significantly enriched for double-stranded DNA viruses (20 of 32 observed; 14 of 31 expected; odds ratio = 3.7 (confidence interval 1.5–9.9); Fig. 2d; Fisher's exact test $P = 2.9 \times 10^{-3}$), compatible with the suggestion that DNA viruses are able to persist more readily in smaller populations than RNA viruses due to longer periods of latency[20–22]. Although the differences reported here may not be indicative of historical exposure, they do support the possibility that rainforest HG and AG populations (at least in southwest Uganda) have faced notable differences in viral exposure, with rainforest HG populations exhibiting a higher viral burden, particularly when considering DNA viruses.

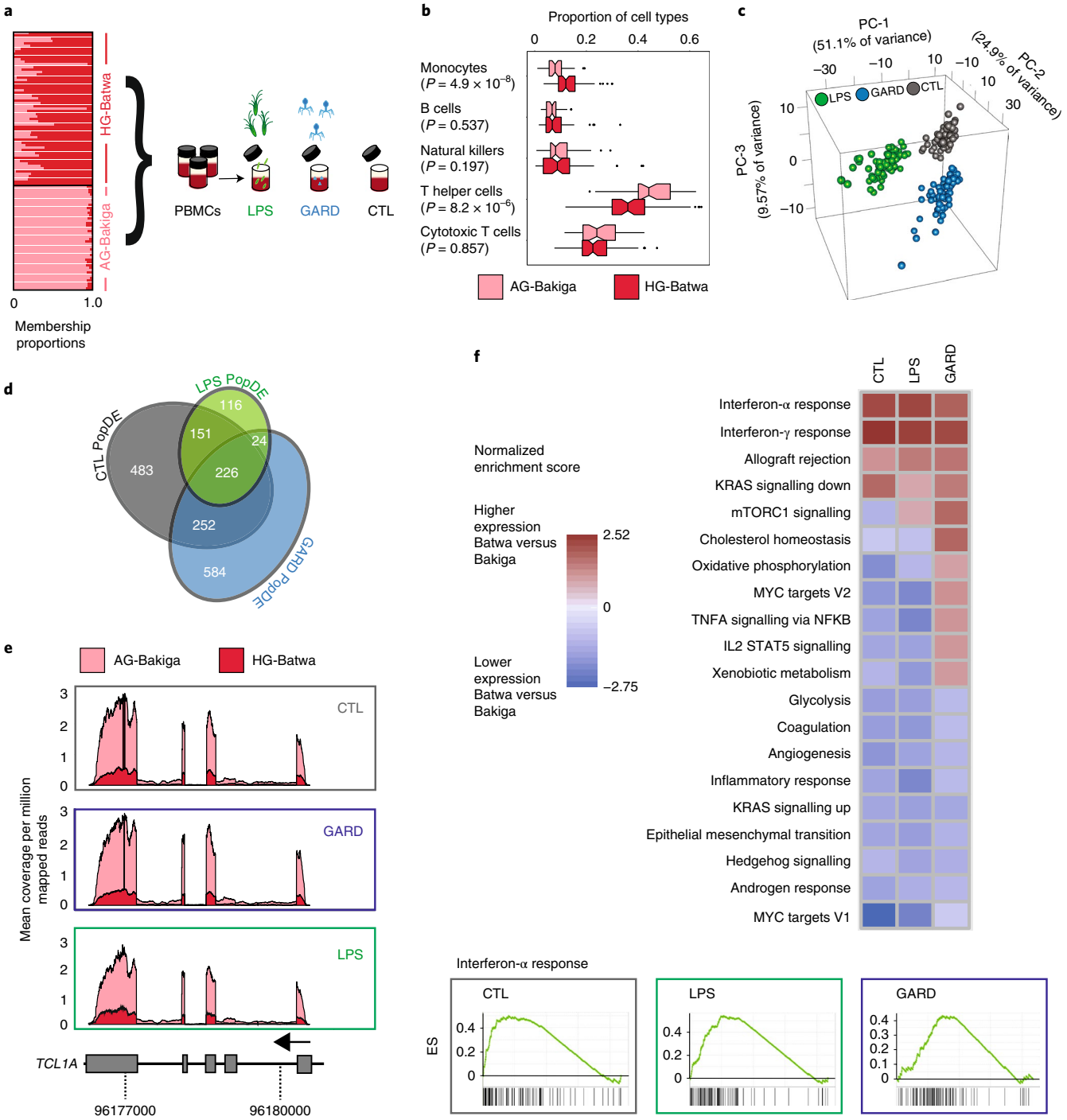

Fig. 1 | Transcriptional differences between Batwa-HG and Bakiga-AG populations. a, Schematics of the study design. The structure plot to the left shows the proportion of HG-ancestry (dark pink) and AG-ancestry (light pink) for each individual included in the study. Their placement along the *y* axis corresponds to how they self-identified. **b**, Boxplots of the proportions of the main cell types found in PBMCs in the Batwa (dark pink) and the Bakiga (light pink). The upper and lower ends of the whiskers correspond to plus or minus 1.5 times the interquartile range, respectively. **c**, Principal component analysis of gene-expression data. The first three principle components (PC) separate non-infected PBMCs from PBMCs stimulated with either LPS or GARD. **d**, Venn diagram of PopDE genes detected in each condition. **e**, Example of a PopDE gene (*TCL1A*) in which gene expression is higher in the AG population (light pink) than the HG population (dark pink) in all conditions. Expression is shown as the mean coverage per genomic position (corrected by total mapped reads) per individual in each population. **f**, GSEA for PopDE genes in all three conditions. The heatmaps show the enrichment scores for all pathways enriched at an FDR < 5% in at least one of the conditions. Positive and negative scores represent enrichments among genes that are more highly or lowly expressed in HG-Batwa than AG-Bakiga individuals, respectively. Example of an enrichment plot for genes involved in the interferon-α response pathway. Genes are ranked (left to right) from those with the strongest statistical evidence for upregulation in the HG-Batwa versus AG-Bakiga to those with the strongest statistical support for downregulation in the HG-Batwa versus AG-Bakiga.

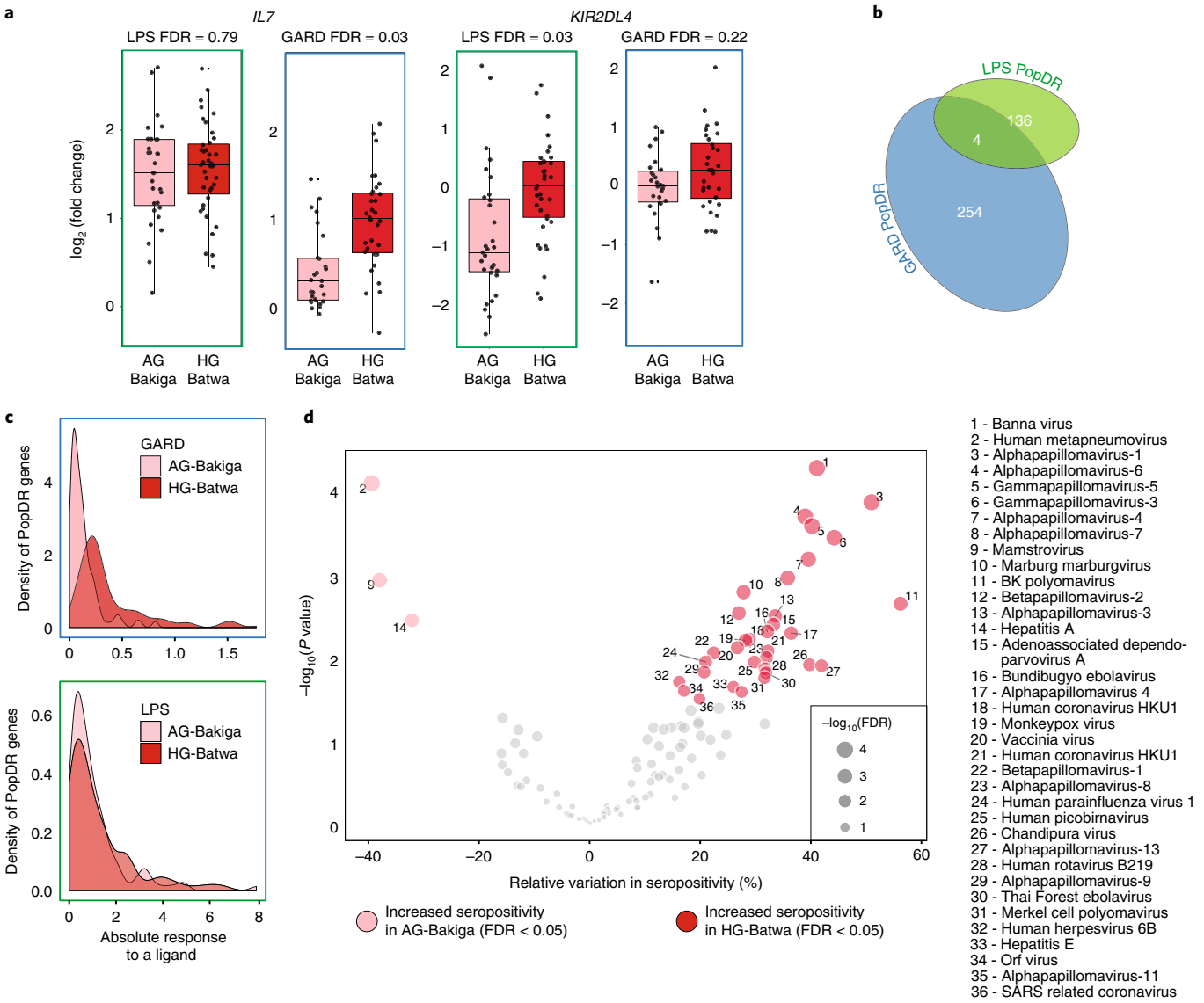

Fig. 2 | Differences in immune response between HG and AG populations. a, Examples of two PopDR genes involved in immune response. The *y* axis shows the $\log_2$ (fold change) in gene-expression levels in response to LPS and GARD, for individuals from each of the two populations (*x* axis). The upper and lower ends of the whiskers correspond to plus or minus 1.5 times the interquartile range, respectively. **b**, Venn diagram showing the number of PopDR genes identified in the LPS and GARD conditions. **c**, Density plots showing the distributions of the absolute response to LPS and GARD of PopDR genes in each population. **d**, A volcano plot showing an increase in seropositivity in the HG-Batwa population for 32 of the 130 viruses tested. Double-stranded DNA viruses showing a notable dependence on ancestry are marked in bold.

