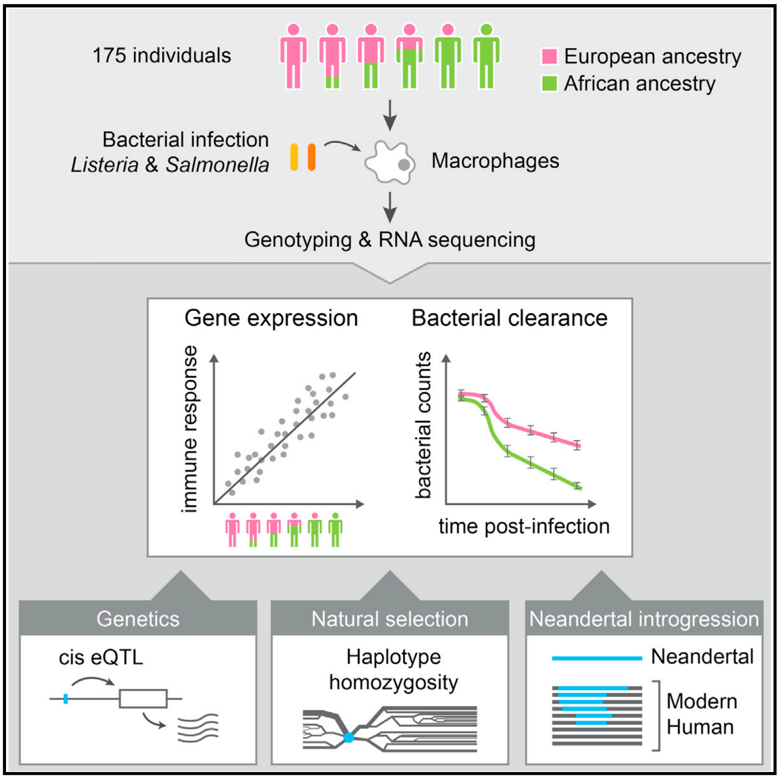
Cell

Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens

Graphical Abstract



Highlights

- Thousands of genes show population differences in transcriptional response to infection
- African ancestry is associated with a stronger inflammatory response
- Population differences in immune response are often genetically controlled
- Natural selection contributed to ancestry-associated differences in gene regulation

Authors

Yohann Nédélec, Joaquín Sanz, Golshid Baharian, ..., Jenny Tung, Vania Yotova, Luis B. Barreiro

Correspondence

luis.barreiro@umontreal.ca

In Brief

Differences in the transcriptional response to infection in human populations are under strong genetic influence, dictated by their ancestry and by recent natural selection events.

Data Resources

GSE81046

Nédélec et al., 2016, Cell *167*, 657–669
October 20, 2016 © 2016 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2016.09.025

Cell

Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens

Yohann Nédélec,[1,2,11] Joaquín Sanz,[1,2,11] Golshid Baharian,[1,2,11] Zachary A. Szpiech,[3] Alain Pacis,[1,2] Anne Dumaine,[2] Jean-Christophe Grenier,[2] Andrew Freiman,[4] Aaron J. Sams,[5] Steven Hebert,[2] Ariane Pagé Sabourin,[2] Francesca Luca,[4,6] Ran Blekhman,[7] Ryan D. Hernandez,[3,8] Roger Pique-Regi,[4,6] Jenny Tung,[9] Vania Yotova,[2] and Luis B. Barreiro[2,10,12,*]

[1]Department of Biochemistry, Faculty of Medicine, Université de Montréal, Montreal, QC H3T1J4, Canada
[2]Department of Genetics, CHU Sainte-Justine Research Center, Montreal, QC H3T1C5, Canada
[3]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA 94143, USA
[4]Department of Obstetrics and Gynecology, Wayne State University, Detroit, MI 48202, USA
[5]Department of Biological Statistics & Computational Biology, Cornell University, Ithaca, NY 14850, USA
[6]Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI 48202, USA
[7]Department of Genetics, Cell Biology, and Development and Department of Ecology, University of Minnesota, Twin Cities, MN 55108, USA
[8]Institute for Human Genetics and Quantitative Biosciences Institute, University of California, San Francisco, San Francisco, CA 94143, USA
[9]Departments of Evolutionary Anthropology and Biology and Duke Population Research Institute, Duke University, Durham, NC 27708, USA
[10]Department of Pediatrics, Faculty of Medicine, Université de Montréal, Montreal, QC H3T1J4, Canada
[11]Co-first author
[12]Lead Contact
*Correspondence: luis.barreiro@umontreal.ca

http://dx.doi.org/10.1016/j.cell.2016.09.025

SUMMARY

Individuals from different populations vary considerably in their susceptibility to immune-related diseases. To understand how genetic variation and natural selection contribute to these differences, we tested for the effects of African versus European ancestry on the transcriptional response of primary macrophages to live bacterial pathogens. A total of 9.3% of macrophage-expressed genes show ancestry-associated differences in the gene regulatory response to infection, and African ancestry specifically predicts a stronger inflammatory response and reduced intracellular bacterial growth. A large proportion of these differences are under genetic control: for 804 genes, more than 75% of ancestry effects on the immune response can be explained by a single *cis*- or *trans*-acting expression quantitative trait locus (eQTL). Finally, we show that genetic effects on the immune response are strongly enriched for recent, population-specific signatures of adaptation. Together, our results demonstrate how historical selective events continue to shape human phenotypic diversity today, including for traits that are key to controlling infection.

INTRODUCTION

As our primary interface with the environment, the immune system is thought to have evolved under strong selective pressure from pathogens (Barreiro and Quintana-Murci, 2010; Fumagalli et al., 2011; Karlsson et al., 2014). When human populations migrated out of Africa, they encountered markedly different pathogenic environments, likely resulting in population-specific selection on the immune response (Barreiro and Quintana-Murci, 2010; Fumagalli et al., 2011; Karlsson et al., 2014). Substantial evidence supports this hypothesis at the genetic level. However, we still know little about the extent to which neutral or adaptive inter-population genetic differences affect the actual immune response to pathogens.

Addressing this gap is not only important for understanding recent human evolution, but may also help reveal the molecular basis of ancestry-related differences in disease susceptibility. Individuals from different populations vary considerably in their susceptibility to many infectious diseases, chronic inflammatory disorders, and autoimmune disorders. For tuberculosis, systemic lupus erythematosus, systemic sclerosis, psoriasis, and septicemia, African American (AA) and European American (EA) individuals exhibit an up to 3-fold difference in prevalence (reviewed in Brinkworth and Barreiro, 2014; Pennington et al., 2009; Richardus and Kunst, 2001). These observations argue in favor of significant ancestry-related differences in immune response, especially in susceptibility to inflammation (Pennington et al., 2009; Richardus and Kunst, 2001).

Such differences almost certainly involve major contributions from the environment. However, genome-wide association studies (GWAS) also support a key role for genetic factors, as many of the GWAS-variants associated with infectious, autoimmune, and inflammatory diseases present extreme differences in allele frequency ($F_{st} > 0.4$) between human populations, again supporting a possible history of population-specific selection (Brinkworth and Barreiro, 2014).

GWAS results also indicate that susceptibility to many common immune-related diseases is primarily controlled by noncoding variants (Gusev et al., 2014; Hindorff et al., 2009; Schaub et al., 2012). Thus, many ancestry-related differences in disease susceptibility may result from genetically controlled transcriptional differences in immune responses to inflammatory signals. This idea is consistent with recent expression quantitative trait locus (eQTL) mapping studies in innate immune cells exposed to immune antigens or live infectious agents (Barreiro et al., 2012; Çalışkan et al., 2015; Fairfax et al., 2014; Lee et al., 2014). Such immune "response eQTL" studies have identified hundreds of genetic variants that both explain variation in the host immune response and are significantly enriched among GWAS-associated loci. However, because studies to date have mostly focused on individuals of European ancestry, the degree to which such variants contribute to population differences in the immune response remains unclear.

Here, we report an RNA-sequencing (RNA-seq)-based immune response eQTL study to test for the effects of African versus European ancestry on the transcriptional response to several live bacterial pathogens. We integrate statistical and evolutionary genetic analyses with primary macrophage gene expression levels, before and after infection, to characterize ancestry-related differences in the immune response. Our analyses address three fundamental questions about recent evolution in the human immune system: (1) the degree to which innate immune responses are differentiated by European versus African ancestry, (2) the genetic variants that account for such differences, and (3) the evolutionary mechanisms (neutral genetic drift versus positive selection) that led to their establishment in modern human populations. Finally, to facilitate the use of our data by the research community, we have developed an accessible, publicly available browser for exploring our results: the ImmunPop QTL browser (http://www.immunpop.com).

RESULTS

Transcriptional Response of Macrophages to *Listeria* and *Salmonella*

We infected monocyte-derived macrophages—phagocytic cells that are essential for fighting foreign invaders, tissue development, and homeostasis (Okabe and Medzhitov, 2016)—derived from 80 AA and 95 EA individuals (Table S1) with either *Listeria monocytogenes* (a Gram-positive bacterium) or *Salmonella typhimurium* (a Gram-negative bacterium). Following 2 hr of infection, we collected RNA-seq data from matched non-infected and infected samples, for a total of 525 RNA-seq profiles across individual-treatment combinations (mean = 36 million reads per sample; see the STAR Methods; Figure S1A). Each individual was genotyped for over 4.6 million single nucleotide polymorphisms (SNPs), with additional imputation to ~13 million SNP genotypes (see the STAR Methods). After quality control (Figure S1A), we were able to study 171 individuals with high-quality RNA-seq data, among which 168 were also successfully genotyped.

The first principal component of the resulting gene expression data accounted for 85% of the variance in our dataset and separated non-infected macrophages from macrophages infected with either *Listeria* or *Salmonella* (Figure 1A). We found extensive differences in gene expression levels between infected and non-infected cells, with 5,201 (44%) and 6,701 (56%) differentially expressed genes after infection with *Listeria* and *Salmonella*, respectively (see the STAR Methods, false discovery rate [FDR] < 0.01 and $|\log_2(\text{fold change})| > 0.5$; Table S2A). As expected, the sets of genes that responded to either infection were strongly enriched (FDR < 0.01) for gene sets involved in immune function, including the regulation of inflammatory responses, cytokine production, T cell activation, and apoptosis (Table S3).

Ancestry-Related Differences in the Innate Immune Response to Infection

We first aimed to characterize European versus African ancestry-related transcriptional differences in non-infected and infected macrophages. Because self-identified ethnicity is an imprecise proxy for the actual genetic ancestry of an individual, we used the genotype data to estimate genome-wide levels of European and African ancestry in each sample using the program ADMIXTURE (Alexander et al., 2009). Consistent with previous reports (Bryc et al., 2010; Tishkoff et al., 2009), we found that many self-identified AA individuals have a high proportion of European ancestry (mean = 30%, range 0.9%–100%; Figure S1B). In contrast, self-identified EA showed more limited levels of African admixture (mean = 0.4%, range 0%–18%; Figure S1B). Thus, we used these continuous estimates (as opposed to a binary classification of individuals into African or European ancestry) to identify ancestry-associated differentially expressed genes (i.e., pop-DE genes: genes for which gene expression levels are linearly correlated with ancestry levels; see the STAR Methods for details on the nested linear model used for this analysis).

Of the 11,914 genes we tested, we identified 3,563 pop-DE genes (30%) in at least one of the experimental conditions, explaining a mean 8.2% of expression variance (range 1.8%–44%) (FDR < 0.05: 1,745 in non-infected [NI], 1,336 in *Listeria*-infected [L], and 2,417 in *Salmonella*-infected [S] macrophages) (Figures 1B and 1C; Table S2B). These differences primarily influence mean gene expression levels across transcript isoforms, as opposed to the proportion of isoform usage within genes. Specifically, among genes with at least two annotated isoforms (n = 10,223), only 62, 39, and 48 genes exhibited evidence for ancestry-associated differential isoform usage, in the non-infected, *Listeria*-infected, and *Salmonella*-infected conditions, respectively (multivariate generalization of the Welch's t test; FDR < 0.05) (Figures 1D and S2A; Table S2D). These results were unaltered by using an alternative identification approach (Wilcoxon rank sum test, as in Lappalainen et al., 2013; see the STAR Methods for details) or when relaxing the FDR threshold used to define significance (Figure S2B). Despite the low number of genes showing ancestry-associated differences in isoform usage, many of these genes are key regulators of innate immunity, including *OAS1* that encodes isoforms with varying enzymatic activity against viral infections (Bonnevie-Nielsen et al., 2005).

Next we sought to identify genes for which the response to infection (i.e., fold change in gene expression in infected versus

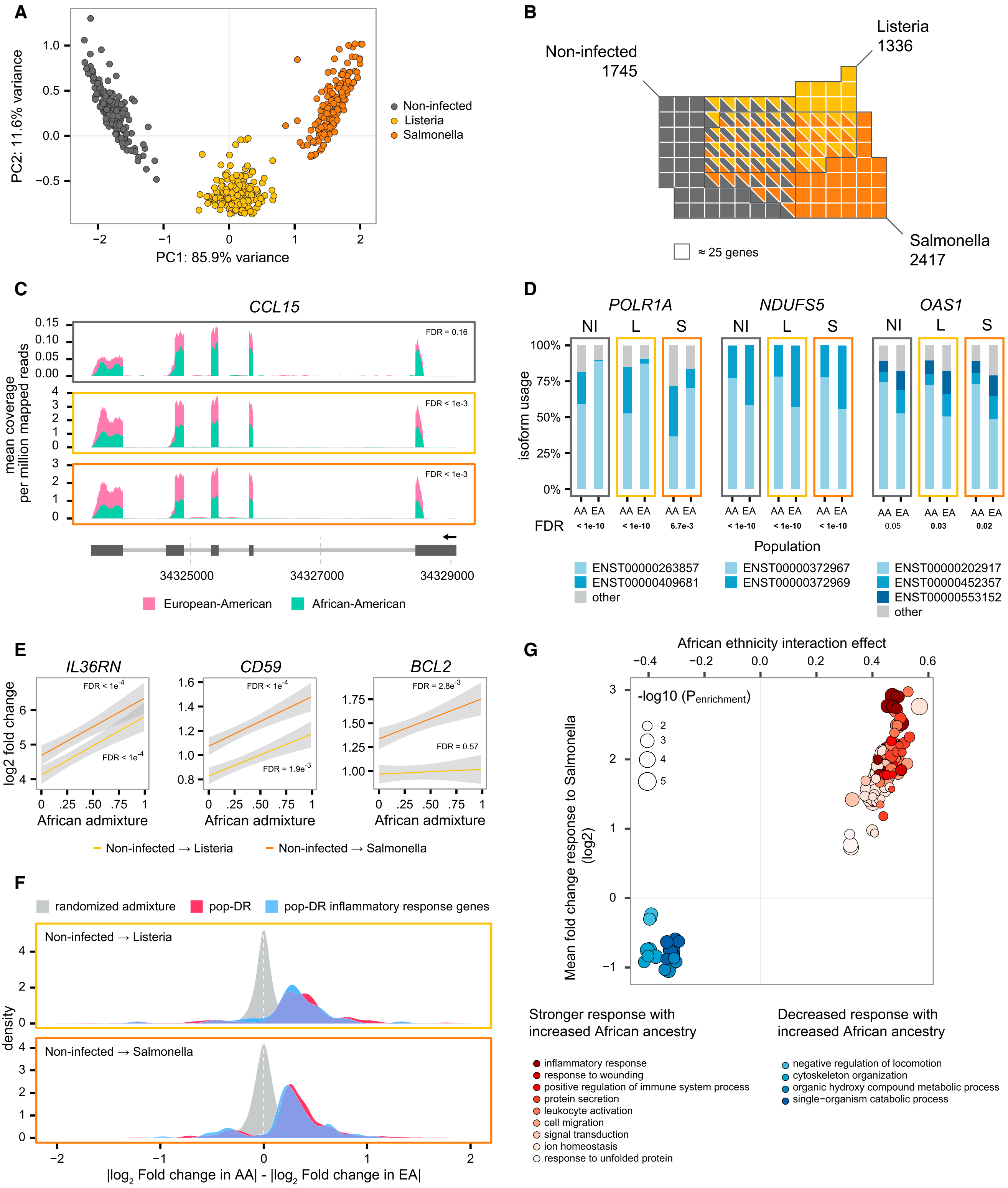


Figure 1. European and African Ancestry-Associated Differences in Immune Response

(A) Principal component analysis of gene expression data from all samples. PC1 (x axis) and PC2 (y axis) clearly separate non-infected macrophages from macrophages infected with either *Listeria* or *Salmonella*.

(legend continued on next page)

non-infected macrophages, cultured in parallel) significantly correlates with ancestry (see the STAR Methods). We term these genes "population differentially responsive" (pop-DR) genes. We detected 1,005 and 206 pop-DR genes (FDR < 0.05) in response to *Salmonella* and *Listeria*, respectively (Figure 1E; Table S2C) (the increased power for *Salmonella* likely results from the stronger transcriptional response induced by *Salmonella* relative to *Listeria*, see Figure 1A). These genes explain a mean 7.4% (range 2.6%–24%) of variance in transcriptional response to infection. Overall, we found that macrophages from individuals of African ancestry produced a markedly stronger transcriptional response to both bacterial infections (Mann-Whitney test, $p < 1 \times 10^{-15}$, Figure 1F). GO term enrichment analyses further revealed that genes related to inflammatory processes were the most enriched among pop-DR genes showing a stronger response to infection in African-descent individuals (Figures 1G and S2C). Together, these results indicate that increased African ancestry predicts a stronger inflammatory response to infection.

We hypothesized that ancestry-associated differences in the transcriptional response to infection could translate into ancestry-associated differences in the ability of macrophages to clear the infection. We tested this hypothesis in a subset of 89 individuals by quantifying the number of bacteria remaining inside the macrophages right after the infection step (T0), 2 hr (T2), and 24 hr (T24) post-infection. For both bacteria, increased African ancestry predicted improved control of intracellular bacterial growth. This effect was particularly noticeable in our infection experiments with *Listeria*. Despite no significant difference in the initial number of bacteria infecting macrophages (Figure 2A, p = 0.95), the number of bacteria inside the macrophages of individuals with high levels of African ancestry at T24 was 3.2-fold lower than that of Europeans (Figure 2A, $p = 2.0 \times 10^{-4}$).

Finally, we tested if pop-DE genes were enriched among GWAS-associated genes. We found seven diseases for which susceptibility genes reported by GWAS were significantly enriched among genes classified as pop-DE, in at least one experimental condition (Figure 2B). Contributing to these enrichments are several HLA genes (*HLA-DQA1*, *HLA-DPA1*, *HLA-DRB1*, *HLA-DPB1*, *HLA-DRA*), known to be the main genetic risk factors for several immune disorders. Strikingly, six of these seven diseases (all but Parkinson's disease) are immune-related and tightly connected to a dysregulated inflammatory response. Further, among the diseases most significantly enriched for pop-DE genes were rheumatoid arthritis, systemic sclerosis, and ulcerative colitis, all of which have been reported to differ in incidence or disease severity between AA and EA individuals (Brinkworth and Barreiro, 2014; Pennington et al., 2009). Thus, ancestry-associated gene regulatory differences likely contribute to known ethnic disparities in inflammatory and autoimmune disease susceptibility, in part through affecting the ability of macrophages to control bacterial infections.

Gene Expression QTL in Non-infected and Infected Macrophages

To identify whether pop-DE and pop-DR genes are explained by genetic differences between European and African populations, we first mapped genetic variants that are associated with gene expression levels (i.e., eQTL) or transcript isoform usage (alternative splicing QTL [asQTL]) in the complete sample. To do so, we used a linear regression model that accounts for population structure and principal components of the expression data, thus limiting the effect of unknown confounding factors (see the STAR Methods for details). Given that our sample size is too small to robustly detect *trans*-acting eQTL, we focused our analyses on local associations that, for simplicity, we refer to as *cis*-eQTL. We define *cis*-eQTL and *cis*-asQTL here as SNPs located in the gene body or in the 100 kb flanking the gene of interest.

We identified *cis*-eQTL for 1,647 genes (14% of all genes tested; FDR < 0.01) in at least one of the experimental conditions (875 in non-infected macrophages, 1,087 in the *Listeria*-infected condition, and 983 in the *Salmonella*-infected condition; Figure 3A; Table S4A; Figure S3A for number of eQTL found at more relaxed cutoffs). Similarly, we detected a large number of *cis*-asQTL affecting the ratio of alternative isoforms used for the same gene (1,120 genes, 10% of all genes tested; FDR < 0.01 [Figure 3A; Table S4C]: 886 in non-infected macrophages, 746 in *Listeria*-infected samples, and 615 in *Salmonella*-infected samples).

Out of all genes with *cis*-eQTL, a large fraction (21.8%) were associated with an eQTL only in infected macrophages. In contrast, only 7.3% of genes showed an infection-specific *cis*-asQTL (Figures 3A and 3B). Infection-specific *cis*-eQTL

(B) Venn diagram illustrating the number of pop-DE genes (FDR < 0.05) in non-infected (black), *Listeria*-infected (yellow), and *Salmonella*-infected (orange) macrophages.

(C) Example of a gene, the chemokine *CCL15*, for which expression levels in all conditions are significantly associated with levels of African versus European ancestry. The average sequencing depth for each base (normalized per million mapped reads) is shown on the y axis.

(D) Example of three genes (*POLR1A*, *NDUFS5*, and *OAS1*) for which ancestry predicts differences in isoform usage.

(E) Example of three immune-related pop-DR genes. The y axis shows the log2 fold changes in gene expression levels in response to *Listeria* and *Salmonella*, as a function of continuous differences in African ancestry (x axis).

(F) Absolute difference in the log2 fold change response to *Salmonella* (top panel) and *Listeria* infection (bottom panel) between European and African individuals (x axis), among all pop-DR genes (red) and pop-DR genes associated with the inflammatory response (blue). The null expectation from permuting admixture levels across individuals is shown in light gray for comparison. A shift in the distribution to the right reflects a stronger response to infection in African-ancestry individuals.

(G) GO enrichment analysis for genes showing a significant interaction between ancestry and the response to *Salmonella*. Only GO terms with an enrichment at FDR < 0.1 are displayed, and GO terms are color-coded into functionally related terms based on the overlap among gene sets (Bindea et al., 2009). For each GO term, the average interaction effect is plotted on the x axis and the mean log2 fold change in gene expression levels in response to infection is plotted on the y axis. See also Tables S2 and S3.

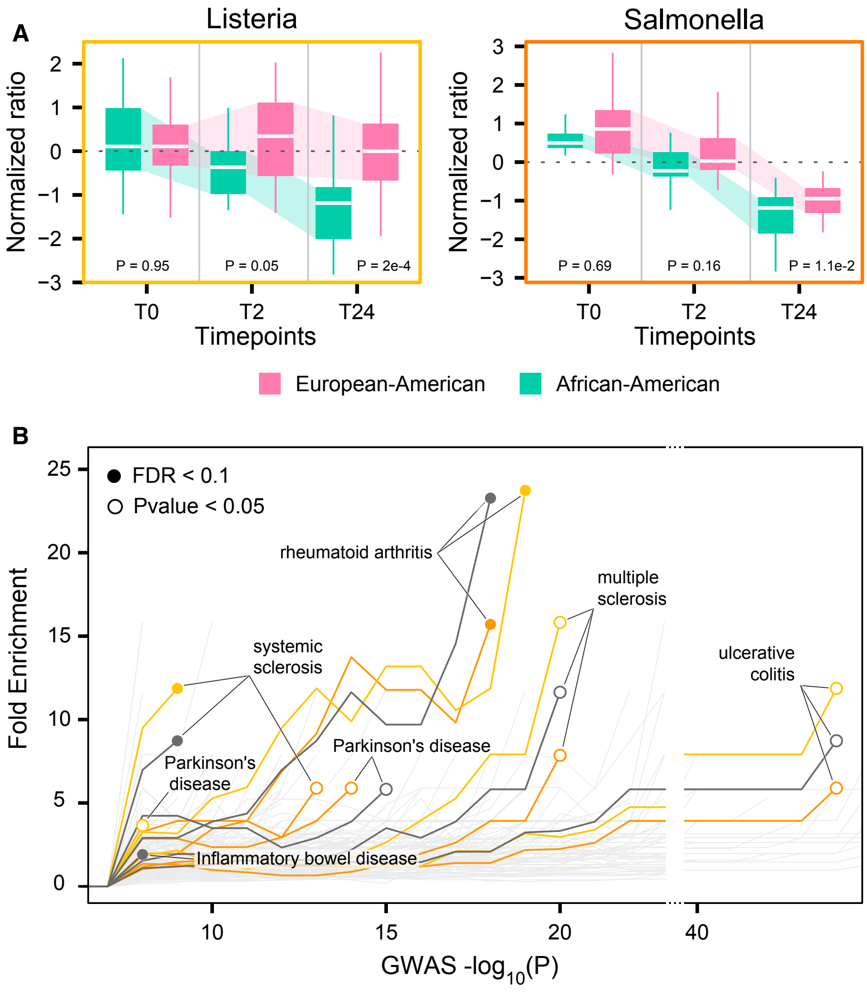


Figure 2. Increased African Ancestry Predicts Improved Control of Bacterial Growth inside Macrophages

(A) Boxplots showing the quantile normalized number of bacteria inside infected macrophages (y axis) immediately after infection (T0), 2 hr post-infection (T2), and 24 hr post-infection (T24) (x axis). We quantile normalized data across individuals and time points. Analyses were conducted using a continuous measure of ancestry; however, for visualization purposes, African and European samples were defined as those with an estimated African ancestry >75% (green) and <25% (pink), respectively.

(B) Fold enrichment of pop-DE genes (y axis) among genes identified in disease susceptibility GWAS, at progressively stringent p value thresholds (x axis). Grey, yellow and orange lines show significant enrichments (filled circles at an FDR < 0.1 and open circles at a nominal p < 0.05) for pop-DE genes identified in non-infected and *Listeria*- and *Salmonella*-infected macrophages, respectively. Light gray lines show non-significant diseases/traits.

See also Figure S2.

were further supported by analysis of allele-specific expression (ASE) levels, which provides independent but complementary evidence for functional *cis*-regulatory variation. As expected, genes with *cis*-eQTL were significantly enriched for genes with ASE, compared to the background of all 9,588 genes tested (Figure S3B, Fisher's exact test, $p < 1 \times 10^{-15}$ for all conditions). Further, genes harboring infection-specific eQTL also tended to exhibit infection-specific ASE in the same condition (*Listeria* or *Salmonella*) in which the eQTL was identified (Figure 3C, ~27 fold-enrichment of infection-specific ASE among infection-specific eQTL, relative to shared eQTL; $p < 1 \times 10^{-15}$). Thus, in agreement with previous studies (Fairfax et al., 2014; Lee et al., 2014), genotype-environment (G × E) interactions are common in the context of immune responses to infection, albeit more so for mean expression levels than alternative isoform usage.

A complementary approach to identifying G × E interactions for expression levels is to directly map response eQTL (reQTL): QTL associated with the magnitude of change in expression levels after infection (Barreiro et al., 2012; Çalışkan et al., 2015; Lee et al., 2014). In contrast to condition-specific eQTL (an extreme case of G × E interaction), reQTL can capture more subtle interaction effects: eQTL can be shared across conditions as long as their effect size differs between infected and non-infected samples. We detected 244 and 503 genes with a *cis*-reQTL (FDR < 0.01, Table S4B) for the response to *Listeria* and *Salmonella*, respectively. Interestingly, among genes associated with a *cis*-reQTL, we found several key regulators of the immune response, including the transcription factors *STAT4* and *IRF2* (Figure 3D). We also found *cis*-reQTL for known susceptibility loci for ulcerative colitis (e.g., *HLA-A*, *HLA-DQA2*, *PMPCA*), systemic lupus erythematosus (*ITGAX*, *HLA-DQA1*), and the infectious diseases hepatitis B and leprosy (e.g., *HLA-C*, *NOD2*).

To investigate the regulatory mechanisms that account for immune reQTL, we next profiled the genome-wide chromatin accessibility landscape of non-infected and *Listeria* and *Salmonella*-infected cells using assay for transposase-accessible chromatin using sequencing (ATAC-seq) (Buenrostro et al., 2013). This approach allowed us to identify transcription factor (TF) binding motifs likely to be occupied by their respective TFs, in both conditions (see the STAR Methods). We found that SNPs within accessible TF binding sites were greater than four times more likely to be identified as reQTL (Figure 3E). Further, reQTL in our analyses were strongly enriched (>20-fold) for PU.1 binding sites (a pioneer TF involved in regulating enhancer activity in macrophages) (Garber et al., 2012) and for virtually all TFs that orchestrate innate immune responses to infection (Figure 3E) (e.g., nuclear factor κB [NF-κB]: >50-fold; AP1: >55-fold; and IRFs: 14-fold for *Salmonella* only). In striking contrast, we found no such enrichment for eQTL identified in non-infected

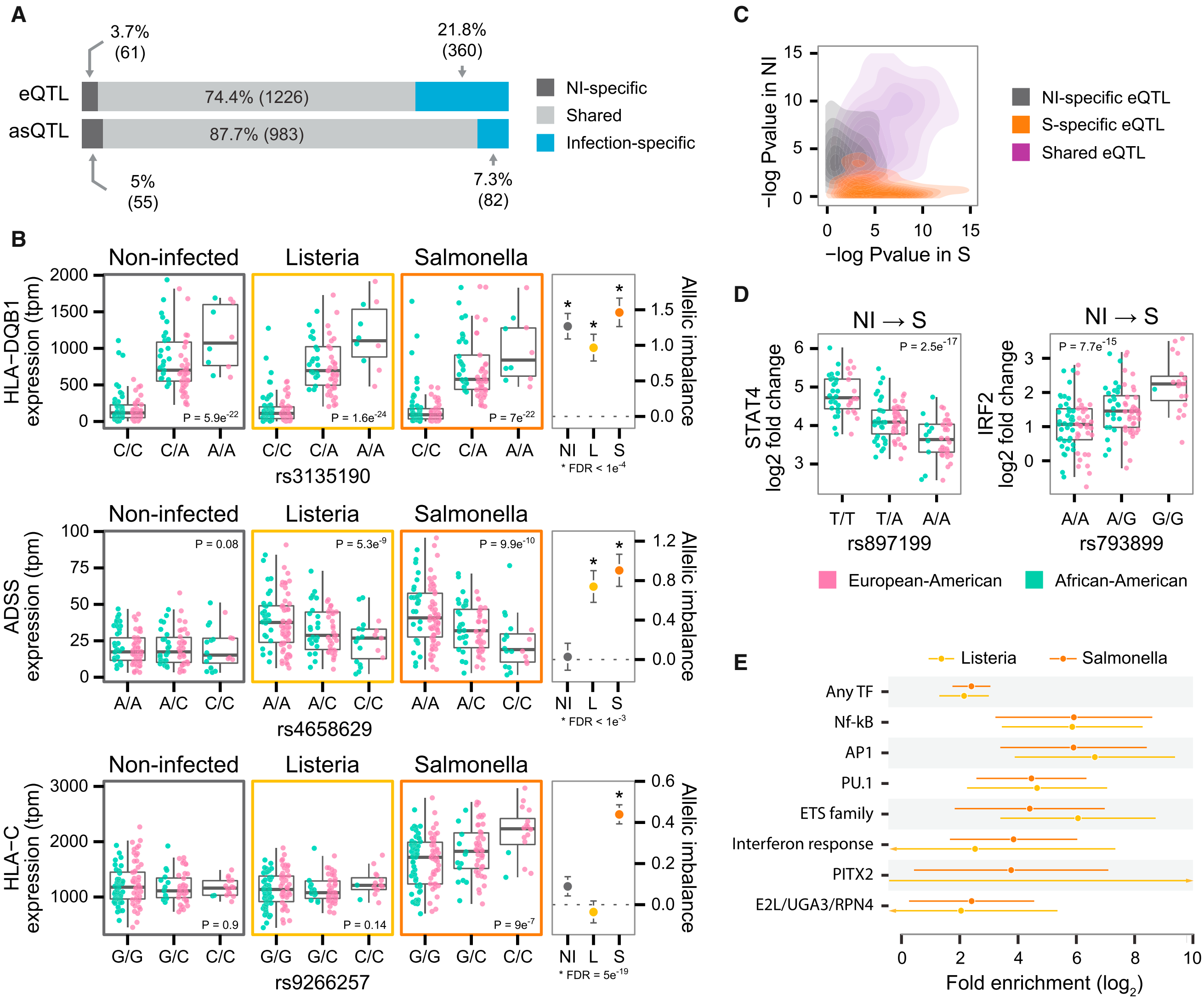


Figure 3. eQTL and ASE Analyses Reveal Extensive *cis*-Regulation of Gene Expression Responses to Pathogens in Macrophages

(A) Schematic representation of the number of *cis*-eQTL and *cis*-asQTL shared across all conditions, or only found in non-infected macrophages or *Listeria* and/or *Salmonella* infected macrophages. Infection-specific eQTL were defined as those showing very strong evidence of eQTL in the infected state (FDR always lower than 0.01), and very limited in the non-infected state (FDR always higher than 0.3).