Genetic variation contributes to ancestry-associated differences in immune regulation. Next, we aimed to identify components of the HG and AG transcriptional immune response driven by either genetic or environmental factors between HG and AG populations. To limit the effects of unknown confounding factors, we used a linear regression model that accounts for population structure and principal components of the expression data (see Methods). We first identified genetic variants that are associated with differences in gene-expression levels (expression quantitative trail loci, eQTL) in our complete sample. We focused specifically on *cis*-eQTL, which we defined as SNPs located either within or flanking (±100 kilobases, kb) the gene of interest. We identified a total of 3,941 genes (37.6% of all genes tested) that are associated with at least one *cis*-eQTL (FDR < 0.05) in at least one condition. Consistent with previous findings[23–26], a large fraction of *cis*-eQTLs (14.7%) were observed only in stimulated samples (Fig. 3a,b for example), highlighting the key importance of gene–environment interactions to the transcriptional regulation of innate immune responses.

We then tested whether PopDE and PopDR genes were more likely to be influenced by genetic variants than expected by chance. We found that PopDE and PopDR genes were significantly enriched among the set of genes associated with *cis*-eQTLs (>1.6-fold enrichment; $P < 1.0 \times 10^{-10}$; Fig. 3c). These results suggest that the differences in transcriptional responses to viral and bacterial stimuli identified in HG- and AG-ancestry individuals are driven, at least partly, by genetic regulatory variants. To explicitly quantify the minimum contribution of identified *cis*-eQTL to the transcriptional differences detected between populations, we used the following approach. First, we estimated in each condition the proportion of variance explained (PVE) by HG-ancestry among PopDE genes. Then, we re-calculated HG-ancestry PVE after regressing out the effect of the single *cis*-SNP for each gene that

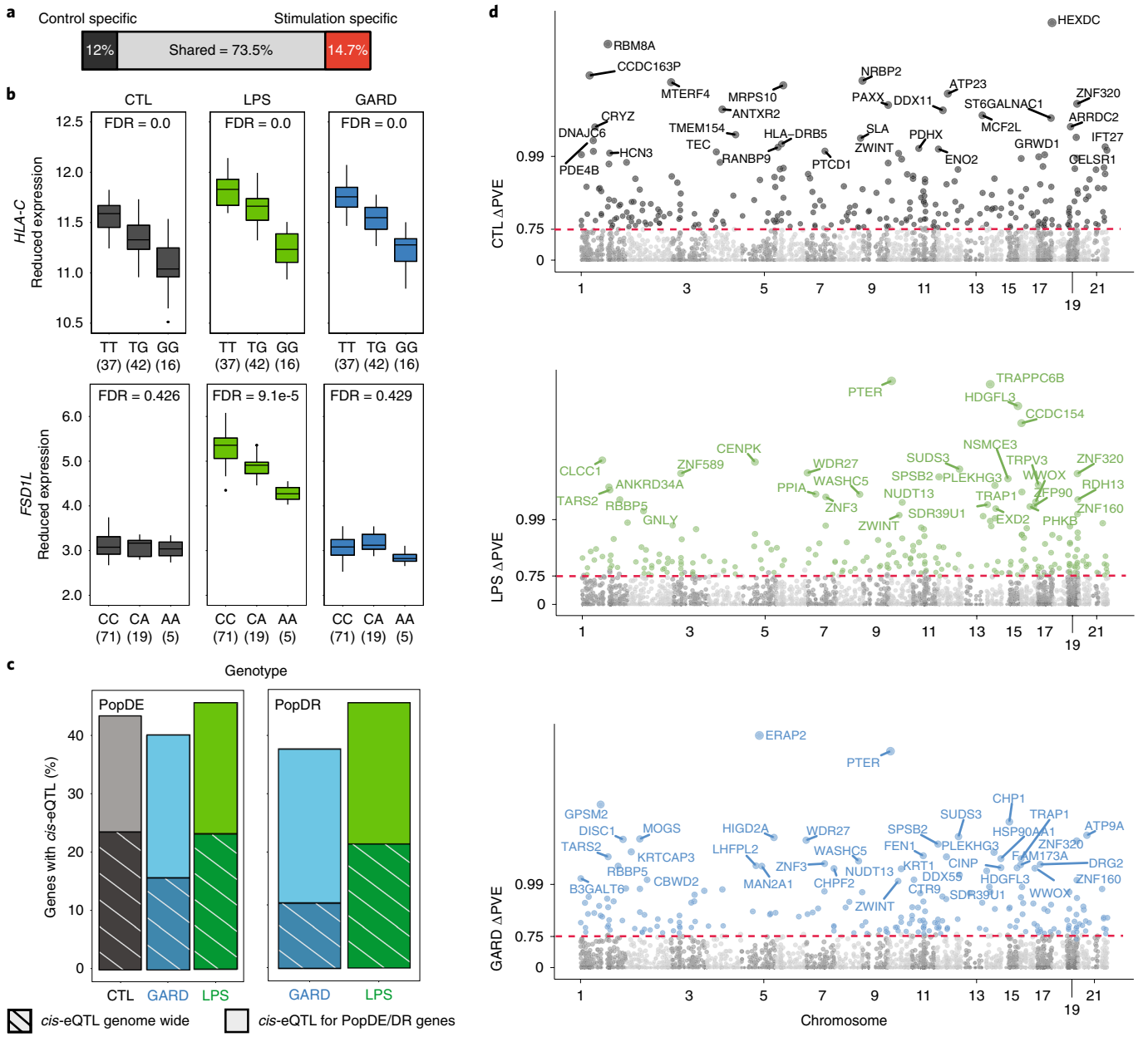

Fig. 3 | Analysis of the contribution of genetics to differences in immune response between the HG-Batwa and the AG-Bakiga. a, Schematic representation of the number of *cis*-eQTL shared across all conditions, or only found in non-infected PBMCs, or found in LPS and/or GARD-stimulated PBMCs (stimulation-specific eQTL). Stimulation-specific eQTL were defined as those showing very strong evidence of eQTL in the stimulated cells (FDR < 0.05) and very limited evidence in the non-infected cells (FDR always higher than 0.25). **b**, Example of two *cis*-eQTL. The top example, *HLA-C*, was found across all experimental condition (CTL-FDR = 0.0, LPS-FDR = 0.0, GARD-FDR = 0.0). The bottom example, fibronectin type III and SPRY domain containing 1 like (*FSD1L*) was detected exclusively in the LPS condition. In this example, expression is in $\log_2$(c.p.m.) (CTL-FDR = 0.426, LPS-FDR = 9.09^{-5}, GARD-FDR = 0.429). The upper and lower ends of the boxplot whiskers correspond to plus or minus 1.5 times the interquartile range, respectively. **c**, Bar graphs showing an enrichment of genes containing *cis*-eQTLs among PopDE/PopDR genes (totality of bars) compared with genome-wide expectations (stripes). **d**, Manhattan plot showing ΔPVE of *cis*-eQTL (normalized as $-\log_{10}(1-\Delta PVE$ for easier viewing) on the *y* axis across all chromosomes for CTL (grey), GARD (blue) and LPS (green). Coloured points have an FDR < 0.1 and a delta-PVE > 0.75. Points are labelled with the corresponding gene name when PVE is > 0.99.

was most strongly associated with the target gene's expression level (the SNP with the lowest FDR, regardless of significance level). The difference between HG-ancestry PVE values before and after regressing out the *cis*-eQTL effect (normalized by the original PVE value) quantifies the proportion of ancestry-associated effects on gene expression that stems from the strongest *cis*-associated variant. Hereafter we refer to this score as ΔPVE. Using this approach, we estimated that *cis*-regulatory variants explain, on average, ~34% of the PopDE signal in each condition (average ΔPVE = 36.7%, 37.5% and 34.2% among PopDE genes (FDR < 0.2) in control, GARD and LPS conditions, respectively; Supplementary Fig. 4). From this analysis, we identified a set of 475 PopDE genes across conditions for which a single *cis*-eQTL is enough to explain almost all ancestry effects on gene-expression levels (ΔPVE > 75%; FDR < 0.1; hereafter referred to as high-ΔPVE variants) on gene-expression levels (Fig. 3d).