(B) Examples of a *cis*-eQTL observed in all conditions (*HLA-DQB1*), a *cis*-eQTL observed only in infected macrophages (*ADSS*), and a *cis*-eQTL observed only in *Salmonella*-infected macrophages (*HLA-C*). Pink and green dots inside the boxplots distinguish African (>75% African ancestry) and European (<25% African ancestry) samples, respectively.

(C) Plot contrasting the evidence for ASE (-log10 p values) in non-infected macrophages (y axis) and in macrophages infected with *Salmonella* (x axis), for genes where we identified *cis*-eQTL in both conditions (purple), genes for which *cis*-eQTL were only found in non-infected macrophages (gray), and genes for which *cis*-eQTL were only found in *Salmonella*-infected macrophages (orange). Qualitatively similar results were obtained when contrasting non-infected and *Listeria*-infected cells (Figure S3C).

(D) Examples of two *cis*-reQTL where genotype (x axis) has a significant effect on the response of *STAT4* (left) and *IRF2* (right) to *Salmonella* infection.

(E) reQTL enrichments (x axis) in actively regulated TF binding sites annotated by ATAC-seq footprinting. Error bars show 95% confidence intervals. Binding sites were grouped into functionally overlapping “TF clusters” using sequence similarity and co-localization in the genome (Table S6; STAR Methods).

See also Table S4.

macrophages (p > 0.05 for NF-κB, AP1, and IRFs) (Figure S3D). These results show that reQTL variants are often conditionally silent in resting macrophages but become functionally relevant post-infection, and this transition is explained by disruption of binding sites for immune response-activated TFs.

Genetic Basis of Ancestry-Associated Differences in the Immune Response to Pathogens

We hypothesized that differences in allele frequencies for some of the eQTL identified above could explain the observed ancestry-associated differences in the transcriptional response

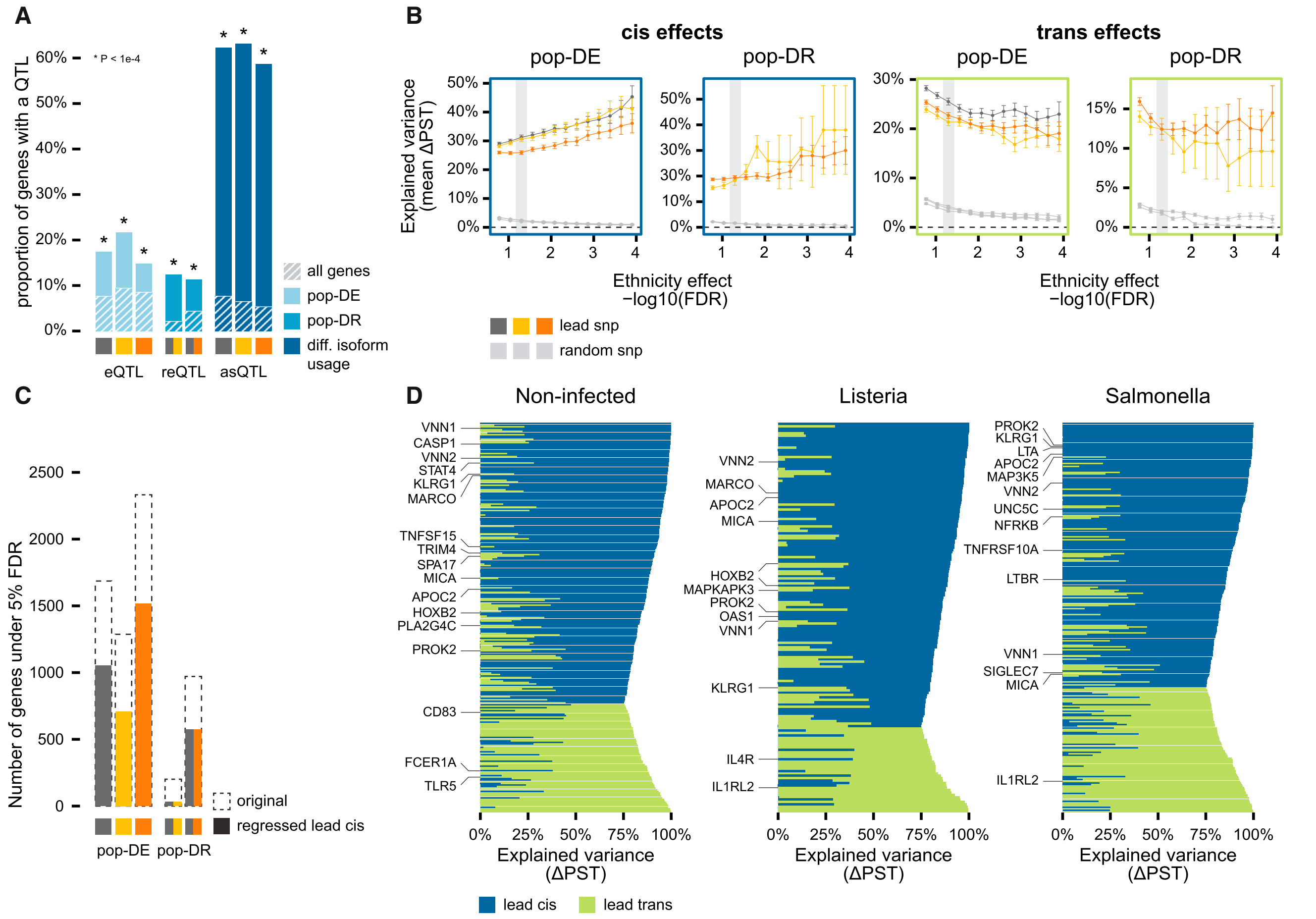


Figure 4. Contribution of *cis* and *trans* Genetic Variation to pop-DE and pop-DR Genes

(A) The proportion of pop-DE, pop-DR, and genes that exhibit ancestry-associated isoform usage that are associated with a *cis*-eQTL, *cis*-reQTL, or *cis*-asQTL, respectively (FDR < 0.01). Null expectations (based on the genome-wide proportion of genes associated with each QTL class) are shown in gray. Similar results are obtained when focusing on transcriptional QTL identified at an FDR of 0.05 (Figure S4B).

(B) Average ΔP_{ST} obtained (±SE) when regressing out the genotype effect of the lead *cis*- or the lead *trans*-SNP for pop-DE and pop-DR genes (y axis), defined using progressively stringent FDR cutoffs (x axis). Colored lines show average ΔP_{ST} values based on the real data; gray lines show the same values when regressing out the genotype effect of the lead SNP identified based on permuted genotypes.

(C) Number of genes identified as pop-DE and pop-DR at an FDR < 0.05 (y axis) before (dashed bars) and after (filled bars) regressing out the effect of the lead *cis*-SNP associated with these genes.

(D) Examples of genes for which the lead *cis*- (blue) or the lead *trans*-SNP (green), explains at least 75% of the differences in gene expression associated with African versus European ancestry.

to infection. In support of this hypothesis, we found that pop-DE genes were enriched up to 3.3-fold for genes with *cis*-eQTL ($p < 1 \times 10^{-10}$), and pop-DR genes were enriched up to 5.8-fold for genes with *cis*-reQTL ($p < 1 \times 10^{-10}$) (Figures 4A and S4A). Additionally, ~60% of genes that exhibited ancestry-associated isoform usage were associated with an asQTL (up to 24-fold enrichment, $p < 1 \times 10^{-10}$). Thus, although rare, ancestry-associated changes in isoform usage are largely genetically driven.

To explicitly quantify the contribution of our eQTL set to transcriptional differences detected between populations, we devised a new score based on P_{ST} estimates (Leinonen et al., 2013; Pujol et al., 2008). P_{ST} is the phenotypic analog of the population genetic parameter F_{ST}, providing a measure of the proportion of overall gene expression variance explained by between-population phenotypic divergence (as opposed to within-population diversity). P_{ST} values range from 0 to 1, with values close to 1 implying that the majority of a gene's expression variance is due to differences between populations. Our score, deltaP$_{ST}$ (ΔP_{ST}), is defined as the difference between P_{ST} values before and after regressing out the effect of the *cis*-SNP that was most strongly associated with the target gene's expression level (regardless of significance level), divided by the P_{ST} value observed before removing the genotype effect. ΔP_{ST} therefore quantifies the proportion of ancestry-associated expression level differences that stem from the strongest *cis*-associated variant.

Among all pop-DE genes, we found that *cis*-regulatory variants explained an average of 31%, 31%, and 26% of ancestry-related differences in expression observed in non-infected, *Listeria*-infected and *Salmonella*-infected samples, respectively (Figure 4B). Further, the larger the effect of ancestry in the original pop-DE analysis, the larger the contribution of *cis*-regulatory variation to these differences: for pop-DE genes identified at a stringent FDR of 1×10^{-4}, *cis*-regulatory variation explained close to 50% (on average) of ancestry effects (Figure 4B). We observed a similar pattern for pop-DR genes after regressing out the genotype effect of the lead *cis*-reQTL SNP (Figure 4B). In support of the substantial role of *cis*-regulatory variation in explaining pop-DE and pop-DR genes, gene expression values for 30% and 45% of pop-DE and pop-DR genes, respectively, were no longer significantly associated with ancestry once we regressed out *cis*-genetic effects (Figure 4C). Importantly, ΔP_{ST} values never exceeded 5% when we regressed out either (1) the genotype effect of randomly selected SNPs matched for the allele frequency of the lead *cis*-SNP, or (2) lead *cis*-SNPs identified after permuting the genotype data. Thus, our results cannot be simply explained by population structure (Figure 4B).

Based on their known importance in the genetic control of gene regulation and because of power limitations, our main analysis of ancestry-associated gene expression patterns focused on the role of *cis*-eQTL. However, in a separate analysis, we recalculated ΔP_{ST} using the lead *trans*-SNP for each gene in place of the lead *cis*-SNP (although only 51, 21, and 22 *trans*-eQTL genes survived genome-wide multiple testing correction ($FDR < 0.1$) in non-infected, *Listeria*-infected and *Salmonella*-infected samples, respectively). Intriguingly, we found that lead *trans*-SNPs accounted for an average of ~23% and ~20% of ancestry effects on gene expression levels for pop-DE and pop-DR genes, respectively (Figure 4B; at least 2-fold more than estimates based on permuted data, $p < 1 \times 10^{-10}$). These results suggest that lead *trans*-SNPs, although difficult to detect at a genome-wide significance level, are enriched for true *trans*-associations that could be resolved with larger sample sizes. Together, a single *cis*- or *trans*-acting variant was sufficient to explain almost all ancestry effects ($\Delta P_{ST} > 75\%$) on gene expression levels for 804 pop-DE genes and pop-DR genes (Figure 4D), including for master regulators of the immune response such as *CASP1*, *STAT4*, and *MICA*. Our results thus provide a comprehensive genome-wide map of *cis*- and *trans*-genetic variants associated with African and European ancestry-related differences in the immune response to infection.

Natural Selection and Genetic Ancestry Effects on Gene Expression Divergence

Finally, we sought to determine the impact of recent local positive selection in either African or European populations on ancestry-related divergence in gene expression levels. To do so, we first calculated F_{ST} values between the Yoruba African (YRI) and the western European population (CEU) in Phase 3 data from the 1000 Genomes Project (Auton et al., 2015). To generate gene-specific estimates, we averaged F_{ST} values for variants within a window of 10 kb around the transcription start site (TSS) of each gene we analyzed (11,914 genes). As a complementary approach, we also calculated integrated haplotype scores (iHS) for all SNPs with a minor allele frequency (MAF) >5% in the CEU and YRI samples. In contrast to F_{ST}, iHS is a within-population measure of recent positive selection that is not affected by the levels of population differentiation (Voight et al., 2006).

Our analyses identified significantly higher mean F_{ST} values among genes that were pop-DE, pop-DR, or showing differences in isoform usage between populations ($p \leq 1 \times 10^{-3}$; Figures 5A and S5A for similar results when using alternative window sizes). Further, variants identified as *cis*-eQTL were significantly enriched (~2-fold) for high iHS values (i.e., iHS > 99th percentile of genome-wide distribution, Figure 5B, $p < 1 \times 10^{-8}$), consistent with the importance of regulatory genetic variation in recent human evolution (Fraser, 2013). *cis*-reQTL and *cis*-asQTL were even more strongly enriched among high iHS values (up to 3.6-fold; Figure 5B, $p < 1 \times 10^{-5}$).

Overall, within the set of *cis*-eQTL-, *cis*-reQTL-, or *cis*-asQTL-associated genes, 258 carried a signature of recent positive selection in either CEU or YRI samples ($|iHS| \geq$ 99th percentile of the genome-wide distribution) (Figure 5C; Table S5A). These variants were also significantly enriched for high XP-EHH values (Sabeti et al., 2007) (~6-fold, $p < 1 \times 10^{-10}$, Figure S5C), further supporting that these variants have been important in recent, population-specific human adaptation. However, because outlier methods for detecting selection can be susceptible to false positives (Kelley et al., 2006), we complemented our iHS analysis with a model-based approach. Specifically, we compared the observed iHS value for each putatively selected allele to those observed under neutral coalescent simulations matched to the known demographic history of African and European populations (Gutenkunst et al., 2009), the candidate allele's observed frequency, and the local recombination rate. The vast majority (92%) of all sites tested exhibited significantly larger observed iHS statistics than expected under a neutral model ($p < 0.01$, Table S5B), providing strong convergent support for recent positive selection at these loci. Far more of these genes are pop-DE or pop-DR than expected by chance (47% and 23%, respectively: Figure 5D, $p < 0.001$), showing that natural selection has contributed to present-day inter-population differences in innate immune responses to infection.

Neanderthal ancestry makes up ~2% of the ancestry of living humans found outside of Africa (Kelso and Prüfer, 2014). It is therefore plausible that interbreeding between Neanderthal and modern human populations could also contribute to some of the ancestry-related differences in gene expression we observed, especially if it enabled the ancestors of modern Europeans to more rapidly adapt to a new pathogen environment (Ségurel and Quintana-Murci, 2014). To test this hypothesis, we identified sites where the derived allele is shared between Neanderthals and non-African populations, but is absent in sub-Saharan Africans samples considered. This class of sites, which we call ''Neanderthal-like sites'' (NLS), is a conservative indicator of Neanderthal introgression (Sankararaman et al., 2014). Among the 18,862 NLS tested in our *cis*-QTL analyses, 297 were significantly associated with transcriptional variation of 145 genes (NLS-QTL). Among these 145 genes, 46% ($FDR < 0.05$) were differentially regulated in at least one experimental condition (non-infected, *Listeria*-infected, *Salmonella*-infected, or in the response to either type of infection) between Europeans

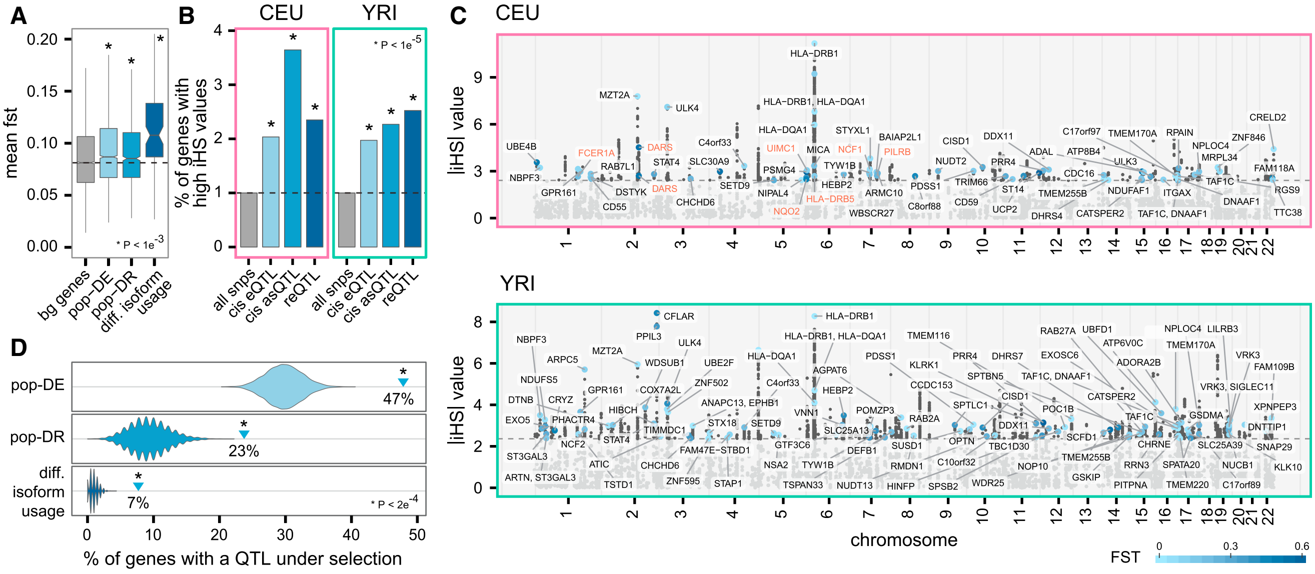


Figure 5. Natural Selection on eQTL and Its Contribution to Ancestry-Associated Regulatory Differences

(A) Mean F_{ST} values in a window of ±10 kb around the TSS of all genes, pop-DE genes, pop-DR genes, and genes showing differences in isoform usage between populations (top panel).

(B) Proportion of all SNPs, *cis*-eQTL, *cis*-reQTL, and *cis*-asQTL with an iHS value above the 99th percentile of the genome-wide distribution in the CEU ($|iHS| > 2.70$) and the YRI ($|iHS| > 2.68$) populations. See Figure S5B for similar results considering QTL identified at an FDR < 0.05 (instead of 0.01).

(C) Manhattan plot of a genome-wide scan for selection in CEU (top) and YRI (bottom) for SNPs identified as regulatory QTL in macrophages. The dashed line represents the 99th percentile of the genome-wide distribution. Darker shades of blue represent larger F_{ST} values for SNPs with elevated |iHS| values; blue circled dots highlight genes that show one or more transcriptional associations with African versus European ancestry. Genes in red are regulated by NLS with elevated |iHS| values in CEU ($|iHS| > 2.7$), supporting adaptive introgression from Neanderthals into the ancestors of modern Europeans.

(D) Proportion of genes regulated by eQTL and targeted by recent positive selection (among the 258 represented by the blue circles in C) that are pop-DE, pop-DR, or show population differences in isoform usage (blue triangles), compared to random expectations when sampling the same total number of genes 10,000 times from all genes tested (violin plots).

See also Table S5.

and Africans (63% at a more relaxed FDR < 0.1). Thus, a non-negligible proportion of ancestry-related gene expression divergence probably results from introgression of functional Neanderthal variants into the ancestors of modern Europeans. Interestingly, some of these variants (n = 16) also have elevated iHS values ($|iHS| \geq 2$) (Figure 5C; Table S5A) and therefore represent new candidates for adaptive introgression in humans.

DISCUSSION

Together, our results provide a comprehensive characterization of genes for which the transcriptional responses of primary cells to live pathogenic bacteria differs depending on European versus African ancestry. We show that 34% of genes expressed in macrophages show at least one type of ancestry-related transcriptional divergence, whether in the form of differences in gene expression (30%), the transcriptional response to infection (9.3%), or, less commonly, differences in isoform usage (1%). Notably, the modest contribution of differences in isoform usage to ancestry-related expression levels differs from previous results in lymphoblastoid cell lines (LCLs), where they were found to be quite common (Lappalainen et al., 2013). The discrepancy between our results and those reported for LCLs may be related to differences in the experimental procedures used to produce the two sets of LCL lines, which were generated more than 20 years apart (Dausset et al., 1990).

One of the most striking observations from our study was the markedly stronger response to infection induced in macrophages from individuals of African descent, particularly among inflammatory response genes. This result agrees with previous reports showing that AAs have higher frequencies of alleles associated with an increased pro-inflammatory response (Ness et al., 2004), increased levels of circulating C-reactive protein (Kelley-Hedgepeth et al., 2008), and a much higher rate of inflammatory diseases than EA individuals (Pennington et al., 2009). Although the exact causal link between ancestry and the pro-inflammatory response has yet to be established, we speculate that the stronger inflammatory response associated with African ancestry accounts for the increased ability of macrophages in African ancestry individuals to control bacterial growth post-infection.

Nevertheless, the evolutionary pressures that explain these differences remain an open question. One possibility is that, after human populations migrated out of Africa, they were exposed to lower pathogen levels (Guernier et al., 2004), which reduced the need for strong, costly pro-inflammatory signals. Change in this direction may have been favored due to the detrimental consequences of acute or chronic inflammation, which are key contributors to the development of autoinflammatory and autoimmune diseases (Okin and Medzhitov, 2012). This hypothesis is consistent with previous reports showing a signature of positive selection in Europeans on a high-frequency non-synonymous variant in the Toll-like receptor 1 gene, which is also associated with impaired NF-κB-mediated signaling (Barreiro et al., 2009). Alternatively, the weaker inflammatory response detected in Europeans could have resulted from relaxation of selective constraint in an environment where the pathogen burden was reduced, or at least different in nature, from that found in Africa.

Because our samples were derived from individuals with their own unknown life histories and environmental exposures, the ancestry-related differences we observed could be explained by both environmental and genetic factors. However, our eQTL analyses suggest that genetic contributions are probably substantial. We estimate that, on average, ∼30% and 20% of ancestry-associated expression differences in pop-DE genes are accounted for by *cis*- and *trans*-regulatory variants, respectively. Further, among the genes with the most robust association with genetic ancestry (pop-DE genes with $\mathrm{FDR} < 1 \times 10^{-4}$), putatively *cis*-acting variants explain up to ∼50% of ancestry effects. Notably, these numbers probably underestimate the true genetic contribution to ancestry-related differences in gene expression, given our low power to detect *trans* associations, our exclusion of non-SNP regulatory variants, which may also influence gene expression (Gymrek et al., 2016), our conservative assumption that genes have only one major *cis*-eQTL (many genes have at least two independent *cis*-eQTL) (Lappalainen et al., 2013); and the fact that we limited our *cis*-eQTL mapping to a 100-kb window around the targeted gene.

The extent to which positive selection has contributed to recent human evolution remains a matter of intense debate (Enard et al., 2014; Fagny et al., 2014; Hernandez et al., 2011). Here, we show that variants associated with regulatory QTL are strongly enriched for signatures of recent selection, supporting an important role of adaptive regulatory variation in recent human evolution. More specifically, our results suggest that a significant fraction of population differences in transcriptional responses to infection are a direct consequence of local adaptation driven by regulatory variants. Notably, several positively selected regulatory QTL (or SNPs in strong LD with them [$r^2 > 0.8$]) have been associated with common diseases by GWAS, further reinforcing the link between past selection and present-day susceptibility to disease (Barreiro and Quintana-Murci, 2010; Brinkworth and Barreiro, 2014). Some examples include positively selected variants affecting the expression of *HLA-DQA1*, the major genetic susceptibility factor for celiac disease (Abadie et al., 2011), *ERAP2*, a susceptibility factors for Crohn's disease (Jostins et al., 2012), and the transcription factor *IRF5*, which is associated with systemic lupus erythematosus, rheumatoid arthritis, ulcerative colitis, and systemic sclerosis (reviewed in Eames et al., 2016).

Finally, our results provide new insight into the contribution of adaptive introgression from admixture with Neanderthals to the diversification of the immune system among modern human populations. We found 17 positively selected NLS regulatory-QTL (associated with 16 genes) that are candidates for adaptive introgression in humans. These genes include previously identified candidates such as *TLR1* (Dannemann et al., 2016; Deschamps et al., 2016) but also a large set of loci that have not previously been associated with adaptive introgression. For example, one of the strongest signatures of selection was found for eQTL for *DARS*, a gene associated with neuroinflammatory and white matter disorders (Wolf et al., 2015). However, in agreement with evidence that most introgressed variation from Neanderthals was probably deleterious (Sankararaman et al., 2014; Vernot and Akey, 2014), as putative cases of adaptive introgression remain relatively rare.

All data generated in this study are freely accessible via a custom web-based browser that enables easy querying and visualization of ancestry-related transcriptional differences and associated QTL. This resource, the ImmunPop QTL browser (http://www.immunpop.com), should serve as a useful tool for fine mapping of genetic association signals and for the continued quest to understand how pathogens have shaped global human population diversity today.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Sample Collection
- METHOD DETAILS
  - ○ Isolation of Monocytes and Differentiation of Macrophages
  - ○ Bacterial Preparation and Infection of Macrophages
  - ○ Estimation of the Number of Infected Macrophages
  - ○ RNA Extraction, Library Preparation, and Sequencing
  - ○ ATAC-Seq Library Preparation and Sequencing
  - ○ DNA Extraction and Genome-wide Genotyping
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Imputation
  - ○ Estimation of Genome-wide Admixture Levels
  - ○ Estimation of Gene- and Isoform-Level Expressions
  - ○ Differences in Expression between Populations and in Response to Infection
  - ○ False Discovery Rate Estimations
  - ○ Differential Isoform Usage between Populations
  - ○ Enrichment of GWAS-Associated Genes among pop-DE Genes
  - ○ Genotype-Phenotype Association Analysis
  - ○ Allele-Specific Expression Detection
  - ○ ATAC-Seq Data Processing and Footprinting Analysis
  - ○ Enrichment of TF Binding Sites among eQTL and reQTL
  - ○ Genetic Control of Ancestry Effects on Gene Expression
  - ○ Gene Ontology Enrichment Analysis
  - ○ Signatures of Selection
  - ○ Coalescence Neutral Simulations for Evaluating Putatively Selected eQTL Sites

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2016.09.025.

AUTHOR CONTRIBUTIONS

L.B.B. conceived and directed the study. A.D., A.P.S., V.Y., and F.L. performed experimental work. Y.N., J.S., and G.B. led the computational analyses with contributions from Z.A.S., A.F., A.P., A.J.S., J.-C.G., R.B., R.D.H., R.P.-R., and L.B.B. Y.N. developed and implemented the ImmunPop QTL browser with help from S.H. L.B.B., J.T., G.B., and J.S. wrote the paper, with input from all authors.

ACKNOWLEDGMENTS

We thank members of the L.B.B. lab for helpful discussions and comments on the manuscript, Danielle Malo for a gift of the *Listeria* and *Salmonella* strains used in this study, and L. Tailleux for advice on the infection experiments. This study was funded by grants to L.B.B. from the Canadian Institutes of Health Research (301538 and 232519), the Human Frontiers Science Program (CDA-00025/2012), and the Canada Research Chairs Program (950-228993). We thank Calcul Québec and Compute Canada for providing access to the supercomputer Briaree from the University of Montreal. Y.N. was supported by a fellowship from the Réseau de Médecine Génétique Appliquée (RMGA); J.S. by a fellowship from the Fonds de recherche du Québec-Nature et technologies (FRQNT), and G.B. and J.S. by a fellowship from the Fonds de recherche du Québec-Santé (FRQS).

Received: April 13, 2016
Revised: July 28, 2016
Accepted: September 15, 2016
Published: October 20, 2016

REFERENCES

Abadie, V., Sollid, L.M., Barreiro, L.B., and Jabri, B. (2011). Integration of genetic and immunological insights into a model of celiac disease pathogenesis. Annu. Rev. Immunol. *29*, 493–525.

Aitchison, J. (1982). The statistical analysis of compositional data. J. R. Stat. Soc. B *44*, 139–177.

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

Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R.; 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Barreiro, L.B., and Quintana-Murci, L. (2010). From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat. Rev. Genet. *11*, 17–30.

Barreiro, L.B., Ben-Ali, M., Quach, H., Laval, G., Patin, E., Pickrell, J.K., Bouchier, C., Tichit, M., Neyrolles, O., Gicquel, B., et al. (2009). Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. PLoS Genet. *5*, e1000562.

Barreiro, L.B., Tailleux, L., Pai, A.A., Gicquel, B., Marioni, J.C., and Gilad, Y. (2012). Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. Proc. Natl. Acad. Sci. USA *109*, 1204–1209.

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.H., Pagès, F., Trajanoski, Z., and Galon, J. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics *25*, 1091–1093.

Bonnevie-Nielsen, V., Field, L.L., Lu, S., Zheng, D.J., Li, M., Martensen, P.M., Nielsen, T.B., Beck-Nielsen, H., Lau, Y.L., and Pociot, F. (2005). Variation in antiviral 2′,5′-oligoadenylate synthetase (2‘5’AS) enzyme activity is controlled by a single-nucleotide polymorphism at a splice-acceptor site in the OAS1 gene. Am. J. Hum. Genet. *76*, 623–633.

Brinkworth, J.F., and Barreiro, L.B. (2014). The contribution of natural selection to present-day susceptibility to chronic inflammatory and autoimmune disease. Curr. Opin. Immunol. *31*, 66–78.

Bryc, K., Auton, A., Nelson, M.R., Oksenberg, J.R., Hauser, S.L., Williams, S., Froment, A., Bodo, J.M., Wambebe, C., Tishkoff, S.A., and Bustamante, C.D. (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc. Natl. Acad. Sci. USA *107*, 786–791.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218.

Çalışkan, M., Baker, S.W., Gilad, Y., and Ober, C. (2015). Host genetic variation influences gene expression response to rhinovirus infection. PLoS Genet. *11*, e1005111.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al.; 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

Dannemann, M., Andrés, A.M., and Kelso, J. (2016). Introgression of Neandertal- and Denisovan-like Haplotypes contributes to adaptive variation in human Toll-like receptors. Am. J. Hum. Genet. *98*, 22–33.

Dausset, J., Cann, H., Cohen, D., Lathrop, M., Lalouel, J.M., and White, R. (1990). Centre d’etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. Genomics *6*, 575–577.

Delaneau, O., Zagury, J.F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods *10*, 5–6.

Deschamps, M., Laval, G., Fagny, M., Itan, Y., Abel, L., Casanova, J.L., Patin, E., and Quintana-Murci, L. (2016). Genomic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. Am. J. Hum. Genet. *98*, 5–21.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Eames, H.L., Corbin, A.L., and Udalova, I.A. (2016). Interferon regulatory factor 5 in human autoimmunity and murine models of autoimmune disease. Transl. Res. *167*, 167–182.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. Math. Geol. *35*, 279–300.

Enard, D., Messer, P.W., and Petrov, D.A. (2014). Genome-wide signals of positive selection in human evolution. Genome Res. *24*, 885–895.

Excoffier, L., and Lischer, H.E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol. Ecol. Resour. *10*, 564–567.

Fagny, M., Patin, E., Enard, D., Barreiro, L.B., Quintana-Murci, L., and Laval, G. (2014). Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. Mol. Biol. Evol. *31*, 1850–1868.