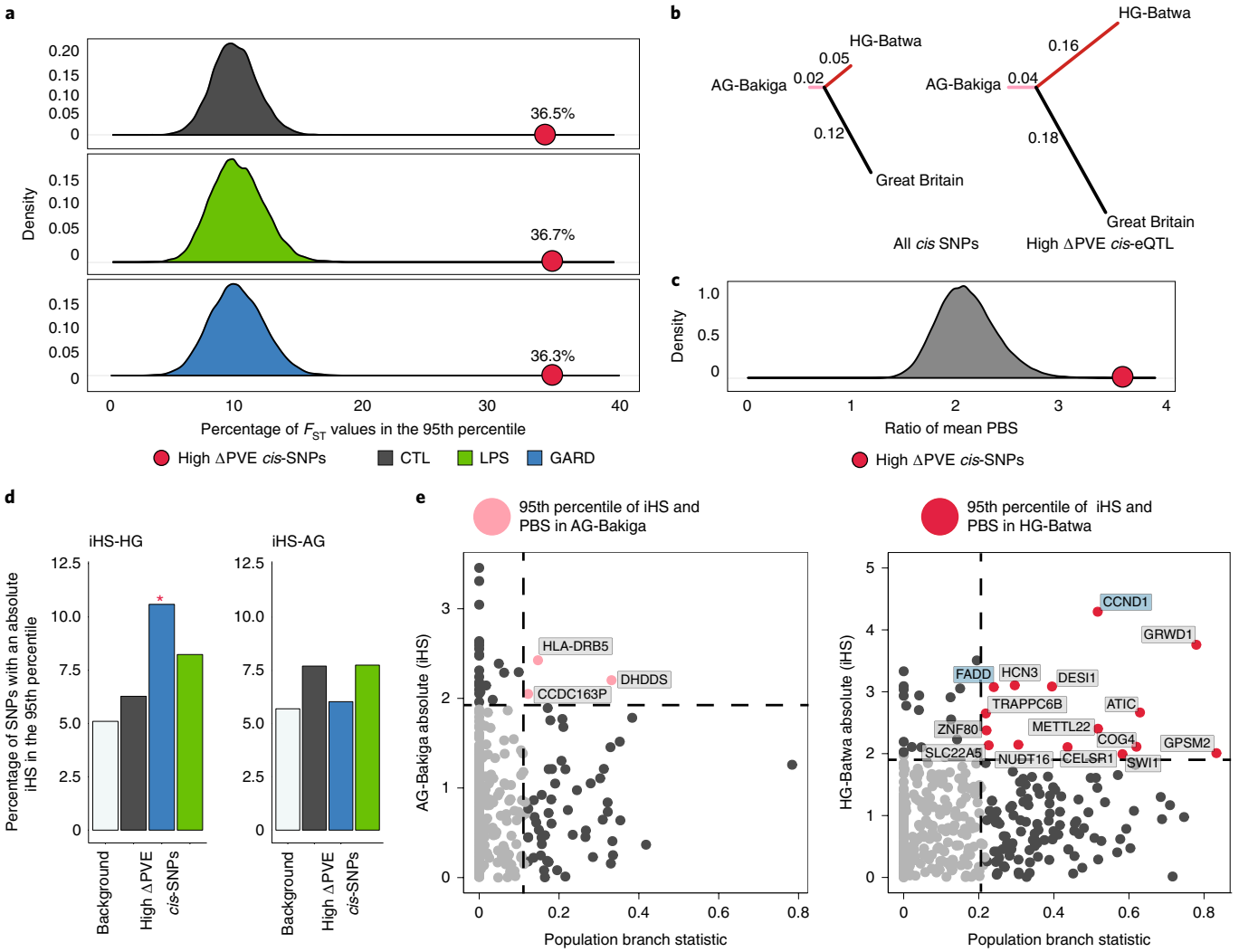

Fig. 4 | Evidence of selection driving population differences in immune response. a, This density plot shows the distribution of the percentage of SNPs with extreme values of F_{ST} (for example, in the 95th percentile) for a set of randomly sampled *cis*-SNPs equally sized sets of SNPs matched for allele frequencies with high-ΔPVE SNPs. A total 10,000 iterations were run to obtain the distribution for each condition. The red point on each graph shows the percentage of high-ΔPVE SNPs in the 95th percentile. High-ΔPVE variants in all conditions had significantly more SNPs in the 95th percentile (F_{ST} comparison chi-squared statistic; CTL P value = 2.2^{-16}, LPS P value = 2.2^{-16}, GARD P value = 2.2^{-16}). **b**, A tree diagram illustrating the mean values of the population branch statistic for the HG-Batwa, AG-Bakiga and a cohort from Great Britain as an outgroup. This figure illustrates a greater mean PBS score in the HG-Batwa population among high-ΔPVE variants. **c**, The distribution of the ratio of mean PBS in the HG-Batwa to the AG-Bakiga for a set of randomly sampled *cis*-SNPs equally sized sets of SNPs matched for allele frequencies with high-ΔPVE SNPs. A total 100,000 iterations were run to obtain the distribution and to calculate the P value. The red point shows the ratio of mean PBS values represented as the branch lengths in the tree graph. **d**, A bar graph illustrating the percentage of high-ΔPVE SNPs that have an iHS value in the 95th percentile compared to a background of all top *cis*-SNPs. For iHS, only values in the GARD-stimulated cells in the HG-Batwa population had significantly more SNPs in the 95th percentile (HG-Batwa iHS comparison chi-squared statistic; CTL P value = 0.446, LPS P value = 0.080, GARD P value = 0.002; AG-Bakiga iHS comparison chi-squared statistic; CTL P value = 0.586, LPS P value = 0.929, GARD P value = 0.210). **e**, PBS values for selection between populations graphed against absolute iHS values showing selection within each population for high-ΔPVE variants. The pink (AG-Bakiga) and red (AG-Batwa) dots represent high-ΔPVE SNPs in the 95th percentile of both PBS and iHS. Among this group, points are labelled with the corresponding gene name.

Positive selection has helped shape immune response differences. We next examined whether positive selection has contributed to the identified differences in immune response between the HG and AG populations. To do this we focused specifically on the set of 475 high-ΔPVE variants, which represent a genetic substrate on which natural selection could potentially act to drive differences in immune response between the two population groups. Given that AG populations have recently shifted their mode of subsistence (from hunting and gathering to agriculture), they are suggested to have experienced commensurate changes in pathogen burden and new selection pressures[1–4]. Under this scenario, we would expect to observe stronger evidence of positive selection on high-ΔPVE SNPs in the AG-Bakiga population relative to that observed for the HG-Batwa population. Surprisingly, our data suggest the opposite.

We found that high-ΔPVE SNPs were significantly more likely to have extreme levels of population differentiation (F_{ST} value above the 95th percentile of the genome-wide distribution) as compared to equally sized sets of SNPs matched for allele frequencies with high-ΔPVE SNPs (Fig. 4a; >3.4-fold enrichment in all conditions; $P < 10^{-4}$). This result suggests a driving role for evolutionary processes in shaping HG-Batwa and AG-Bakiga population divergence in immune regulation but does not alone distinguish the population

lineage(s) on which the selection occurred. We therefore also calculated the population branch statistic (PBS; ref. [27]), which provides an estimate of the magnitude of allele frequency change for each SNP that occurred along each population lineage following divergence from a common ancestor. Using this statistic, we found that the majority of the allele frequency divergence among high-ΔPVE SNPs occurred along the HG-Batwa lineage (mean PBS HG-Batwa = 0.16; mean PBS AG-Bakiga = 0.04; Mann–Whitney *t*-test $P = 1.2 \times 10^{-14}$) and not in the lineage leading to the AG-Bakiga population (Fig. 4b). Importantly, the relative difference in the branch length leading to the HG-Batwa lineage versus the AG-Bakiga lineage among high-ΔPVE SNPs is significantly greater than that on the basis of genome-wide expectations (4.0 versus 2.3 in average out of 100,000 sets of randomly sample sets of 475 SNPs matched for allele frequencies to high-ΔPVE SNPs, $P = 2.5 \times 10^{-4}$).

Additionally, we observed a significant enrichment of extreme integrated haplotype score (iHS) values (a neutrality test devised to detect recent positive selection events within a population)[28] among high-ΔPVE SNPs only in the HG-Batwa population. Specifically, we found that extreme iHS variants in the HG-Batwa population (>95th percentile) were significantly enriched (2.1-fold) among high-ΔPVE SNPs associated to GARD PopDE genes as compared to the set of all *cis*-SNPs (chi-squared test, $P = 1.75 \times 10^{-3}$; Fig. 4c). No such enrichments were observed in the AG-Bakiga population. Finally, more high-ΔPVE SNPs and associated genes show strong signatures of natural selection (95th percentile for both PBS and iHS) in the HG-Batwa ($n = 15$) than in the AG-Bakiga ($n = 3$) (Fig. 4d), further supporting the conclusion that positive selection in the HG-Batwa lineage has at least partly led to the extreme levels of population differentiation observed in the set of high-PVE variants.

Finally, we expanded this evolutionary analysis to include available genome-wide SNP genotype data[29] from rainforest HG-Baka and AG-Nzebi and AG-Nzime from west Central Africa. Specifically, we tested whether the set of Batwa–Bakiga high-ΔPVE variants are similarly enriched for signatures of positive selection in the HG-Baka as they are in the HG-Batwa. They are not (Supplementary Fig. 5), suggesting that Batwa-specific selection on these loci probably occurred subsequent to the estimated ~12–18 × 10³ yr divergence of eastern and western African HGs[30].

Discussion

Our study provides the first genome-wide functional genomic comparison of variation in early immune responses to infection between human HG and AG populations in Africa. Altogether, our results demonstrate that positive natural selection has contributed to present-day differences in innate immune responses between the HG-Batwa and the AG-Bakiga. Yet since functional evolutionary change occurred disproportionately on the HG-Batwa lineage, our results do not provide support for the long-standing hypothesis that selective pressures imposed by pathogens were particularly acute (at least in this region of the world) for AG populations due to the emergence of new crowd epidemic diseases.