Fairfax, B.P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R., McGee, C., and Knight, J.C. (2014). Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science *343*, 1246949.

Fraser, H.B. (2013). Gene expression drives local adaptation in humans. Genome Res. *23*, 1089–1096.

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admetlla, A., Pattini, L., and Nielsen, R. (2011). Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genet. *7*, e1002355.

Garber, M., Yosef, N., Goren, A., Raychowdhury, R., Thielke, A., Guttman, M., Robinson, J., Minie, B., Chevrier, N., Itzhaki, Z., et al. (2012). A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. Mol. Cell *47*, 810–822.

Guernier, V., Hochberg, M.E., and Guégan, J.F. (2004). Ecology drives the worldwide distribution of human diseases. PLoS Biol. *2*, e141.

Gusev, A., Lee, S.H., Trynka, G., Finucane, H., Vilhjálmsson, B.J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium; SWE-SCZ Consortium (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am. J. Hum. Genet. *95*, 535–552.

Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., and Bustamante, C.D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. *5*, e1000695.

Gymrek, M., Willems, T., Guilmatre, A., Zeng, H., Markus, B., Georgiev, S., Daly, M.J., Price, A.L., Pritchard, J.K., Sharp, A.J., and Erlich, Y. (2016). Abundant contribution of short tandem repeats to gene expression variation in humans. Nat. Genet. *48*, 22–29.

Harvey, C.T., Moyerbrailean, G.A., Davis, G.O., Wen, X., Luca, F., and Pique-Regi, R. (2015). QuASAR: quantitative allele-specific analysis of reads. Bioinformatics *31*, 1235–1242.

Hernandez, R.D., Kelley, J.L., Elyashiv, E., Melton, S.C., Auton, A., McVean, G., Sella, G., and Przeworski, M.; 1000 Genomes Project (2011). Classic selective sweeps were rare in recent human evolution. Science *331*, 920–924.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA *106*, 9362–9367.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G.R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. *44*, 955–959.

Hudson, R.R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics *18*, 337–338.

Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Schumm, L.P., Sharma, Y., Anderson, C.A., et al.; International IBD Genetics Consortium (IIBDGC) (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature *491*, 119–124.

Karlsson, E.K., Kwiatkowski, D.P., and Sabeti, P.C. (2014). Natural selection and infectious disease in human populations. Nat. Rev. Genet. *15*, 379–393.

Kelley, J.L., Madeoy, J., Calhoun, J.C., Swanson, W., and Akey, J.M. (2006). Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res. *16*, 980–989.

Kelley-Hedgepeth, A., Lloyd-Jones, D.M., Colvin, A., Matthews, K.A., Johnston, J., Sowers, M.R., Sternfeld, B., Pasternak, R.C., and Chae, C.U.; SWAN Investigators (2008). Ethnic differences in C-reactive protein concentrations. Clin. Chem. *54*, 1027–1037.

Kelso, J., and Prüfer, K. (2014). Ancient humans and the origin of modern humans. Curr. Opin. Genet. Dev. *29*, 133–138.

Krishnamoorthy, K., and Yu, J. (2004). Modified Nel and Van der Merwe test for the multivariate Behrens–Fisher problem. Stat. Probab. Lett. *66*, 161–169.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Lee, M.N., Ye, C., Villani, A.C., Raj, T., Li, W., Eisenhaure, T.M., Imboywa, S.H., Chipendo, P.I., Ran, F.A., Slowikowski, K., et al. (2014). Common genetic variants modulate pathogen-sensing responses in human dendritic cells. Science *343*, 1246980.

Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics *28*, 882–883.

Leinonen, T., McCairns, R.J., O'Hara, R.B., and Merilä, J. (2013). Q(ST)-F(ST) comparisons: evolutionary and ecological insights from genomic heterogeneity. Nat. Rev. Genet. *14*, 179–190.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lischer, H.E.L., and Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics *28*, 298–299.

Ness, R.B., Haggerty, C.L., Harger, G., and Ferrell, R. (2004). Differential distribution of allelic variants in cytokine genes among African Americans and White Americans. Am. J. Epidemiol. *160*, 1033–1038.

Okabe, Y., and Medzhitov, R. (2016). Tissue biology perspective on macrophages. Nat. Immunol. *17*, 9–17.

Okin, D., and Medzhitov, R. (2012). Evolution of inflammatory diseases. Curr. Biol. *22*, R733–R740.

Pennington, R., Gatenbee, C., Kennedy, B., Harpending, H., and Cochran, G. (2009). Group differences in proneness to inflammation. Infect. Genet. Evol. *9*, 1371–1380.

Pique-Regi, R., Degner, J.F., Pai, A.A., Gaffney, D.J., Gilad, Y., and Pritchard, J.K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. *21*, 447–455.

Pujol, B., Wilson, A.J., Ross, R.I., and Pannell, J.R. (2008). Are Q(ST)-F(ST) comparisons for natural populations meaningful? Mol. Ecol. *17*, 4782–4785.

Richardus, J.H., and Kunst, A.E. (2001). Black-white differences in infectious disease mortality in the United States. Am. J. Public Health *91*, 1251–1253.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al.; International HapMap Consortium (2007). Genome-wide detection and characterization of positive selection in human populations. Nature *449*, 913–918.

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., and Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. Nature *507*, 354–357.

Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. Genome Res. *22*, 1748–1759.

Ségurel, L., and Quintana-Murci, L. (2014). Preserving immune diversity through ancient inheritance and admixture. Curr. Opin. Immunol. *30*, 79–84.

Shabalin, A.A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics *28*, 1353–1358.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA *100*, 9440–9445.

Szpiech, Z.A., and Hernandez, R.D. (2014). selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. Mol. Biol. Evol. *31*, 2824–2827.

Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.M., Doumbo, O., et al. (2009). The genetic structure and history of Africans and African Americans. Science *324*, 1035–1044.

van de Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J.K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat. Methods *12*, 1061–1063.

Vernot, B., and Akey, J.M. (2014). Resurrecting surviving Neandertal lineages from modern human genomes. Science *343*, 1017–1021.

Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. PLoS Biol. *4*, e72.

Wang, L., Wang, S., and Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics *28*, 2184–2185.

Wen, X., Lee, Y., Luca, F., and Pique-Regi, R. (2016). Efficient integrative multi-SNP association analysis via deterministic approximation of posteriors. Am. J. Hum. Genet. *98*, 1114–1129.

Wolf, N.I., Toro, C., Kister, I., Latif, K.A., Leventer, R., Pizzino, A., Simons, C., Abbink, T.E., Taft, R.J., van der Knaap, M.S., and Vanderver, A. (2015). DARS-associated leukoencephalopathy can mimic a steroid-responsive neuroinflammatory disorder. Neurology *84*, 226–230.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Antibody against CD14	BD Biosciences	Cat#:555398; RRID: AB_395799
Antibody against CD1a	BD Biosciences	Cat#:555807; RRID: AB_396141
Antibody against CD83	BD Biosciences	Cat#:556855; RRID: AB_396526
Antibody against HLA-DR	BD Biosciences	Cat#:555561; RRID: AB_395943
Chemicals, Peptides, and Recombinant Proteins		
FICOLL-PAQUE PREMIUM	GE Healthcare	Cat#:17-5446-52
GENTAMICIN REAGENT SOLUTION LIQUID 10ML	Thermo Fisher Scientific	Cat#:15710-054
IGEPAL CA-630	Sigma Aldrich	Cat#:I3021-50ML
SYBR Green I Nucleic Acid Gel Stain - 10,000X concentrate in DMSO (500ul)	Thermo Fisher Scientific	Cat#:S7563
Triton X-100	Sigma Aldrich	Cat#:X100-500ML
FBS premium, US origin	WISENT	Cat#:80150
Human CD14 microbeads for Macs	Miltenyi Biotec	Cat#:130-050-201
L-glutamine	WISENT	Cat#:609-065-EL
NEBNext High-Fidelity 2X PCR Master Mix	New England Biolabs	Cat#:M0541S
Recombinant Human M-CSF, Animal- Free	R&D Systems	Cat#:AFL216
RPMI-1640	HyClone	Cat#:SH30096.01
Tryptic Soy Broth (TSB) (BBL Trypticase Soy Broth)	BD Biosciences	Cat#:211768
Trypticase Soy agar	BD Biosciences	Cat#:221283
Critical Commercial Assays		
Nextera DNA Sample Preparation Kit (24 samples)	Illumina	Cat#:FC-121-1030
Nextera index kit (24 indexes, 96 samples)	Illumina	Cat#:FC-121-1011
TRUSEQ RNA SAMPLE PREP KIT V2, SET A	Illumina	Cat#:RS-122-2001
DNA extraction kit (Gentra Systems)	QIAGEN	Cat#:1042606
MinElute PCR Purification Kit (50)	QIAGEN	Cat#:28004
miRNeasy Mini kit	QIAGEN	Cat#:217004
RNA Nano Chips	Agilent Technologies	Cat#:5067-1521
Deposited Data		
Raw and analyzed data	This study	GEO: GSE81046
Experimental Models: Organisms/Strains		
Listeria monocytogenes	This study	N/A
Salmonella typhimurium	This study	N/A
Software and Algorithms		
Impute2 v 2.3.0	Howie et al. (2012)	https://mathgen.stats.ox.ac.uk/impute/impute_v2.html
shapeIT v2.r790	Delaneau et al. (2013)	https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html
PLINK 1.9	Chang et al. (2015)	https://www.cog-genomics.org/plink2
RseQC	Wang et al. (2012)	http://rseqc.sourceforge.net
R		https://www.r-project.org/
edgeR	Robinson et al. (2010)	https://bioconductor.org/packages/release/bioc/html/edgeR.html
Limma	Ritchie et al. (2015)	https://bioconductor.org/packages/release/bioc/html/limma.html

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Matrix eQTL	Shabalin (2012)	http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/
Sva	Leek et al. (2012)	https://www.bioconductor.org/packages/release/bioc/html/sva.html
cbcbSEQ	Okrah n. d.	https://github.com/kokrah/cbcbSEQ
WASP	van de Geijn et al. (2015)	https://www.encodeproject.org/software/wasp/
SAMtools	Li et al. (2009)	http://samtools.sourceforge.net
QuASAR	Harvey et al. (2015)	https://github.com/piquelab/QuASAR
ADMIXTURE	Alexander et al. (2009)	https://www.genetics.ucla.edu/software/admixture/
Trim Galore!	Krueger n.d.	http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
Picard-tools	Broad Institute	http://broadinstitute.github.io/picard/
Bowtie 2	Langmead and Salzberg (2012)	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Centipede	Pique-Regi et al. (2011)	http://centipede.uchicago.edu/
STAR	Dobin et al. (2013)	https://github.com/alexdobin/STAR
RSEM	Li and Dewey (2011)	http://deweylab.github.io/RSEM/
Vcftools	Danecek et al. (2011)	http://vcftools.sourceforge.net/
CLUEGO	Bindea et al. (2009)	http://apps.cytoscape.org/apps/cluego
TORUS	Wen et al. (2016)	https://github.com/xqwen/dap/tree/master/torus_src
PGDSpider	Lischer and Excoffier (2012)	http://www.cmpg.unibe.ch/software/PGDSpider/
Arlequin	Excoffier and Lischer (2010)	http://cmpg.unibe.ch/software/arlequin35/
Selscan v1.1.0b	Szpiech and Hernandez, 2014	https://github.com/szpiech/selscan
Other		
eQTL visualization web browser		http://www.immunpop.com

CONTACT FOR REAGENT AND RESOURCE SHARING

Reagent and resource requests should be addressed and will be fulfilled by the Lead Contact, Luis Barreiro (luis.barreiro@umontreal.ca).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Sample Collection

Buffy coats from 175 healthy donors were obtained from the Indiana Blood Center (Indianapolis, IN, USA). A signed written consent was obtained from each participant and the project was approved by the ethics committee at the CHU Sainte-Justine (protocol #4022). All individuals recruited in this study were males, self-identified as African-American (AA) (n = 76) or European-American (EA) (n = 99) between the age of 18 and 55 years old. The average age across AA and EU samples was similar (34.2 years (AA) versus 35 years (EA), t test, $p = 0.7$). We decided to only focus on males to avoid the potentially confounding effects of sex-specific differences in immune responses to infection. Only individuals self-reported as currently healthy and not under medication were included in the study. In addition, each donor's blood was tested for Hepatitis B, Hepatitis C, Human Immunodeficiency Virus (HIV), and West Nile Virus, and only samples negative for all of the tested pathogens were used.

METHOD DETAILS

Isolation of Monocytes and Differentiation of Macrophages

Blood mononuclear cells were isolated by Ficoll-Paque centrifugation. Monocytes were purified from peripheral blood mononuclear cells by positive selection with magnetic CD14 MicroBeads (Miltenyi Biotech) using the autoMACS Pro Separator. The purity of the isolated monocytes was verified using an antibody against CD14 (BD Biosciences) and only samples showing > 90% purity were used to differentiate into macrophages. Monocytes were then cultured for 7 days in RPMI-1640 (Fisher) supplemented with 10% heat-inactivated FBS (FBS premium, US origin, Wisent), L-glutamine (Fisher) and M-CSF (20ng/mL; R&D systems). Cell cultures were fed every 2 days with complete medium supplemented with the cytokines previously mentioned. Before infection, we checked the differentiation/activation status of the monocyte-derived macrophages by flow cytometry, using antibodies against CD1a, CD14, CD83, and HLA-DR (BD Biosciences). Only samples presenting the expected phenotype for non-activated macrophages (CD1a+, CD14+, CD83, and HLA-DRlow) were used in downstream experiments.

Bacterial Preparation and Infection of Macrophages

We infected macrophages with two bacteria, *Salmonella typhimurium* and *Listeria monocytogenes*. The day prior to infection, aliquots of *Salmonella typhimurium* and *Listeria monocytogenes* were thawed and bacteria were grown overnight in Tryptic Soy Broth (TSB) media. Bacterial culture was diluted to mid-log phase prior to infection and supernatants density was checked at OD_{600}.

Monocyte-derived macrophages were infected at a multiplicity of infection (MOI) of 10:1 for *Salmonella typhimurium* and an MOI of 5:1 for *Listeria monocytogenes* for 2h at 37°C. A control group of non-infected macrophages was treated the same way but with only medium without bacteria. After 2 hr in contact with the bacteria, macrophages were washed and cultured for another hour in the presence of $50\mu g/ml$ gentamycin in order to kill all extracellular bacteria present in the medium. The cells were then washed a second time and cultured in complete medium with $3\mu g/ml$ gentamycin for an additional 2h, the time point we refer to in the main text. A control group of non-infected macrophages was treated the same way but with only medium without bacteria. We note that we did not run technical replicates for the infections because we could not derive sufficient macrophages from one individual to perform multiple infections with both bacteria. However, the impact of technical confounds are reduced by our large set of biological replicates (and are probably overall small, given our power to detect so many eQTL and ancestry-associated responses).

Estimation of the Number of Infected Macrophages

To determine bacterial counts in infected cells, monolayers of $2 \cdot 10^6$ infected macrophages in 6-well plates were used. Culture medium was removed and replaced with 1ml of 1% Triton X-100 in distilled water. Serial 10-fold dilutions were made, in duplicates, in Trypticase Soy broth and plated on Trypticase Soy agar plates. Plates were kept at 37°C and counted after 24h. Enumeration of intracellular bacteria was performed at T0, corresponding to the percentage of infected macrophages, and T2 and T24, corresponding to the number of bacteria inside the macrophages 2- and 24 hr post-infection, respectively. Data was collected for T0 for all the samples (to control for variation in the number of infected macrophages among individuals), and for T2 and T24 for a subset of 89 individuals for which enough macrophages were available to perform the experiment.