While it is difficult to contest the premise that the advent of agriculture led to the emergence of new pathogens and to the increased pathogenicity of others, it is likely that other, perhaps yet unknown, diseases have simultaneously been consistently more prevalent in HG populations. In particular, our serological data suggest that differences in viral exposure may have been a primary contributing factor to the divergence of HG-Batwa and AG-Bakiga immune responses. This notion is consistent with recent claims that viruses have been the primary drivers of adaptive evolution in mammals[31] and one of the main selective pressures during recent human evolution[32]. Interestingly, viral burden differences have also been reported in other HG–AG population comparisons[33,34]. For example, estimated ebolavirus seroprevalence was as high as 37.5% in Aka rainforest HG groups from west Central Africa compared to 13.2% among neighbouring Monzombo and Mbati agriculturalists.[34]

We chose to work with the HG-Batwa and AG-Bakiga for two reasons. First, while these two populations live in a relatively remote area of southwest Uganda, samples collected from this region could be transferred to a cell culture laboratory within 24 h—a critical factor needed to ensure the viability of PBMCs—and processed identically, limiting possible batch effects that otherwise can affect inter-population functional comparisons. Second, while the long-term ecological histories of these two populations are distinct, they have shared similar environments and subsistence modes since 1992, when the HG-Batwa were evicted from Bwindi Impenetrable Forest. Thus, potential proximate environmental effects have been minimized to the greatest possible degree, facilitating our study of the genetic basis of functional genomic variation.

Our study is still not free of challenges. First, our relatively small sample size—an inherent constraint when studying HG populations especially—limits our power to detect eQTL. It is likely that we are underestimating the true genetic contribution to ancestry-related differences in gene expression. Moreover, our ability to detect recent events of positive selection (such as those suggested to have occurred on immune system loci following the advent of agriculture) is bounded by the limited power of the currently available neutrality tests[28], especially if selection occurred on standing genetic variation[35].

We also note that the HG-Batwa are estimated to have experienced a 7.1- to 11-fold reduction in effective population size (N_e) over the past 20 × 10³ yr, versus a mild expansion (1.2- to 2.2-fold) for the AG-Bakiga over the same period[30]. However, this difference is unlikely to account for our observation of disproportionate functional evolutionary change on the HG-Batwa lineage. First, HG and AG populations in Central Africa (including the Batwa and the Bakiga) have similar mutational loads, suggesting that their demographic differences were not sufficiently long and/or to greatly influence the efficacy of selection[30]. Moreover, even if the estimated differences in recent N_e history had markedly affected selection efficacy, the expected direction would be for reduced levels of natural selection on the HG-Batwa lineage—the opposite of our result.

Finally, we also emphasize that these population lineages diverged more than 60,000 years ago, long before the origins of agriculture in Africa[29,36,37]. Thus, a substantial proportion of the functional genetic divergence we observed probably reflects earlier (pre-agriculture) evolutionary responses to long-standing ecological differences facing each lineage. Still, our results are in direct opposition to a priori expectations of radical shifts in selection pressures on human immune systems following the agricultural transition, suggesting that the reality may be much less straightforward. Future studies of denser time-course immune responses to a larger array of pathogenic stimuli, in additional cell types, and on additional pairs of HG and AG populations will help to more precisely characterize the effects of agriculture on the evolution of human immune systems.

Methods

Sample collection. Blood samples were taken from 103 individuals: 59 HG-Batwa and 44 AG-Bakiga (Bantu-speaking) individuals (Supplementary Fig. 1). We restricted our sample collection to adults. For the HG-Batwa, we only collected samples from individuals who had lived in the forest and who were born before the 1991 formation of Bwindi Impenetrable Forest National Park, a time point known well to the HG-Batwa.

Genome-wide genotyping and imputation. From the 99 individuals that were included in the sample-set used for PopDE analyses, a subset of 96 individuals (54 Batwa and 42 Bakiga, samples labelled as EQTL_set = 1 in Supplementary Table 1) were successfully genotyped on the Illumina HumanOmni1-Quad genotyping array, as previously described[17]. Briefly, genotypes of 928,705 SNPs were called in all samples using the Illumina Genome Studio v.2010. SNPs were excluded if they had a call rate <98% across all samples or if they exhibited significant deviation from Hardy–Weinberg equilibrium ($P < 1 \times 10^{6}$) in any of the individual populations. Data were phased using shapeIT (v.2.r790) and imputation was

performed using Impute2 (v.2.3.0)[38] against multi-ethnic reference panel data that includes all populations from phase 3 of the 1000 Genomes Project. In the absence of whole-genome sequencing data from the Batwa and the Bakiga themselves, we decided to use an ancestrally include reference panel as this approach has been shown to improve imputation accuracy[39]. Post-imputation, we removed genotype calls with likelihood lower than 0.9. In addition, we excluded sex chromosomes and we removed SNP positions with an information metric lower than 0.5, with minor allele frequencies below 0.1, with greater than 5% of individuals missing genotype calls or with deviating from Hardy–Weinberg equilibrium in at least one of the studied populations ($P < 1 \times 10^{6}$). After all of these filters were applied, 5,036,671 SNPs were maintained. Further, for the *cis*-eQTL analysis, only SNPs within 100 kb of a gene body were considered (2,284,380 SNPs).

Admixture and relatedness estimations. Admixture was estimated using a non-hierarchical clustering analysis of the SNP data using the software ADMIXTURE (ref. [18]), on the basis of independent SNPs (linkage disequilibrium > 0.3) from the genotyping chip dataset for the set of 96 individuals that were successfully genotyped. For the three individuals for which genotype data was not available (T15, T30 and T62, included in Pop_DE set but absent from EQTL_set), admixture values were estimated from RNA-seq data. Importantly, the correlation between admixture estimates calculated using the microarray genotype data and genotypes obtained from the RNA-seq data is extremely high ($r = 0.978$, $P < 1 \times 10^{-16}$). Accordingly, when excluding these three samples from the PopDE analyses the effect sizes obtained for ancestry-associated differences in gene expression are virtually unchanged ($R^2 > 0.97$ across all conditions; Supplementary Fig. 6).

A pair-wise relatedness matrix among genotyped individuals was computed using PLINK (ref. [39]). As expected, we found that the mean relatedness within each population was modest in both cases but significantly larger among HG-Batwa (mean relatedness among HG-Batwa samples, 6.9%; 0.6% among AG-Bakiga). To ensure that our results were not affected by the increased number of related individuals in the HG-Batwa population, we re-ran our PopDE analyses excluding strongly related individuals (pi_hat > 0.375). This yielded 57, 58 and 62 samples in CTL, GARD and LPS condition, respectively (18, 12 and 21 samples removed in each condition, either because of high relatedness or absent genotypes, of which 17, 10 and 20 were Batwa). The results of the PopDE analyses remained largely unaffected by the removal of these related samples ($r > 0.94$ for the correlation of the estimated effect sizes when using all the samples versus those obtained when we excluded closely related individuals; Supplementary Fig. 7).

Characterization of cell type composition. PBMCs were isolated from whole blood by Ficoll-Paque centrifugation and cryopreserved. Cell-type composition of each PBMC sample was quantified using the following conjugated antibodies: CD3-FITC (clone UCHT1, BD Biosciences), CD20-PE (clone L27, BD Biosciences), CD8-APC (clone RPA-T8, BD Biosciences), CD4-V450 (clone L200, BD Biosciences), CD16-PE (clone 3G8, Biolegend), CD56-APC (clone HCD-56) and CD14-Pacific Blue (clone M5E2, Biolegend). We selected these cell types because they are by far the most common cell types found in PBMCs: collectively, almost 100% of PBMCs can be assigned to one of these types. A few rarer cell types can also be found in PBMCs but they account for so few of the total pool that they have negligible effects on overall estimates of PBMC gene expression. Antibodies were incubated for 20 min. Fluorescence was analysed on a total of 30,000 cells for each population per sample with a FACSFortesa (BD Biosciences) and the FlowJo software (Treestar). Supplementary Fig. 2 illustrates what combinations of markers were used to define each of the cellular populations we considered in this study. We note that we only quantified cellular composition of PBMCs at steady-state as in our in vitro experimental system changes in cellular composition following immune stimulation are negligible because (1) new cells cannot be recruited to the site of infection, as it would happen in vivo; and (2) none of cell types found in PBMCs proliferates in response to LPS or GARD.

Ligand stimulation. PBMCs were cultured in RPMI-1640 (Fisher) supplemented with 10% heat-inactivated FBS (FBS premium, US origin, Wisent) and 1% L-glutamine (Fisher). For each of the tested individuals, PBMCs (2 million per condition) were stimulated for 4 h at 37 °C with 5% CO_2 with the immune challenges gardiquimod (GARD, 0.5 μg ml^{-1}, TLR7 and TLR8 agonist) or lipopolysaccharide-EB (LPS, 0.25 μg ml^{-1}, TLR4 agonist). A control group of non-stimulated PBMCs were treated the same way but with only medium. We chose the 4 h time point to focus on the early transcriptional response to stimulation. This choice was on the basis of our own experience that indicates that the 4 h time point strikes a balance between the ability to detect biologically relevant gene regulatory responses to Gard/LPS, while being early enough to avoid notable cell death (which can lead to substantial alterations in gene-expression profiles that may be orthogonal to immune response itself)[40,41].