RNA Extraction, Library Preparation, and Sequencing

Total RNA was extracted from the non-infected and infected macrophages 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 for RNA degradation (RNA integrity number > 8) were kept for further experiments. RNA-sequencing libraries were prepared using the Illumina TruSeq protocol. Once prepared, indexed cDNA libraries were pooled (6 libraries per pool) in equimolar amounts and were sequenced with single-end 100bp reads on an Illumina HiSeq2500. Samples were carefully balanced across flow cells and sequencing lanes. Specifically, we multiplexed infected and non-infected samples from the same individual in the same lane, and balanced the number of African Americans and European Americans in each of the flowcells. Additionally, we multiplexed non-infected and infected macrophages (*Salmonella* and *Listeria*) from one European American and one African American in each lane. Because we had a larger number of European ancestry samples than African ancestry samples, the ideal 50-50 ratio was significantly violated for samples sequenced in two of 16 total flowcells. Yet, these samples account for only 5% of all the RNA-seq libraries sequenced. Sequencing libraries from both infected and non-infected conditions were always prepared in parallel with a balanced amount of samples derived from EA and AA individuals.

ATAC-Seq Library Preparation and Sequencing

ATAC-seq libraries were generated from 100,000 cells, as previously described in (Buenrostro et al., 2013) and sequenced on an Illumina HiSeq 2500 using 100-bp paired-end reads. We found high concordance between the ATAC-seq signals for the two biological replicates sequenced for each of the conditions (Spearman r > 0.80), which allowed us to merge them for downstream footprint analyses.

DNA Extraction and Genome-wide Genotyping

DNA was extracted from each of the blood samples using the PureGene DNA extraction kit (Gentra Systems). Each individual was genotyped for over 4.6 million single nucleotide polymorphisms (SNPs), using the Illumina HumanOmni5Exome BeadChip, which interrogates > 4.3 million whole-genome variants, plus the content of the Illumina exome BeadChip. Genotypes were called in all samples together using Genome Studio v2010. All samples had genotype call rates (CR) above 98%, with the exception of 2 samples that were excluded from further analysis. SNPs with > 5% of missing data or deviating from Hardy–Weinberg equilibrium in at least one of the studied populations (at a $p < 10^{-5}$) were excluded. In total, 4,452,246 SNPs passed our quality-control filters. Since samples were collected anonymously, we systematically tested for relatedness in our samples by estimating the pair-wise genome-wide identity by state (IBS) between all possible pairs of individuals using PLINK (Chang et al., 2015). We found 2 pairs of individuals that appeared to be genetically identical, suggesting that these pairs of sample are from the same individual that donated blood twice during our recruitment process. Therefore, we randomly excluded the data of one individual from each of these pairs. All other samples were unrelated as defined by an estimated proportion of IBS < 0.2. Finally, all samples were confirmed to be males on the basis of the genotype data from the X chromosome. After various quality control checks, we ended up with 171 individuals for which genotype data was available for eQTL analyses.

QUANTIFICATION AND STATISTICAL ANALYSIS

Imputation

Imputation was done using Impute2 (ver. 2.3.0) (Howie et al., 2012), on the pre-filtered genotype data and using as reference panels phased genotype data from phase 3 of the 1000 Genomes project (downloaded from: https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html). Our genotype data was phased (per chromosome) using shapeIT (version 2.r790). Post-imputation, we removed genotype calls with likelihood lower than 0.9. In addition, we removed positions with an information metric lower than 0.5, more than 5% of missing genotype calls or deviating from Hardy–Weinberg equilibrium in at least one of the studied populations (at $P < 10^{-5}$). After all filters, we kept 13,846,937 SNPs.

Estimation of Genome-wide Admixture Levels

Self-reported EA and AA have variable degrees of African and European ancestry. In particular, the genome-wide levels of European genetic ancestry among self-reported AAs average 30% and can attain close to 100% in some individuals (Bryc et al., 2010). Thus, instead of relying on self-reported ancestry labels, we calculated the actual proportion of European and African ancestry for each of the samples included in the study using the unsupervised clustering algorithm ADMIXTURE (Alexander et al., 2009). We included 56 Yoruban samples in our analyses to have a group of African individuals that are arguably not admixed. A total of 86,329 unlinked SNPs (*i.e.*, r^2 between all pairs < 0.1) were used for ancestry assignments, assuming K = 2 ancestral clusters. The estimated ancestry proportions were used to assess differences in immune responses between populations, unless mentioned otherwise.

Estimation of Gene- and Isoform-Level Expressions

Adaptor sequences and low quality score bases (Phred score < 20) were first trimmed using Trim Galore (version 0.2.7). The resulting reads were then mapped to the human genome reference sequence (Ensembl GRCh37 release 75) using STAR (2.4.1d) (Dobin et al., 2013) with a hg19 transcript annotation GTF file downloaded from Ensembl (date: 2014-02-07).

The following parameters were used for STAR index generation (other than default):

- genomeSAindexNbases 2
- genomeChrBinNbits 14
- sjdbOverhang 99

In order to obtain aligned reads in transcriptome coordinates, we used the following options specifically recommended for downstream analysis with RSEM:

- outSAMattributes NH HI
- outFilterMultimapNmax 20
- outFilterMismatchNmax 999
- outFilterMismatchNoverLmax 0.05
- alignIntronMin 20
- alignIntronMax 1000000
- alignSJoverhangMin 8
- alignSJDBoverhangMin 1
- quantMode TranscriptomeSAM

Transcript- and gene-level expression estimates were calculated using RSEM (version 1.2.21) (Li and Dewey, 2011), with default parameters considering a mean and standard deviation of 178bp and 58bp, respectively, for insert sizes across our RNA-seq libraries.

Differences in Expression between Populations and in Response to Infection

Quality Control

A total of 22 RNA-seq libraries (out of 525 in total) were removed from downstream analyses because the genotype calls made on the RNA-seq data did not match those obtained from the genotyping array (n = 12), the non-infected samples were clustering close to infected samples in a principal component analysis (n = 4) or the *Listeria*-infected samples were clustering together with *Salmonella*-infected samples (n = 6).

We subset our phenotype data by keeping protein-coding genes that were sufficiently expressed: median TPM value above 0.5 in at least one of the three conditions.

Identification of Relevant Technical Confounders

As a preliminary step for the differential expression analysis, we aimed at identifying confounders that amounted to unwanted technical sources of variability in the expression data. To do this in a systematic way, we began by considering the following pool of putatively relevant technical confounders:

- x_1: sequencing flowcell
- x_2: mean GC content estimated per sample (using RseQC (Wang et al., 2012))
- x_3: fraction of uniquely mapped reads
- x_4: RNA concentration post-extraction
- x_5: sequencing lane
- x_6: RNA integrity numbers (RIN)
- x_7: fraction of multiply mapped reads
- x_8: Total number of sequenced reads
- x_9: RNA library concentration used for sequencing
- x_{10}: library insert size (based on Bioanalyzer)

two of which (*i.e.*, sequencing flowcell and lane) are categorical variables, while the rest are continuous variables that were standardized before the analysis (*i.e.*, rescaled to have *mean* = 0 and *sd* = 1). In order to identify the confounding variables, among the above-mentioned list, that explain a significant amount of the variance in the data, we implemented the following iterative procedure:

- *Step 1:* Let $\boldsymbol{M_{ref}} : \boldsymbol{E} \sim \mathbf{1}$ denote the reference model with no covariates, where only an intercept is estimated for the gene expression data $\boldsymbol{E}$. In addition, assume that $\boldsymbol{M_i} : \boldsymbol{E} \sim \boldsymbol{x_i}$ models the gene expression data by only considering the *i*th technical confounder as the covariate, for $\boldsymbol{i} \in \{\mathbf{1}, \ldots, \mathbf{10}\}$. The fraction of variance in the expression data explained by the *i*th technical covariate was then estimated by $\boldsymbol{v_i} = (\boldsymbol{SS_{M_{ref}}} - \boldsymbol{SS_{M_i}})/\boldsymbol{SS_{M_{ref}}}$ for each gene, where $\boldsymbol{SS_{M_{ref}}}$ and $\boldsymbol{SS_{M_i}}$ represent the residual sum of squares in $\boldsymbol{M_{ref}}$ and $\boldsymbol{M_i}$, respectively.
- *Step 2:* For each technical confounder listed above, the following procedure was repeated for $\boldsymbol{N_{iter}} = \mathbf{200}$ iterations per gene: The entries of the original confounder vector $\boldsymbol{x_i}$ were permuted and the permuted vector was denoted by $\tilde{x}_{\mathrm{i}}$. Afterward, the randomized model $\boldsymbol{M_{rand(i)}} : \boldsymbol{E} \sim \boldsymbol{x_i}$ was set up. The expected amount of variance explained by the randomized confounding variable was then estimated by $\boldsymbol{v_{rand(i)}} = (\boldsymbol{SS_{M_{ref}}} - \boldsymbol{SS_{M_{rand(i)}}})/\boldsymbol{SS_{M_{ref}}}$, where $\boldsymbol{SS_{M_{rand(i)}}}$ denotes the residual sum of squares for $\boldsymbol{M_{rand(i)}}$.
- *Step 3:* For each confounder, the distribution of $\boldsymbol{v_i}$ (*i.e.*, the observed or true fraction of variance explained by the confounder) across all genes was compared to the corresponding distribution of randomized values, $\boldsymbol{v_{rand(i)}}$, through the non-parametric Mann–Whitney U test. The shift between these two distributions at a significance level of p = 0.05 is denoted by δ_i for the ith confounder.
- *Step 4:* We compared the δ_i values across the ten confounders and chose the technical confounder with the maximum shift $\delta_{i^*} := \max_i \delta_i$. If $\delta_{i^*} > 0.01$ (*i.e.*, the contribution of this confounder in explaining the variability in the data is least 1% more than that of an arbitrary random variable), then the confounder was selected and added as a covariate to the reference model $\boldsymbol{M_{ref}}$.
- *Step 5:* We repeated Steps 1 to 3, using the updated reference model. After re-evaluating the distribution shifts δ_i in Step 3, we proceeded as follows: (i) Among the set of confounders currently present in $\boldsymbol{M_{ref}}$, the one with the lowest amount of shift was removed from $\boldsymbol{M_{ref}}$, given that the shift was below 0.01. (ii) Among the set of confounders currently absent in $\boldsymbol{M_{ref}}$, the one which satisfied the selection procedure described in Step 4 was added to $\boldsymbol{M_{ref}}$.
- *Step 6:* Step 5 was repeated until we obtained a reference model where, out of the ten studied confounders, only the covariates present in $\boldsymbol{M_{ref}}$ satisfied the condition mentioned in Step 4 (*i.e.*, their contribution in explaining the variability in the data is least 1% more than that of an arbitrary random variable).

It turned out that only five iterations of the above procedure were sufficient, leading to the following reference model:

$$M_{ref} : E \sim x_1 + x_2 + x_3 + x_4 \tag{1}$$

containing four technical confounders that are controlled for in the downstream analysis (see Figure S1C).

Data Pre-processing

To account for differences in read counts at the tails of the distribution, we normalized the samples using the weighted trimmed mean of M-values algorithm (TMM), as implemented in the R package edgeR (Robinson et al., 2010). Afterward, we log-transformed the data using the voom function in the limma package (Ritchie et al., 2015) and removed the flowcell batch effect using the ComBat function in sva Bioconductor package (Leek et al., 2012). We then applied the voomMod function from package cbcbSEQ (https://github.com/kokrah/cbcbSEQ/blob/master/README.md), specifically devised to work on log-transformed data as opposed to voom which works on count data, to recover new sample weights for the batch-corrected data. Following this pre-processing of the data, we fitted the log-transformed expression estimates to linear models (with design details explained in the subsequent paragraphs), using the lmFit function from the limma package (Ritchie et al., 2015). This function uses the sample weights previously estimated, from the overall mean-variance trend by voomMod, to rescale model residuals and improve the quality of the fit. In these models, the three numerical confounders shown in Equation 1 (*i.e.*, x_2, x_3 and x_4; GC means, RNA concentrations, and fractions of uniquely mapped reads, respectively) are introduced as model covariates. Note that the categorical confounder x_1, or flowcell, has already been corrected for using ComBat. Finally, differential expression effects across conditions (DE) and across populations (pop-DE), along with Ethnicity Condition interaction effects resulting in differential response across populations (pop-DR), were estimated using these linear models. In what follows, each model is elaborately explained.

Ancestry-Related Differential Expression

The following nested linear model was used to identify genes for which expression levels correlated with the African-ancestry levels estimated for each of our samples:

$$M_1 : E(i,j) = \begin{cases} \beta_o(i) + \beta_{Af}^{NI}(i) \cdot Af(j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^{NI}(j) + \varepsilon^{NI}(i,j) & \textit{if} \quad \textit{Condition} = NI \\ \beta_o(i) + \beta_L(i) + \beta_c^L(i) \cdot c_L(j) + \beta_{Af}^{L}(i) \cdot Af(j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^{L}(j) + \varepsilon^{L}(i,j) & \textit{if} \quad \textit{Condition} = L \\ \beta_o(i) + \beta_S(i) + \beta_c^S(i) \cdot c_S(j) + \beta_{Af}^{S}(i) \cdot Af(j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^{S}(j) + \varepsilon^{S}(i,j) & \textit{if} \quad \textit{Condition} = S \end{cases} \tag{2}$$

Here, $E(i,j)$ shows *the expression level of gene i for individual j*, $\beta_{Af}^{NI}(i)$, $\beta_{Af}^{L}(i)$, and $\beta_{Af}^{S}(i)$ indicate *the effects of African admixture (Af) on gene i within each condition*, $\beta_L(i)$ and $\beta_S(i)$ represent the intrinsic infection effects of each pathogen, and $\beta_c^L(i)$ and $\beta_c^S(i)$ are the effects of the standardized bacterial counts (denoted by c_L and c_L) registered in the samples immediately after infection with each pathogen. Furthermore, $\{x_k, k=2,3,4\}$ represents the three numerical covariates previously detected as significant technical confounders; i.e., mean GC content per sample, RNA concentration, and fractions of uniquely mapped reads, with β_{x_k} being their corresponding effects on gene expression. Finally, $\varepsilon^C(i,j)$ represents the residuals at condition C (NI, L or S) for the gene-i individual-j pair, and $\beta_o(i)$ is the global intercept accounting for *the expected expression of gene i in a 100% European non-infected sample (i.e., Af = 0)*. Note that for each individual, we assessed only one sample per condition. In other words, no technical replicates were used in the design.

Fitting the model using the Bioconductor's limma pipeline (Ritchie et al., 2015), we extract the estimates $\beta_{Af}^{NI}(i)$, $\beta_{Af}^{L}(i)$, and $\beta_{Af}^{S}(i)$ of the ethnicity effects across all genes, along with their corresponding p values. Each of these estimates represents the ancestry-related differential expression effects within each condition (pop-DE). Afterward, we control for false discovery rates using an approach analogous to that of Storey and Tibshirani (Storey and Tibshirani, 2003), which makes no explicit distributional assumption for the null-model but instead derives it empirically from 200 permutation tests, where African admixture values are permuted across individuals (see section "Estimation of false discovery rates" below for details). Before proceeding to the table below, which includes results and further details on these effects, some notation is introduced. Let $\langle E_C \rangle_{Af=x}$ denote the expected expression value for a gene in condition $C \in \{NI, L, S\}$ for an individual with African admixture $Af=x$, under M_1. According to this definition, $\langle E_C \rangle_{Af=1} - \langle E_C \rangle_{Af=0}$ represents the expected African ancestry effect within condition C. This effect is denoted by pop-DE:C in the table below and by the β_{Af}^{C} coefficient in model M_1.

Within Condition Ancestry-Associated Differences in Gene Expression

Pop-DE Effect	Linear Model Coefficient M_1	No. Genes under 0.05 FDR	Permuted Variable at Null Model	Estimated Fraction of True Negatives π_o
Pop-DE:NI $\langle E_{NI} \rangle_{Af=1} - \langle E_{NI} \rangle_{Af=0}$	β_{Af}^{NI}	1,745	African ancestry across individuals	0.601
Pop-DE:L $\langle E_{L} \rangle_{Af=1} - \langle E_{L} \rangle_{Af=0}$	β_{Af}^{L}	1,336	African ancestry across individuals	0.658
Pop-DE:S $\langle E_{S} \rangle_{Af=1} - \langle E_{S} \rangle_{Af=0}$	β_{Af}^{S}	2,417	African ancestry across individuals	0.528

Infection Effects: Condition-Related Differential Expression

Contrary to the case of pop-DE analysis, expression levels of samples corresponding to the same individuals are compared in order to test for global infection effects (condition-related DE). To this end, a paired design is used, in which individuals are introduced as additional covariates:

$$M_2 : E(i,j) = \begin{cases} \beta_o(i,j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^{NI}(j) + \varepsilon^{NI}(i,j) & \textit{if} \quad \textit{Condition} = NI \\ \beta_o(i,j) + \beta_L(i) + \beta_c^L(i) \cdot c_L(j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^{L}(j) + \varepsilon^{L}(i,j) & \textit{if} \quad \textit{Condition} = L \\ \beta_o(i,j) + \beta_S(i) + \beta_c^S(i) \cdot c_S(j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^{S}(j) + \varepsilon^{S}(i,j) & \textit{if} \quad \textit{Condition} = S \end{cases} \tag{3}$$

Specifically, $\beta_o(i,j)$ represents the intercept corresponding to gene i and individual j; *i.e.*, the model's expectation for the expression level of gene i at the non-infected sample of individual j. Analyzing model M_2 results in the global (condition-related or ethnicity-independent) estimates of *Salmonella* and *Listeria* infection effects, β_L and β_S, approximated using the within-individual variation in gene expression across conditions.

Similar to the previous model, M_2 is fit using limma; however, the 200 permutation tests implemented here to estimate FDRs are based on random reshuffling of condition labels within each individual (see the table below); moreover, considering the large effect of infection on gene expression, FDRs are obtained from Benjamini-Hochberg's more conservative approach in order to avoid false positives. In the table below, $\langle E_C - E_{NI} \rangle$ shows the expected response upon infection with pathogen C ∈ {L, S} (or equivalently, *C* infection effect), which is denoted by DE:C in the table and by the β_C coefficient in model M**₂**.

Condition-Related Differential Expression Effects

Condition DE Effect	Linear Model Coefficient M_2	No. Genes under 0.05 FDR (BH)	Permuted Variable at Null Model
DE:L $\langle E_L - E_{NI} \rangle$	β_L	10,663	conditions within individual
DE:S $\langle E_S - E_{NI} \rangle$	β_S	10,751	condition within individual

Infection-Ethnicity Interactions: Ancestry-Associated Differential Response to Infection (pop-DR genes)

After obtaining global infection effects, we explored for genes whose response to infection significantly depend on ethnic ancestry. Specifically, we fit the following linear model:

$$M_3 : E(i,j) = \begin{cases} \beta_o(i,j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^{NI}(j) + \varepsilon^{NI}(i,j) & \text{if} \quad Condition = NI \\ \beta_o(i,j) + \beta_L(i) + \beta_c^L(i) \cdot c_L(j) + \beta_{Af}^L(i) \cdot Af(j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^L(j) + \varepsilon^L(i,j) & \text{if} \quad Condition = L \\ \beta_o(i,j) + \beta_S(i) + \beta_c^S(i) \cdot c_S(j) + \beta_{Af}^S(i) \cdot Af(j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^S(j) + \varepsilon^S(i,j) & \text{if} \quad Condition = S \end{cases} \tag{4}$$

which is quite similar to M_2 with the difference that the infection effect of, say, *Listeria* is no longer built in an ethnicity-independent fashion as in model M_2 (*i.e.*, $\langle E_L - E_{NI} \rangle_{M_2} = \beta_L$), since it is in fact dependent on ethnic ancestry as follows: $\langle E_L - E_{NI} \rangle_{M_3} = \beta_L + \beta_{Af}^L(i) \cdot Af$. In this framework, β_{Af}^L and β_{Af}^S denote ethnicity-infection interactions, which represent variations in response to infection observed across ethnic groups (pop-DR). Similar to previous models, 200 permutations were implemented here to estimate FDRs (see the table below for details). According to the notation introduced for models M_1 and M_2, $\langle E_C - E_{NI} \rangle_{Af=x}$ denotes the expected response upon infection with pathogen $C \in \{L, S\}$ for an individual of African admixture $Af = x$. It follows that $\langle E_C - E_{NI} \rangle_{Af=1} - \langle E_C - E_{NI} \rangle_{Af=0}$ represents the infection-ethnicity interaction induced by pathogen *C* (or, *C* infection-ethnicity interaction). This interaction term is denoted by pop-DR:C for pathogen *C* in the table below and by the β_{Af}^C coefficient in model M_3.

Ancestry-Associated Differential Response to Infection

Interaction Effect pop-DR	Linear Model Coefficient M_3	No. Genes under 0.05 FDR (BH)	Permuted Variable at Null Model	Estimated Fraction of True Negatives π_o
Pop-DR:L $\langle E_L - E_{NI} \rangle_{Af=1} - \langle E_L - E_{NI} \rangle_{Af=0}$	β_L	206	African admixture across individuals	0.683
Pop-DR:S $\langle E_S - E_{NI} \rangle_{Af=1} - \langle E_S - E_{NI} \rangle_{Af=0}$	β_S	1,005	African admixture across individuals	0.631

Considering that, and taking *Listeria* as an example, from M**₃** we can build the mentioned expected response upon infection for African-Americans: $\langle E_L - E_{NI} \rangle_{Af=1} = \beta_L + \beta_{Af}^L$, and compare it against the corresponding effect in Europeans: $\langle E_L - E_{NI} \rangle_{Af=0} = \beta_L$, as it is done in Figure 1F, after extracting absolute values.

Applying models M_2 and M_3, instead of M_1, allows us to obtain estimates that are solely based on the within-individual variability. The upside to considering only within-individual variability is that despite the many degrees of freedom consumed by the individual-specific offsets $\beta_o(i,j)$, it augments the statistical power for detecting both global infection effects and ethnicity-infection interaction effects.

False Discovery Rate Estimations

Throughout the paper, (unless stated otherwise), FDRs were calculated separately for each dataset, following a procedure analogous to that proposed by Storey and Tibshirani (Storey and Tibshirani, 2003), which can be described as the following two-component model:

$$F(p) = \pi_o F_o(p) + (1 - \pi_o)F_A(p) \tag{5}$$

where $F_o(p)$ represents the cumulative density of p values for tests truly fulfilling the null hypothesis (*i.e.*, true negatives) and $F_A(p)$ is the equivalent cumulative distribution for tests truly verifying the alternative hypothesis (*i.e.*, true positives). In addition, π_o refers to the fraction of true negatives of the experiment. If the null cumulative distribution $F_o(p)$ is approximately linear (or equivalently, the p values are uniformly distributed under the null hypothesis), the above-mentioned model reduces to Storey and Tibshirani's model, corresponding to the case with $F_o(p) = p$. However, when null distributions deviate from uniform (for example, when the most strongly associated variant is assigned as a single eQTL for a gene), comparisons to empirical, permutation-based null distributions are more appropriate. Indeed, this approach, which requires a minor modification to the method in Storey and Tibshirani, is also appealing because it avoids any assumptions about uniformity. We thus elected to use it here, despite the fact that our empirical nulls are consistently uniform or close to uniform.

Here, we use the empirical cumulative distribution functions (ECDFs) $\widetilde{F_o}(p)$ and $\tilde{F}(p)$ as estimates of F_o(p) and F(p), respectively. To be more specific, $\tilde{F}(p)$ is the ECDF of the actual p values of any effect of interest (either pop-DE, pop-DR or response to infection, for example), whereas $\widetilde{F_o}(p)$ denotes the ECDF obtained from a suitable permutation test performed on that effect.

From Equation 5, the fraction of true negatives under a give p value can be derived as $\pi_o F_o(p)/F(p)$; or $\pi_o \tilde{F}_o(p)/\tilde{F}(p)$, once we accept the ECDF as pertinent estimators of the underlying distributions. Correcting that fraction to ensure monotonicity yields the definition of the tail-area-based false discovery rate FDR(p):

$$FDR(p) = \min_{p' \geq p}\left(\frac{\pi_o \tilde{F}_o(p')}{\tilde{F}(p')}\right) \tag{6}$$

In order to compute $FDR(p)$, we must first estimate π_o. As proposed in (Storey and Tibshirani, 2003), this is achieved by

$$\widehat{\pi}_o(p) = \frac{1 - F(p)}{1 - F_o(p)} \approx \frac{1 - \tilde{F}(p)}{1 - \tilde{F}_o(p)} \tag{7}$$

yielding a biased estimator of π_o, where the amount of bias declines as p approaches the maximum p value registered in the experiment, p_{max}. Therefore, to obtain a better estimation of π_o, the estimator $\widehat{\pi}_o(p)$ is fitted to a suitable smooth function f(p) - typically a decreasing cubic spline- evaluated at $p_{max}: \ \pi_o \simeq f(p_{max})$.

Differential Isoform Usage between Populations

Prior to performing the DIU analysis, we removed the lowly expressed isoforms and only kept those with median TPM value (strictly) above zero in at least one of the three experimental conditions, using isoform-level TPM values estimated by RSEM. In the next step of data pre-processing, the ComBat function in the sva Bioconductor package (Leek et al., 2012) was applied to the log-transformed isoform-level TPM data to remove the flowcell batch effect. Then, the following linear model was designed using limma (Ritchie et al., 2015):

$$M_4 : log_2\left(TPM + 10^{-5}\right)(i,j) = \begin{cases} \beta_o(i) + \beta_{Af}^{NI}(i) \cdot Af(j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^{NI}(j) + \varepsilon^{NI}(i,j) & if \quad Condition = NI \\ \beta_o(i) + \beta_L(i) + \beta_c^L(i) \cdot c_L(j) + \beta_{Af}^{L}(i) \cdot Af(j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^{L}(j) + \varepsilon^{L}(i,j) & if \quad Condition = L \\ \beta_o(i) + \beta_S(i) + \beta_c^S(i) \cdot c_S(j) + \beta_{Af}^{S}(i) \cdot Af(j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^{S}(j) + \varepsilon^{S}(i,j) & if \quad Condition = S \end{cases} \tag{8}$$

where 0.00001 (*i.e.*, $0.001 \times \min_{TPM > 0} TPM$) was added to all the TPM values to avoid instances of $\log(0)$. Note that this model has the same design as model M_1 for pop-DE effects. After obtaining covariate estimates for this model, the effect of the technical confounders (*i.e.*, mean GC content, fraction of uniquely mapped reads, and RNA concentration) were regressed out from the log-transformed isoform-level TPM values. Finally, the obtained values were transformed back, from log_2-scale, to TPM-scale. These values were then used in the downstream DIU analysis. To detect differences in isoform usage between African-Americans and European-Americans, we applied a multivariate generalization of the Welch's t test to the set of 10223 genes, out of the total number of genes, with at least two an-notated isoforms (which remained after the elimination of lowly expressed isoforms in the pre-processing step). To implement the method, we started by calculating the proportional abundance of the different isoforms for each tested gene using the isoform-level TPM values estimated by RSEM. The proportional isoform abundance (or relative isoform usage) for a target gene g with D isoforms is denoted by a vector of size D, where its elements sum to one and the *ith* element denotes the proportional abundance of isoform *i*. Next, we tested whether the means of the two multivariate distributions, associated with African-American and European-American populations, were equal. Specifically, suppose that group *i* consists of n_i samples, and let $\pi_{ij} = (\pi_{ij1}, \ldots, \pi_{ijD})$ be the vector of proportional isoform abundance for sample j of group I, with $i \in \{1, 2\}$ and $j \in \{1, \ldots, n_i\}$. Clearly, $\sum_{d=1}^{D} \pi_{ijd} = 1$, with π_{ijd} denoting the relative isoform usage associated with isoform d. Following this notation, the relative isoform usage data falls into the category of compositional data (Aitchison, 1982), where components (or vector elements) are proportions of total isoform abundance that sum to one. Mathematically, the state space of such compositional data is defined as an open simplex (*i.e.*, a generalization of the notion of a two-dimensional triangle to higher dimensions) as follows (Egozcue et al., 2003):

$$S^D = \left\{ (x_1, \ldots, x_D) \mid x_d > 0 \ \forall \ \ d \in \{1, \ldots, D\}, \ \sum_{d=1}^{D} x_d = 1 \right\} \tag{9}$$

The fact that the proportions have a fixed sum implies that (1) there is dependency between relative isoform usage values within each sample and (2) S^D is not a vector space. This results in specific numerical characteristics, which interfere substantially with the approaches taken in the statistical analysis of compositional data. For one thing, the familiar Euclidean geometry cannot be applied when dealing with compositional data; specifically, although the distance between two real vectors can be easily computed with the standard Euclidean metric, it is not the proper metric to use for compositional data. To illustrate, consider the following pairs of compositions: {(0.25, 0.05, 0.7), (0.25, 0.1, 0.65)} and {(0.25, 0.5, 0.25), (0.25, 0.55, 0.2)}. The Euclidean distance between the compositions in the first pair equals that of the second pair, as the element-wise difference between the compositions is (0, 0.05, 0.05) for both pairs. However, the second component has doubled in the first pair, while it has only increased by ten percent in the second pair. The fold changes associated with the third components are more comparable between the pairs (0.9 and 0.8 for the first and the second pairs, respectively). In other words, while the Euclidean distances between compositions of both pairs are equal, fold changes imply that the actual distance is larger for the first pair. Therefore, the relative variation of components, rather than their absolute differences, provide the basis to the statistical analysis of compositional data. Lending a linear vector space structure to the open simplex, the Aitchison geometry (Aitchison, 1982) provides us with a way to work with compositional data that is analogous to the real space. Any statistical analysis on compositional data can be performed using this vector space structure; however, it is easier to use alternative methods, which transform compositional data to the familiar Euclidean space. Egozcue et al. (Egozcue et al., 2003) proposes the isometric log-ratio transformation (*ilr*), which is obtained with orthogonal coordinates. Using a set of D-1 orthonormal vectors as the basis for S^D, the *ilr* transformation maps the log of a given composition to a vector of size D-1 in the Euclidean space. The advantage of applying this distance-preserving mapping is that it allows for the familiar Euclidean geometry to be applied to the obtained vectors in $\mathbb{R}^{D-1}$, since the vector elements are no longer dependent on one another after the transformation.

As one can come up with more than one orthonormal basis for the open simplex, the *ilr* transformation is not unique. In this paper, we employed a specific one defined by (Egozcue et al., 2003). For any $\boldsymbol{x} = (x_1, \ldots, x_D) \in S^D$,

$$ilr(x) : = \log(x) \cdot U \tag{10}$$

where $U = [U_1, \ldots, U_{D-1}]$ is the $D \times (D-1)$ orthonormal basis, with $U_i \in \mathbb{R}^D$ denoting its *i*th column:

$$U_{ji} = \begin{cases} \frac{1}{\mathrm{i}}\sqrt{\frac{i}{1+i}}, & if \qquad j \le i \\ -\sqrt{\frac{i}{1+i}}, & if \qquad j = i+1 \\ 0, & \qquad otherwise \end{cases} \tag{11}$$

for $j \in \{1, \ldots, D\}$. Initially, and before applying the *ilr*-transformation on the relative isoform usage data, the statistical hypothesis test for differential isoform usage between African-American and European-American groups within each condition was set up as:

$$\begin{aligned} H_o &: \mu_{\pi_1} = \mu_{\pi_2} \\ H_1 &: \mu_{\pi_1} \neq \mu_{\pi_2} \end{aligned} \tag{12}$$

where μ_{π_i} is the mean relative isoform usage for group *i*. To prepare the data for the *ilr* transformation and statistical analysis, two preliminary steps were undertaken as follows. To make sure that the statistical test yielded biologically meaningful results, if the average relative abundance of an isoform across samples was less than 0.05 in both African-American and European-American groups, that isoform was eliminated from statistical testing analysis. Any relative isoform usage value that remained after the isoform-elimination step and was estimated as zero was replaced by a small strictly positive value of 0.0005 to make sure that all samples belong to the open simplex S^D. Note that if, for a specific isoform and a specific sample, the relative abundance is below 0.05 and strictly greater than 0, and if that isoform is not removed in the isoform-elimination step, then the relative abundance value is retained.