Steps for RNA-sequencing. Total RNA was extracted from the non-stimulated and stimulated cells using the miRNeasy kit (Qiagen). RNA quantity was evaluated spectrophotometrically and the quality was assessed with the Agilent 2100 Bioanalyzer (Agilent Technologies). Only samples with no evidence of RNA degradation (RNA integrity number >8) were kept for further experiments. RNA-sequencing libraries were prepared using the Illumina TruSeq protocol. Once prepared, indexed complementary DNA libraries were pooled (six libraries per pool) in equimolar amounts and sequenced with single-end 100 base pair (bp) reads on an Illumina HiSeq2500. In total, we generated RNA-sequencing profiles for 265 samples coming from 101 different individuals.

Adaptor sequences and low-quality score bases (Phred quality score <20) were first trimmed using Trim Galore (v.0.2.7). The resulting reads were then mapped to the human genome reference sequence (Ensembl GRCh37 release 75) using STAR (2.4.1d; ref. [42]) with an hg19 transcript annotation GTF downloaded from ENSEMBL (date: 2 July 2014). Reads matrices were computed using htseq-count (ref. [42]). To ensure stringent quality control of the RNA-seq data we removed from downstream analyses samples: (1) with less than 10 million of sequencing reads, (2) with less than 50% of reads mapping to annotated exons, and (3) samples that in a principal component analysis appeared to be contaminated or had failed to respond to the immune challenges. To check for potential sample mixups, we confirmed that genotype calls from the genotyping array matched those obtained from the RNA-seq data. After these filtering steps we were left with 229 samples (76 CTL, 83 LPS and 70 GARD, samples labelled as PopDE_set = 1 in Supplementary Fig. 1), coming from 99 individuals (42 HG-Bakiga and 57 AG-Batwa).

Identification of PopDE genes. To estimate the effects of HG-ancestry on gene expression (within each experimental condition), gene-expression levels across samples were normalized using the TMM algorithm (weighted trimmed mean of *M*-values), implemented in the edgeR R package[43]. Afterwards, we log-transformed the data and obtained precision-weights using the voom function in the limma package[44]. Only genes showing a median $\log_2$(c.p.m.) >2 within at least one of the experimental conditions were included in the analyses, which resulted in 10,895 protein-coding genes. We decided to focus solely on protein-coding genes to reduce the burden of multiple testing and because it is easier to derive biological interpretations from coding genes. Sequencing flow cell batch effects were removed using the function ComBat, in the sva Bioconductor package[45]. Then, expression was modelled as a function of HG-ancestry levels, while correcting for sex (x_1), proportions of CD4$^+$ T cells (x_2), CD14$^+$ monocytes (x_3), CD20$^+$ B cells (x_4) and the fraction of reads assigned to the transcriptome (x_5). Monocytes, T cells and B cells were included in the model after we identified that they were the only significant drivers of tissue composition effects on gene expression (cell types whose proportion in blood had a significant effect (FDR < 5%) in at least 2.5% of the genes tested, in at least one condition). The fraction of reads assigned to the genome (x_5) was included because this explained a significant (albeit small) fraction of the total variance in gene-expression levels (median = 1.6%, 2.5% and 6.4%, in CTL, LPS and GARD, respectively). We note, however, that when excluding the covariate x_5 from the model below, both effect sizes and *P* values for admixture effects remain almost exactly the same as when correcting for variation in the fraction of reads assigned to the genome ($R^2 > 0.979$ in all three conditions; Supplementary Fig. 8).

Using the weighted fit function from limma (lmFit) and the weights obtained from voom, we fitted the following model:

$$E^{c} = \sum_{i=1}^{5} \beta_i \times x_i + \beta_{HG} \times HG + \varepsilon \quad (1)$$

where E^c represents the vector of flow cell-corrected expression levels of a given gene in condition c, β_i the effects of the covariates and β_{HG} the effect of HG genetic ancestry. The β of these coefficients represent the fold-change (FC) effects associated with unit variation in each of the variables tested: for sex, this corresponds to the average differences in expression between male and female; for HG, the fold change difference in gene expression between HG and AG. For all other variables, since they are standardized, it represent the differences in expression associated to a shift in the covariate equal to one standard deviation.

We note that we did not include age as a covariable in our model because not all HG-Batwa individuals know their calendar ages. However, if differences in mean age between HG-Batwa and the AG-Bakiga individuals was confounding our PopDE results, then we would expect PopDE genes to be enriched among age-associated genes. To test that hypothesis, we retrieved the list age-associated genes reported by Piasecka et al. (at an FDR < 5%)[46]. That study analysed leukocyte gene expression from a panel of 1,000 healthy individuals at both steady-state and on infection with *Escherichia coli* (broadly similar to our LPS condition) and influenza (broadly similar to our Gard condition). We found no evidence that our PopDE genes were enriched among the age-associated genes reported by Piasecka et al. (odds ratio 0.75 [range 0.31–1.9]; $P = 0.51$), suggesting that age variation is unlikely to confound our results.

Estimation of PopDR statistics. To model the effects of HG-admixture on the intensity of the response to either GARD or LPS stimulation (PopDR effects), individual-wise fold-change matrixes were built for each ligand. To do so, the effects of the technical covariates (sex, tissue composition and fraction of mapped reads) were first removed from the flow cell-corrected expression matrixes within each condition. The resulting matrixes were subtracted (LPS-CTL and GARD-CTL,

in $\log_2$ scale) to build corrected fold-change matrixes using only individuals for which pairs of samples CTL versus ligand were available (70 individuals for LPS, 59 for GARD; see Supplementary Fig. 1). Finally, fold-changes were modelled according to a simple design $FC = \beta_{HG} \cdot HG + \varepsilon$, using lmFit, with weights propagated from the ones calculated by voom for each condition. More specifically, voom weights are the inverse of the variance expectation for each RNA-seq entry, obtained from the method defined by Robinson et al.[44]. This means that, for a given fold-change entry $FC = E^{ligand} - E^{CTL}$, we propagate the expected variance of the FC as follows: $\sigma^2(FC) = \sigma^2(E^{ligand}) + \sigma^2(E^{CTL})$.

Since the within condition weights were: $w_{ligand} = 1/\sigma^2(E^{ligand})$ and $w_{CTL} = 1/\sigma^2(E^{CTL})$, $\sigma^2(FC) = 1/w_{CTL} + 1/w_{ligand}$ and, finally:

$$w_{FC} = 1/\sigma^2(FC) = \frac{1}{1/w_{CTL} + 1/w_{ligand}} \quad (2)$$

Power considerations. Power calculations specifically devised for RNA-seq data[47] suggest that we are reasonably powered to detect even modest changes in gene expression between the two population groups. Assuming: (1) that the minimum average read counts among the differently expressed genes is five read counts, (2) the maximum dispersion is 0.5, (3) the total number of genes for testing is 10,895, and (4) that 10% of these genes are expected to be differently expressed between the two populations; our sample size provides 74% power to detect changes in mean gene expression between the two populations above 50% (or 0.58 on a $\log_2$ scale). While these power calculations inherently rely on a large number of assumptions (for example, effect sizes, variance estimates, and so on), our own data provide empirical evidence that we can detect statistically robust differences in gene expression between HG-Batwa and the AGR-Bakiga. Specifically, for PopDE effects, we were able to detect average $\log_2$ fold-change admixture effects as small as 0.28 (that is, a 20% change in mean gene expression between individuals with 100% HG-Batwa ancestry versus 100% AG-Bakiga ancestry). For PopDR effects, with an FDR < 10% we were able to detect mean ancestry effects on ligand response of 0.38 and 0.23 $\log_{FC}$ for LPS and GARD, respectively (Supplementary Fig. 9).

Ligand stimulation effects and DE statistics. To estimate the overall LPS and GARD effects on gene expression, we separated the samples as CTL+GARD and CTL+LPS samples and analysed them following the same analytical procedure used for PopDE, this time according to the following model design:

$$E = \sum_{i=1}^{5} \beta_i \times x_i + \beta_{HG} \times HG + \beta_{stim} \times stim + \varepsilon \quad (3)$$

where stim is a dummy variable capturing the association of each sample to either the CTL condition (stim = 0) or the stimulated condition (stim = 1), and, thus, β_{stim} captures the overall ligand effects on gene expression. Whilst the CTL and LPS samples were sequenced together as part of the same sequencing batch, the GARD samples were sequenced in a later batch. Thus, to avoid the confounding sequencing batch and the effects of GARD-stimulation, we re-sequenced a reduced number of CTL samples along with the GARD batch, of which, five CTL samples passed our QC filters.

We performed the resequencing specifically to estimate the magnitude of the batch effect for each gene (by modelling gene expression as a function of batch, for the five controls sequenced in the first batch and the five otherwise identical controls sequenced in the second batch). We then regressed out these batch effect estimates from the control and GARD samples before identifying GARD-responding genes.

Although this approach is less optimal than sequencing all three conditions together on the same flow cells, we believe that our approach does successfully show true biological effects of GARD-stimulation. For example, gene-set enrichment analysis shows that GARD-responsive genes are strongly enriched for pathways involved in antiviral responses such as defence response to virus (the top-ranked enriched gene ontology term: fold-enrichment = 13.24, FDR = 1.6×10^{-35}), type I interferon signalling (fold-enrichment = 8.5, FDR = 2.0×10^{-22}) and regulation of viral life cycle (fold-enrichment = 7.96, FDR = 3.6×10^{-17}). This observation suggests that differences in expression between GARD and CTL samples reflect a true biological response to the viral ligand. Most importantly, any potential batch effect does not affect our estimation of ancestry effects within CTL, LPS or GARD datasets, which are the effects of primary interest for this study.

False discovery rates in PopDE, PopDR and analyses of differences in gene expression in response to immune stimulation. To avoid biases related to distributional assumptions on statistical significance that might arise as a result of batch removal procedures or data pre-treatment, for all of our PopDE, PopDR and analyses of differences in gene expression in response to immune stimulation, we controlled for multiple testing using a generalization of the FDR method of Storey and Tibshirani, re-calibrated to empirical null *P* value distributions generated via permutation tests, as we previously described[25]. To perform these tests, in the case of PopDE and PopDR effects, HG-Batwa admixture was randomly permuted, while for establishing the null distribution for ligand stimulation effects, condition labels (CTL versus stimulus) were randomly re-assigned within each individual. In this case, whenever one single sample was available for a given individual, it was labelled either as CTL or stimuli (either LPS or GARD), with $P = 0.5$. Permutation tests were repeated 1,000 times per test.

Gene set enrichment analyses. GSEA were run using the javaGSEA Desktop application by the Broad Institute (http://software.broadinstitute.org/gsea/index.jsp v.3.0) against the 'Hallmark gene sets' from the molecular signatures database collection. The GSEA pre-rank mode was used, ranking genes according to *t* statistics for both popDE and PopDR effects. The *t* statistics captures both the significance level and the direction of the effects: large positive and negative values will refer to genes showing a significantly higher or low expression in HG-Batwa as compared to AG-Bakiga, respectively. The complete results of these analyses are shown in Supplementary Table 3.

Antibody profiling. Antibody profiling was performed using VirScan, as previously described[19]. Briefly, we added 2 µl sera to 1 ml VirScan bacteriophage library, diluted to ~(2×10^5)-fold representation (2×10^{10} plaque-forming units for a library of 1×10^5 clones) in phage extraction buffer (20 mM Tris-HCl, pH 8.0, 100 mM NaCl, 6 mM $MgSO_4$), in a single well of a 96-deep-well plate, pre-blocked with 3% bovine serum albumin in TBST buffer. We allowed the serum antibodies to bind the phage overnight on a rotator at 4 °C. To each well, we then added 40 µl of a 1:1 mixture of magnetic protein A:protein G Dynabeads (Invitrogen) and rotated for 4 h at 4 °C to allow sufficient binding of phage-bound antibodies to magnetic beads. Using a 96-well magnetic stand to immobilize the magnetic bead–antibody–phage complexes, we then washed the beads three times with 400 ml of PhIP-Seq wash buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 0.1% NP-40). After the final wash, beads were re-suspended in 40 ml water and phage were lysed at 95 °C for 10 min. For downstream statistical analyses, we also lysed phage from the library before immunoprecipitation (the input library) and after immunoprecipitation using only phage extract buffer without serum ('beads only control'). Each sample was run in duplicate.

Briefly, we performed two rounds of PCR amplification on the lysed phage material using hot start Q5 polymerase. The first round of PCR used the primers IS7_HsF5_2 and IS8_HsF3_2. The second round of PCR used 1 ml first-round product and the primers IS4_HsF5_2 and a unique indexing primer for each sample to be multiplexed for sequencing, where 'xxxxxxx' denotes a unique seven-nucleotide indexing sequence (see below). After the second round of PCR, DNA concentration was quantified using qPCR, and pooled equimolar amounts of all samples were used for gel extraction. The extracted pooled DNA was sequenced by the Harvard Medical School Biopolymers Facility using a 50-bp read cycle on an Illumina HiSeq 2000 or 2500, with the full pool split and run over both lanes of a HiSeq flow cell to obtain 700,000–1,300,000 reads per sample.

IS7_HsF5_2:
ACACTCTTTCCCTACACGACTCCAGTCAGGTGTGATGCTC
IS8_HsF3_2:
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCG-AGCTTATCGTCGTCATCC
IS4_HsF5_2:
AATGATACGGCGACCACCGAGATCTACACTCTT-TCCCTACACGACTCCAGT
Indexing primer:
CAAGCAGAAGACGGCATACGAGATxxxxxxxGTGACTGGAGTTCAGAC GTGT

After sequencing, samples were deconvoluted and reads aligned to the known epitope reference library for quantification and statistical analysis, as previously described. When an antibody to a particular epitope was in the sample serum, the epitope was expected to be enriched above a specific threshold, with the threshold dependent on the relative input count of the particular phage in the input library. *P* values for enrichment were calculated using generalized Poisson regression to obtain a distribution of next-generation sequencing read counts per sample for a given input count.

Analysis of viral epitope burden. The goal of this analysis was to identify viruses differentially associated to either one of the two populations tested. To that end, we first restricted our analysis to a set of 130 viruses known to be present in Africa. The full list of viruses tested can be found in Supplementary Table 4. For these viruses, we obtained an estimation of seropositivity for each individual by counting the number of epitopes for which they tested positive (defined as epitopes detected above background at a $P < 0.05$ in both technical replicates). After filtering out lowly represented viruses (those whose median number of epitopes across all individuals was lower than two), the number of viruses was reduced to 112, for which we quantified the relative deviation of epitope counts per individual, with respect to the overall mean of each virus. Explicitly, let r_i^j represent the number of positive epitopes for virus *i* and individual *j*, and $\langle r^j \rangle_i$ the virus average across all individuals. Thus, the relative deviation in seropositivity for each individual gets defined, for individual *i* and virus *j* as $\delta_i^j = \left(r_i^j - \langle r^j \rangle_i\right) / \langle r^j \rangle_i$. By testing for a linear association between δ_i^j and HG-ancestry, we estimate the inter-population differences in seropositivity relative to the mean epitope prevalence of each virus. We conducted this analysis using the lmFit function, in the R package limma[44]. Finally, false discovery rates associated to these linear models were estimated using Storey and Tibshirani's method implemented in the R package qvalue[48].

Mapping of *cis*-eQTL. *Cis*-eQTL mapping was conducted using the R package Matrix eQTL (ref. [49]). We estimated associations between SNP genotypes (variable G in equation (4) below) and changes in gene-expression levels using a linear regression model where alleles affecting expression, denoted G, were assumed to be additive. This was conducted for each of the conditions separately with individuals from both populations included in the analyses. Associations of SNPs within the gene body or 100 kb upstream and downstream of the transcript start-site and transcript end-site were used to map *cis*-eQTL. SNPs with a minor allele frequency less that 10% were removed from the analyses resulting in 2,284,380 autosomal SNPs that were tested against a total of 10,479 protein-coding genes. To account for false positives resulting from population structure, the first two principal components obtained from a principal component analysis on the genotype data were included in the model (genotype principal components (GPC)). For each library, we also took into account the potential biases and technical confounders. These included, as in the DE analyses, sex (x_1), proportions of CD4$^+$ cells (x_2), CD14$^+$ cells (x_3), CD20$^+$ cells (x_4), the fraction assigned for example the percentage of reads mapping to the transcriptome (x_5), as well as sequencing flow cell, which was accounted for by including in the model as many covariates as sequencing flow cell levels sf_i present in each case ($n_{sf}(c)$):

$$\tilde{E}^c = \sum_{i=1}^{5} \beta_i \times x_i + \sum_{i=1}^{n_{sf}(c)} \beta_{sf} \times x_{sf} + \beta_{GPC1} \times GPC_1 + \beta_{GPC2} \times GPC_2 + \beta_G \times G + \varepsilon \quad (4)$$

In this model, $\tilde{E}^c$ represents a vector of transformed expression values in condition c, which we obtained from the original expression values E^c after accounting for unmeasured-surrogate confounders. Specifically, we extracted the principal components EPC_i from a correlation matrix of the expression table within each condition E^c, and then regressed out the first $n_{EPC}(c)$ of them as follows: $E^c = \sum_{i=1}^{n_{EPC}(c)} \beta_{EPC_i} \times EPC_i + \varepsilon_{EPC}$; to obtain from the residuals of this expression the transformed expression values used in equation (4): $\tilde{E}^c = \varepsilon_{EPC}$. The specific number of principal compents to regress out for each condition was chosen empirically (23,25), on optimization of the signal strength obtained for EQTLs in equation 4. This yielded $n_{EPC}(CTL) = n_{EPC}(GARD) = 8$ and $n_{EPC}(LPS) = 11$.

We decided to do eQTL mapping on the combined dataset because our within-population sample sizes would be too small to provide sufficient mapping power. Indeed, when we re-ran the eQTL mapping on the HG-Batwa and the AG-Bakiga separately, the number of *cis*-eQTL identified within each condition dropped greatly (from >2,000 eQTL-associated genes per condition to only 281–540 at the same FDR cutoff on the population-specific analyses; Supplementary Fig. 10). Importantly, the larger number of eQTL observed in the combined dataset is not a reflection of unaccounted population structure. Indeed, the first two principal components of the genetic data included in our model clearly separate the HG-Batwa from the AG-Bakiga, and PC-1 alone correlates almost perfectly with genetic ancestry (Supplementary Fig. 11; $P < 1 \times 10^{-16}$). Most importantly, the effect sizes of the eQTL obtained using the combined dataset are very strongly correlated with those obtained when performing the mapping on the individual populations ($R > 0.93$ in all conditions tested; Supplementary Fig. 11), which empirically demonstrates that our eQTL are not an artefact due to population structure.

Proportion of variance estimations. To compute the PVE by the different covariates in the PopDE models (Supplementary Fig. 3), we used the method proposed by Shabalin et al.[50] and implemented in the R package relaimpo[51]. According to this approach, the contribution of each covariate to the overall determination coefficient R^2 is calculated on adding sequentially all covariates to the model and calculating their contribution to the increase of R^2 in each case, averaging across all possible covariate orderings. We summed the contributions of the three fractions of cell types included in the models (CD14$^+$, CD4$^+$ and CD20$^+$) to obtain the estimates of tissue composition reported in the Supplementary Fig. 3. The PVE associated with sex (PVE_{sex}), tissue composition ($PVE_{tissue} = PVE_{CD4} + PVE_{CD14} + PVE_{CD20}$) and HG ancestry ($PVE_{HG}$), add up to the total fraction of explained variance for each gene, that is:

$$R^2 = PVE_{sex} + PVE_{tissue} + PVE_{HG} \quad (5)$$

To quantify what fraction of the inter-population differences in gene expression was accounted for by *cis*-eQTL, we first estimated, for each gene, the contribution of HG-ancestry on gene-expression variation within each condition (that is, the PopDE effect sizes $\beta_{HG}^{CTL}, \beta_{HG}^{LPS}, \beta_{HG}^{GARD}$, for genes showing statistical evidence of ancestry effects at a relaxed threshold of FDR < 0.2). The proportion of variance explained by HG ancestry PVE_{HG}^{o} is defined as the increase in variance explained (that is, the increase in R^2) by the PopDE model in equation (1), on adding the HG variable as the last covariable. Then, we fitted an alternative PopDE model for each gene, starting from equation (1) but adding the genotype of the top *cis*-SNP for the gene being tested, G_{Top}, as follows:

$$E^c = \sum_{i=1}^{5} \beta_i \times x_i + \beta_{HG} \times HG + \beta_{GTop}^{c} \times G_{Top} + \varepsilon \quad (6)$$

From this model, an analogous estimate $PVE_{HG}^{G_{Top}}$ was obtained, which captured the relevance, in terms of explained variance, of adding HG ancestry, once the best SNP was already included in the model.

Once the contribution to final variance explained was obtained from both models we retrieved the difference between the two models $\Delta PVE = \left(PVE_{HG}^{o} - PVE_{HG}^{G_{Top}}\right)/PVE_{HG}^{o}$. ΔPVE represents the proportion of the population difference in gene expression that can be attributed to the strongest *cis*-eQTL for the gene of interest.

To assess the statistical significance of ΔPVE, we used the same approach described above but we removed the effect of the strongest *cis*-eQTL identified after randomly shuffling individual labels from the genotype data. Then, to construct a null model that was unbiased by the selection of the best SNP per gene, we built a third linear model, analogous to that of equation (6). Instead of the true, most significant SNP variant for that gene G_{Top}, we used the most significant variant that arises by chance, among all the permuted SNPs, G_{Top}^{Random}:

$$E^c = \sum_{i=1}^{5} \beta_i \times x_i + \beta_{HG} \times HG + \beta_{GTop.Rand}^{c} \times G_{Top}^{Random} + \varepsilon \quad (7)$$

Then, we calculate PVE values on the basis of the HG-admixture effects inferred from equation (7), which we call $PVE_{HG}^{G_{Top.Rand}}$. Finally, we estimate the null-expectation for ΔPVE, which we call ΔPVE_{null}, as follows:

$$\Delta PVE_{null} = \left(PVE_{HG}^{o} - PVE_{HG}^{G_{Top.Rand}}\right)/PVE_{HG}^{o} \quad (8)$$

Comparing the distribution of observed ΔPVE to the distribution of its empirical null-expectation ΔPVE_{null} we obtain empirical one-tailed P values for each test, defined as the fraction of null-tests with $\Delta PVE_{null} > \Delta PVE$. Finally, proper correction for multiple testing (Storey–Tibshirani FDRs) of these empirical P values allows us to stablish an empirical model for statistical significance of these effects (see Supplementary Fig. 4).

Selection statistics. We calculated the selection statistics by using the individuals used to map *cis*-eQTL that had an admixture less than 0.2 or greater than 0.8 to clearly define the two populations. This included 43 AG-Bakiga individuals and 39 HG-Batwa individuals. We calculated the fixation indexes (F_{ST}) using a modified version of Wright's F_{ST} for all SNPs using VCFtools v.0.1.12b (ref. [52]). The integrated haplotype scores (iHS) were calculated using Selscan, which is a program that calculates haplotype-based scans for recent or ongoing signatures of positive selection. This method is on the basis of the knowledge that when adaptive de novo mutations quickly increase in frequency it reduces genetic diversity around this variant faster than recombination can occur. Therefore, this score is a measure of haplotype homozygosity extending from an adaptive locus[53]. To do this, phased genotypes were created using SHAPEITv2 (ref. [54]) for each chromosome independently. We calculated iHS separately for the HG and AG population for all imputed genotypes. When estimating mean F_{ST} and iHS among *cis*-eQTL we combined *cis*-eQTL mapped in all conditions and selected the variant with the lowest P value for a given gene resulting in one *cis*-SNP per gene. The F_{ST} and/or iHS for that SNP was then considered in this analysis. Finally, the PBS was calculated from F_{ST} values using a cohort from Great Britain available from the 1000 Genomes Project as an outgroup. F_{ST} was first used to calculate population divergence as $T = -\log(1 - F_{ST})$ and then PBS was calculated for each SNP for HG-Batwa and AG-Bakiga as:

$$PBS.Batwa = (T.Batwa.Bakiga + T.Batwa.GBR - T.Bakiga.GBR)/2$$

$$PBS.Bakiga = (T.Batwa.Bakiga + T.Bakiga.GBR - T.Batwa.GBR)/2$$

Ethics. The HG-Batwa and AG-Bakiga samples were collected under informed consent (Institutional Review Board protocols 2009-137 from Makerere University, Uganda and 16986A from the University of Chicago). The project was also approved by the Uganda National Council for Science and Technology (HS617).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are available at https://zenodo.org/record/2656662#.XMyCSi3MzOQ.

Code availability

All scripts required to run the analyses described in the manuscript can be found at https://github.com/GFHarrison/Natural-Selection-HG-and-AG-2019.

Received: 4 December 2018; Accepted: 17 June 2019;
Published online: 29 July 2019

References

1. Diamond, J. & Bellwood, P. Farmers and their languages: the first expansions. *Science* **300**, 597–603 (2003).

2. Greger, M. The human/animal interface: emergence and resurgence of zoonotic infectious diseases. *Crit. Rev. Microbiol.* **33**, 243–299 (2007).
3. Pearce-Duvet, J. M. The origin of human pathogens: evaluating the role of agriculture and domestic animals in the evolution of human disease. *Biol. Rev. Camb. Phil. Soc.* **81**, 369–382 (2006).
4. Wolfe, N. D., Dunavan, C. P. & Diamond, J. Origins of major human infectious diseases. *Nature* **447**, 279–283 (2007).
5. Gignoux, C. R., Henn, B. M. & Mountain, J. L. Rapid, global demographic expansions after the origins of agriculture. *Proc. Natl Acad. Sci. USA* **108**, 6044–6049 (2011).
6. Page, A. E. et al. Reproductive trade-offs in extant hunter-gatherers suggest adaptive mechanism for the Neolithic expansion. *Proc. Natl Acad. Sci. USA* **113**, 4694–4699 (2016).
7. Black, F. L. Measles endemicity in insular populations: critical community size and its evolutionary implication. *J. Theor. Biol.* **11**, 207–211 (1966).
8. Anderson, R. M. & May, R. M. *Infectious Diseases of Humans: Dynamics and Control* (Oxford Univ. Press, 1992).
9. Furuse, Y., Suzuki, A. & Oshitani, H. Origin of measles virus: divergence from rinderpest virus between the 11th and 12th centuries. *Virol. J.* **7**, 52 (2010).
10. Matthijnssens, J. et al. Full genome-based classification of rotaviruses reveals a common origin between human Wa-Like and porcine rotavirus strains and human DS-1-like and bovine rotavirus strains. *J. Virol.* **82**, 3204–3219 (2008).
11. Suzuki, Y. & Nei, M. Origin and evolution of influenza virus hemagglutinin genes. *Mol. Biol. Evol.* **19**, 501–509 (2002).
12. Sundararaman, S. A. et al. Genomes of cryptic chimpanzee *Plasmodium* species reveal key evolutionary events leading to human malaria. *Nat. Commun.* **7**, 11078 (2016).
13. Otto, T. D. et al. Genomes of all known members of a *Plasmodium* subgenus reveal paths to virulent human malaria. *Nat. Microbiol.* **3**, 687-697 (2018).
14. Dounias, E. & Froment, A. When forest-based hunter-gatherers become sedentary: consequences for diet and health. *UNASYLVA-FAO* **57**, 26-33 (2006).
15. Barreiro, L. B. & Quintana-Murci, L. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nat. Rev. Genet.* **11**, 17-30 (2010).
16. Karlsson, E. K., Kwiatkowski, D. P. & Sabeti, P. C. Natural selection and infectious disease in human populations. *Nat. Rev. Genet.* **15**, 379-393 (2014).
17. Perry, G. H. et al. Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. *Proc. Natl Acad. Sci. USA* **111**, E3596–E3603 (2014).
18. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
19. Xu, G. J. et al. Comprehensive serological profiling of human populations using a synthetic human virome. *Science* **348**, aaa0698 (2015).
20. McGeoch, D. & Davison, A. J. in *Origin and Evolution of Viruses* (eds Domingo, E., Webster, R. & Holland, J.) Ch. 17 (Academic Press, 1999).
21. McGeoch, D. J., Dolan, A. & Ralph, A. C. Toward a comprehensive phylogeny for mammalian and avian herpesviruses. *J. Virol.* **74**, 10401–10406 (2000).
22. Van Blerkom, L. M. Role of viruses in human evolution. *Am. J. Phys. Anthropol.* **122**, 14–46 (2003).
23. Barreiro, L. B. et al. Deciphering the genetic architecture of variation in the immune response to *Mycobacterium tuberculosis* infection. *Proc. Natl Acad. Sci. USA* **109**, 1204–1209 (2012).
24. Fairfax, B. P. et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).
25. Nédélec, Y. et al. Genetic ancestry and natural selection drive population differences in immune responses to pathogens. *Cell* **167**, 657–669 (2016).
26. Quach, H. et al. Genetic adaptation and neandertal admixture shaped the immune system of human populations. *Cell* **167**, 643–656 (2016).
27. Yi, X. et al. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010).
28. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
29. Patin, E. et al. The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. *Nat. Commun.* **5**, 3163 (2014).
30. Lopez, M. et al. The demographic history and mutational load of African hunter-gatherers and farmers. *Nat. Ecol. Evol.* **2**, 721-730 (2018).
31. Enard, D., Cai, L., Gwennap, C. & Petrov, D. A. Viruses are a dominant driver of protein adaptation in mammals. *eLife* **5**, e12469 (2016).
32. Enard, D. & Petrov, D. A. RNA viruses drove adaptive introgressions between Neanderthals and modern humans. Preprint at *bioRxiv* https://doi.org/10.1101/120477 (2017).
33. Gonzalez, J. P., Nakoune, E., Slenczka, W., Vidal, P. & Morvan, J. M. Ebola and Marburg virus antibody prevalence in selected populations of the Central African Republic. *Microbes Infect.* **2**, 39–44 (2000).
34. Johnson, E., Gonzalez, J.-P. & Georges, A. Filovirus activity among selected ethnic groups inhabiting the tropical forest of equatorial Africa. *Trans. R. Soc. Trop. Med. Hyg.* **87**, 536–538 (1993).
35. Prezeworski, M., Coop, G. & Wall, J. D. The signature of positive selection on standing genetic variation. *Evolution* **59**, 2312–2323 (2005).
36. Mellars, P. Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. *Proc. Natl Acad. Sci. USA* **103**, 9381–9386 (2006).
37. Verdu, P. et al. Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Curr. Biol.* **19**, 312–318 (2009).
38. Storey, J. D. & Tibshirani, R. in *Functional Genomics* (eds Brownstein, M. J. and Kohdursky, A. B.) 149–157 (Springer, 2003).
39. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
40. Snyder-Mackler, N. et al. Social status alters immune regulation and response to infection in macaques. *Science* **354**, 1041–1045 (2016).
41. Sams, A. J. et al. Adaptively introgressed Neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans. *Genome Biol.* **17**, 246–261 (2016).
42. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
43. Anders, S. et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc.* **8**, 1765–1786 (2013).
44. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
45. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).
46. Piasecka, B. et al. Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges. *Proc. Natl Acad. Sci. USA* **115**, E488–E497 (2018).
47. Guo, Y., Zhao, S., Li, C.-I., Sheng, Q. & Shyr, Y. RNAseqPS: a web tool for estimating sample size and power for RNAseq experiment. *Cancer Inform.* **13**, 1–5 (2014).
48. Bindea, G. et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091–1093 (2009).
49. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
50. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
51. Lindeman, R. H., Merenda, P. F. & Gold, R. Z. *Introduction to Bivariate and Multivariate Analysis* (Scott, Foresman and Co, 1980).
52. Grömping, U. Relative importance for linear regression in R: the package relaimpo. *J. Stat. Softw.* **17**, 1–27 (2006).
53. Jeffrey, C. Genome-wide association study and meta-analysis finds over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* **41**, 703–707 (2009).
54. Szpiech, Z. A. & Hernandez, R. D. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol. Biol. Evol.* **31**, 2824–2827 (2014).

Acknowledgements

The authors thank the Batwa and Bakiga communities and all individuals who participated in this study; also the Batwa Development Program, J. Byaruhanga, M. Magambo, P. Byamugisha, S. Twesigomwe, J. Safari and L. Busingye for expert assistance during the sample collection process in Uganda. We thank S. Nanyunja for technical laboratory assistance. We thank J. Tung and L.B.B. laboratory members for critical reading of the manuscript. We thank Calcul Québec and Compute Canada for providing access to the supercomputer Briaree from the University of Montreal. This work was supported by NIH R01-GM115656 to G.H.P and L.B.B., a fellowship from the Réseau de Médecine Génétique Appliquée and the Fonds de Recherche du Québec–Santé to G.F.H, and 1 F32 GM125228-638 01A1 to C.M.B. RNA-seq data have been deposited in Gene Expression Omnibus (accession number GSE120502). The 1M SNP genotype data are available at the European Genome–Phenome archive, www.ebi.ac.uk/ega/ (accession numbers EGAS00001000605 and EGAS00001000908).

Author contributions

L.B.B. and G.H.P conceived and coordinated the study, and performed field work in Uganda. S.L.N facilitated samples collection. J.B., A.D. and V.Y. performed cell culture experiments. G.F.H. and J.S. conducted most data analysis, with support from F.C.G. and C.M.B and input from co-authors. M.J.M, Y.L. and S.J.E. generated VirScan data. E.S. and L.Q.M. contributed to data generation. G.F.H, L.B.B. and G.H.P. wrote the paper with input from all co-authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-019-0947-6.

Correspondence and requests for materials should be addressed to L.B.B.

nature research

Corresponding author(s): Dr. Luis Barreiro

Last updated by author(s): Dec 7, 2018

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Trim Galore (version 0.2.7); STAR (2.4.1d); edgeR R package; limma package; sva Bioconductor package; cytoscape app ClueGO (vesion 2.3.3); ShapeIT v2; ADMIXTURE; Plink; R package Matrix eQTL; R package relaimpo; VCFtools v0.1.12b; Selscan
Data analysis	Data analyses were performed using the software packages described above. For identifying genes different expressed between population we used linear models. A detailed description of each of the models used is provided in the material and methods.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

RNA-Sequencing data is deposited in the Gene Expression Omnibus with the accession number GSE120502. The 1M SNP genotype data are available at the European Genome-Phenome archive: www.ebi.ac.uk/ega/ (accession numbers EGAS00001000605 and EGAS00001000908.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	In our study we compared the immune response of two populations that have historically resided in different ecologies and maintained different sustenance strategies. The goal was to determine how variation in immune response has been shaped by natural selection. The populations included in this study were the was the Batwa hunter-gatherers (N = 56) and the Bakiga agriculturalists (N = 47). We separated peripheral blood mononuclear cells (PBMC) and serum from whole blood. We challenged PBMCs with ligands to mimic infection with either a bacteria (lippopolysacharide) or a virus (Gardiquimod). We then collected genotypes, RNA-sequencing profiles, and sequenced for anti-viral antibodies.
Research sample	Our research samples included whole blood from 103 individuals belonging to two human populations residing in Uganda: the Batwa hunter-gatherer population and the Bakiga agricultural population. Samples were all adults and included both men and women. We chose these populations because they have historically practiced different sustenance strategies and occupied different ecologies but now live in close proximity to one another.
Sampling strategy	Decisions on sample size were based on our past experience with mapping eQTL in the context of immune responses to infection.
Data collection	Blood samples were taken from a total of 103 individuals, 59 HG-Batwa (Hunter-gatherer) and 44 AG-Bakiga (Bantu speaking agriculturalist) individuals. We restricted our sample collection to adult individuals. For the HG-Batwa, we only collected samples from individuals who had lived in the forest and that were born prior to the 1991 formation of Bwindi Impenetrable Forest National Park, a time point known well to the HG-Batwa. PBMCs were collected and processed for both populations simultaneously during the same field expedition to minimize technical variability.
Timing and spatial scale	PBMCs were collected by Dr. Barreiro and Dr. Perry within a two weeks time window back in 2011. Cryopreserved PBMCs were processed in the lab and stimulated over a period of ~2 months.
Data exclusions	Samples were excluded if the data did not meet quality control thresholds.
Reproducibility	This is a population-level analyses and therefore we do not have multiple samples from the same individual. Reproducibility comes from the large number of samples obtained in each of the populations studied, and the statistical model that provide statistical support to the differences in gene expression and immune response detected.
Randomization	To avoid batch effects among different experimental conditions, cDNA libraries were pooled with six libraries per pool in equimolar amounts.
Blinding	N/A

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions	Samples were collected from multiple settlements around the Bwindi Impenetrable Forest National Park in Uganda.
Location	Bwindi Impenetrable Forest National Park in Uganda
Access and import/export	All biological samples were sent to the US under an export permit obtained from the Ugandan government.
Disturbance	NA

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Covariates included in our models included sex, genetic ancestry and variation in cellular composition of PBMCs, as measured by flow cytometry.
Recruitment	We restricted our sample collection to adult individuals. For the HG-Batwa, we only collected samples from individuals who had lived in the forest and that were born prior to the 1991 formation of Bwindi Impenetrable Forest National Park, a time point known well to the HG-Batwa.
Ethics oversight	Samples were collected under informed consent (Institutional Review Board protocols 2009-137 from Makerere University, Uganda and 16986A from the University of Chicago). The project was also approved by the Uganda National Counsel for Science and Technology (HS617). This study was approved by the ethics committee of CHU Sainte Justin (project # 2016-1215).

Note that full information on the approval of the study protocol must also be provided in the manuscript.