After performing the *ilr* transformation on each sample, a multivariate normal distribution on $\mathbb{R}^{D-1}$ was assumed for the *ilr*-transformed relative isoform usage data:

$$ilr(\pi_{i1}), \ldots, ilr(\pi_{in_i}) \sim \mathcal{N}_{D-1}\left(\mu_{ilr(\pi_i)}, \Sigma_i\right) \quad for \quad i=1,2, \tag{13}$$

where Σ_i is the covariance matrix for group *i*, with $\Sigma_1 \neq \Sigma_2$. Consequently, differential isoform usage boils down to testing the equality of means of two multivariate normal populations, with distinct covariance matrices. This is mathematically shown by:

$$\begin{aligned} H_o &: \mu_{ilr(\pi_1)} = \mu_{ilr(\pi_2)} \\ H_1 &: \mu_{ilr(\pi_1)} \neq \mu_{ilr(\pi_2)} \end{aligned} \tag{14}$$

Where $\mu_{ilr(\pi_i)}$ is the mean of *ilr*-transformed relative isoform usage vectors of group *i*. This problem is referred to as the multivariate Behrens-Fisher problem, and different approaches have been proposed to tackle the multi-dimensional case. In this paper, we adopted the method proposed by (Krishnamoorthy and Yu, 2004), which reduces to the well-known Welch's t test for one-dimensional data (or equivalently, when D = 2). This test, referred to as T_{KY} herein, cannot be employed when $D-1 \geq \min\{n_1, n_2\}$ (a case that results in either of the estimated covariance matrices to be singular and non-invertible). This is not a concern in our analysis, since we have a large number of samples per group. The result of differential isoform usage test is reported in Table S2D, where estimation of isoform expression values was done using the RSEM software package. Out of the 10223 genes tested, 62, 39, and 48 genes showed statistically significant DIU between African-American and European-American populations, in the non-infected, *Listeria*-infected, and *Salmonella*-infected samples, respectively (FDR <0.05).

To verify the robustness of our results, we re-conducted our DIU analysis with the approach adopted by (Lappalainen et al., 2013). Specifically, we used the Mann-Whitney U test to compare the distributions of relative abundances, for each isoform, between African-American and European-American populations so as to detect transcripts with significantly different ratios between populations. Afterward, the Benjamini–Hochberg FDR method was applied to adjust the p-values obtained from these individual comparisons (*i.e.*, between-population comparisons per isoform). Subsequently, a gene was labeled DIU provided that at least one of its isoforms showed significant evidence of differential usage between African-Americans and European-Americans. In particular, using this approach 93, 70, and 123 genes were detected with significant DIU (FDR < 0.05) between African-American and European-American populations, in the non-infected, *Listeria*-infected, and *Salmonella*-infected samples, respectively. Figure S2B compares the number of DIU genes detected using the multivariate Welch's t test (our original approach) with that obtained by the rank sum test at different FDR thresholds.

Enrichment of GWAS-Associated Genes among pop-DE Genes

To identify enrichment of pop-DE genes among genes that were previously found to be associated with complex human disease and traits, we used data from the GWAS catalog (Hindorff et al., 2009). Since each GWAS has a different distribution of P-values and significance cutoffs, we chose to use a set of $-log_{10}(p)$ cutoffs in the range of 8-60 (plotted along the x axis in Figure 2B). For a given disease, we identified the overlap between the genes significantly associated with the disease at each cutoff and pop-DE genes, and calculated a fold enrichment (plotted along the y axis in Figure 2B), defined as the ratio of observed/expected overlap between the two gene sets. We used a Fisher Exact Test to calculate a P-value for each cutoff, and corrected these P-values for multiple tests using the FDR approach within each disease.

Genotype-Phenotype Association Analysis

eQTL, asQTL and reQTL mappings were performed against a set of 11,927 protein coding genes. We examined associations between SNP genotypes and the phenotype of interest using a linear regression model, in which phenotype was regressed against genotype. In particular, expression levels were considered as the phenotype when searching for eQTL and asQTL, while to identify reQTL, fold changes in response to infection were treated as the quantitative trait to be mapped. In all cases, we assumed that alleles affected the phenotype in an additive manner. For the eQTL and asQTL analyses, we mapped *Salmonella*-infected, *Listeria*-infected, and non-infected macrophages, separately. All regressions were performed using the R package Matrix eQTL (Shabalin 2012). To avoid low power caused by rare variants, only SNPs with a minor allele frequency of 5% across all individuals were tested. Local associations (*i.e.*, putative *cis* QTL) were tested against all SNPs located within the gene body or 100Kb upstream and downstream of the transcript start site (TSS) and transcript end site (TES) of each tested gene. We recorded the minimum P-value (*i.e.*, the strongest association) observed for each gene, which we used as statistical evidence for the presence of at least one eQTL for that gene. *Trans*-eQTL were defined as SNPs located > 500kb of the gene they regulate and could be on the same or different chromosomes. To estimate an FDR, we permuted the phenotypes (expression levels, fold changes or percent of isoform usage) ten times, re-performed the linear regressions, and recorded the minimum P-values for the gene for each permutation. These sets of minimum P-values were used as our empirical null distribution and FDRs were calculated using the method described in the section "Estimation of FDRs."

Consistent with previous reports (Barreiro et al., 2012), we found that we could increase the power to detect *cis*-eQTL by accounting for unmeasured-surrogate—confounders. To identify such confounders, we first performed principal component analysis (PCA) on a correlation matrix based on genes expressions, for non-infected, *Salmonella*- or *Listeria*-infected samples. Subsequently, we regressed out up to 15 principal components before performing the association analysis for each gene. This specific number of

PCs was chosen since it empirically led to the identification of the largest number of eQTL in each condition. The exact number of PCs regressed in each of the analyses can be found in the table below. Note that for the *trans* analysis we did not regress any PCs to avoid inadvertently removing the effect of true *trans* signals.

Principal Components Regressed

Analysis	Condition	Regressed PCs	No. Genes under 0.01 FDR	Estimated Fraction of True Negatives π_o
Cis eQTL	non-infected	1 to 3	875	0.56
	Listeria	1 to 7	1,087	0.47
	Salmonella	1 to 5	983	0.47
Cis reQTL	fold change *Listeria*	1 to 10	244	0.69
	fold change *Salmonella*	1 to 7	503	0.66
Cis asQTL	non-infected	1 to 3	886	0.67
	Listeria	1 to 3	746	0.66
	Salmonella	1 to 3	615	0.65

Importantly, although the PC corrections clearly increase power to detect eQTL, they do not affect the underlying structure of the expression data. Indeed, over 80% of the eQTL observed before any PC correction are also observed after PC correction at the same FDR cutoff. A similar approach was used for asQTL and reQTL mapping, with the only difference being that for those analyses the PCA were performed in a matrix of isoform proportional abundance or fold-change responses, respectively.

Mapping was performed combining AA and EA samples to increase power. To avoid spurious outcomes resulting from population structure, the first five eigenvectors obtained from a PCA on the genotype data were included in the linear model as well. For each library, we also took into account the potential biases and significant technical confounders identified before the DE analyses; *i.e.*, bacteria counts used when infecting the macrophages (c), sequencing flowcell (x_1), mean GC content estimated per sample (x_2), proportion of uniquely mapped reads (x_3), and RNA concentration (x_4). The covariate subscripts or superscripts corresponding to the experimental condition, in which they were measured, are dropped in the following models for simplicity:

- eQTL models, non-infected condition:

$$Gene\ expression \sim Genotype + \sum_{i=1}^{4} x_i + EV1 + \ldots + EV5 \tag{15}$$

- eQTL models, *Listeria* and *Salmonella* conditions:

$$Gene\ expression \sim Genotype + c + \sum_{i=1}^{4} x_i + EV1 + \ldots + EV5 \tag{16}$$

- reQTL models, response to *Listeria* and *Salmonella* infections:

$$Gene\ fold\ change \sim Genotype + c + \sum_{i=1}^{4} x_i + EV1 + \ldots + EV5 \tag{17}$$

- asQTL models, non-infected condition:

$$Transcript\ proportion \sim Genotype + \sum_{i=1}^{4} x_i + EV1 + \ldots + EV5 \tag{18}$$

- asQTL models, *Listeria* and *Salmonella* conditions:

$$Transcript\ proportion \sim Genotype + c + \sum_{i=1}^{4} x_i + EV1 + \ldots + EV5 \tag{19}$$

Identification of Condition-Specific eQTL and asQTL

We classified condition-specific *cis*-QTL using a conservative criterion aimed at minimizing the risk that true eQTL in both resting and infected macrophages are only identified in one condition because of incomplete power. Specifically, we defined condition-specific QTL when we found strong evidence (FDR < 0.01) of a *cis*-QTL in one condition and no statistical evidence, using a relaxed FDR threshold (0.3), supporting a *cis*-QTL for the same gene in the other conditions.

Multiple Testing Correction

To estimate a FDR, we permuted the phenotypes ten times and used the distribution of the acquired minimum p value per gene to calculate the FDR associated with the p value obtained from the real data, as described above.

Allele-Specific Expression Detection

The sequenced samples were preprocessed using WASP (van de Geijn et al., 2015) program in order to correct for mapping biases toward the reference sequence. To this end, we removed all the monomorphic sites and hence only the positions showing polymorphism in at least one of the 171 samples were included in the analysis for correction. The resulting fastq files from WASP were mapped the same way the original alignment files were obtained (*i.e.*, using the STAR pipeline). Allele counts per sample for positions that overlap with the Omni5Exome-4v1-1 genotyping array were obtained using SAMtools mpileup (Li et al., 2009) v0.1.19-44428cd, with minimum base quality of 13.

Genotype calls obtained from these steps were then used as the input files for ASE identification with the QuASAR software (Harvey et al., 2015), which can jointly genotype and detect allelic imbalances for each SNP. Starting with three samples from each individual, corresponding to the two bacterial infections and the non-infected control, QuASAR can simultaneously identify heterozygous SNPs and ASE by taking into account base-calling errors and over-dispersion in the ASE ratio. The prior genotype probabilities in QuASAR are obtained from the 1000 Genomes Project minor allele frequencies assuming Hardy–Weinberg equilibrium; however, as we had the genotype information available, we manually input the prior genotype probabilities. Specifically, the prior genotype probabilities in QuASAR are indicated in a matrix of three columns, where the columns denote homozygous reference, heterozygous, and homozygous alternate probabilities, respectively, and each row corresponds to an exonic location. As an illustration, to input our genotype information for a heterozygous exonic location, we set the corresponding row equal to $(\gamma, 1-2\gamma, \gamma)$ where $\gamma = 0.001$ accounts for genotyping error. This is done through changing the gmat argument of the *fitAseNullMulti()* function. Manually setting up the genotype probabilities, as mentioned above, assures that the prior genotype information will not change drastically as QuASAR iterates through the EM algorithm steps; moreover, it enables us to estimate the base-calling error rate. In the subsequent step of inferring ASE, we set the min.cov argument of the *aseInference()* function equal to 10 to only assess the sites represented in at least 10 reads across all the samples.

QuASAR outputs the allelic imbalance estimate for each exonic location as $\log(p/1-p)$, where p denotes the proportion of the number of reference reads over the total number of reads, with no allelic imbalance (*i.e.*, $p = 0.5$) resulting in an effect size estimate of zero. After obtaining estimates of allelic imbalance and the corresponding standard errors from QuASAR for each (heterozygous) exonic SNP in each sample, we used the CRG Alignability tracks in the framework of the GEM (GEnome Multitool) project to only keep the exonic SNPs with a mapability score of S = 1; *i.e.*, with only one match in the genome. We then performed meta-analysis on this output to aggregate the effect sizes across samples for each exonic SNP. Specifically, we only retained the exonic SNPs with a frequency of at least three samples and adopted the inverse-variance weighting method for each of these exonic locations to combine the effect sizes across samples, where each effect size was weighted by its inverse variance. A Z-score is then calculated for each weighted mean effect size to allow for a Z-test of significant deviation from zero, to test the null hypothesis of no allelic imbalance at each exonic SNP.

ATAC-Seq Data Processing and Footprinting Analysis

ATAC-seq paired-end reads were mapped to the human reference genome (GRCh37/hg19) using Bowtie 2 (Langmead and Salzberg, 2012) with the following parameters: -N 1 -X 2000–no-mixed–no-discordant–no-unal. Only reads that had a paired and unique alignment were retained. PCR duplicates were removed using Picard's MarkDuplicates tool.

To detect TF binding footprints in the ATAC-seq data we used the program Centipede (Pique-Regi et al., 2011). We ran Centipede separately for each of the three conditions. We started by defining the set of transcription factors that were active (*i.e.*, had motif instances with footprints) before and after infection with *Listeria* and *Salmonella* using a reduced set of high-confidence motif instances for each TF. Using these reduced set of motifs, we calculated a Z-score corresponding to the PWM effect in the prior probability in Centipede's logistic model. The Z-score corresponds to the parameter in:

$$log\left(\pi_l/1-\pi_l\right) = \alpha + \beta \bullet PMW_{score_l} \tag{20}$$

where π_l represents the prior probability of binding in Centipede model in motif location l. We considered a TF binding site as active if the estimate was supported at Bonferroni-corrected $P < 10^{-5}$. In total, 369, 420 and, 422 motifs were identified as active in non-infected, *Listeria*-infected and *Salmonella*-infected macrophages, respectively. We then scanned the entire genome for motif instances matching the original PWM for these active motifs, separately for each of the three conditions.

Footprints were grouped into clusters using sequence similarity. The positional overlap of predicted bound regions of active motifs was first determined using bedtools multiinter. Using the overlap scores, footprints were then divided into clusters using R function hclust with a distance cutoff of 0.9. Well-supported footprints (posterior Pr > 0.9) were used for the enrichment analysis.

Enrichment of TF Binding Sites among eQTL and reQTL

To estimate the enrichment level of particular transcription factor binding locations among eQTL and reQTL, we used the method described in (Wen et al., 2016) and implemented in TORUS (https://github.com/xqwen/dap/tree/master/torus_src).

This method uses a hierarchical model that aggregates eQTL signals across all genetic variants to model the characteristics shared among those most likely to be causal. This is an iterative process starting from eQTL summary statistics calculated with matrix-eQTL using a comprehensive set of imputed genotypes. Using the deterministic approximation of posteriors (DAP) approach, we then learn a prior for each genetic variant using a logistic that can use different types annotations that are informative for determining SNPs that are more likely to disrupt transcription. In our case, we seek to determine if SNPs in binding sites for certain TFs are more likely to be eQTL or reQTL, and therefore determine the likely molecular mechanisms underlying our QTL signals.

The putative binding sites (*i.e.*, ATAC-seq footprints) were determined using ATAC-seq, as described in the section above. To analyze eQTL in non-infected, *Salmonella*-infected, and *Listeria*-infected macrophages, we used the footprints derived in each condition. To analyze reQTLs to *Salmonella* and *Listeria* infection, we used the footprints collected in *Salmonella*-infected and *Listeria*-infected macrophages, respectively. To avoid spurious enrichments resulting from the fact that several TF binding sites are non-randomly distributed with respect to the TSS, we used a background model that captures the effects of distance to the TSS.

Footprints corresponding to different types of transcription factors were analyzed separately. For each annotation, we also calculated a 95% confidence interval. The "enrichment" parameter represents the log-odds that genetic variants in a particular annotation are more likely to harbor causal SNPs for an eQTL compared to a baseline background model that takes into account distance to TSS. In our application, higher enrichment for genetic variants in a specific transcription factor binding sites provide evidence for a likely causal mechanism underlying many of the measured eQTLs and reQTLs.

Genetic Control of Ancestry Effects on Gene Expression

To determine the extent to which a set of genetic variants control the signal associated with ethnic admixture in gene expression variation, a comparison should be made between the models M_1, M_2, and M_3 -previously introduced to produce estimates of pop-DE, condition-DE and pop-DR effects, respectively, and their corresponding extensions, which take the effect of SNPs into consideration.

Control of Ancestry-Related Differential Expression (pop-DE genes) by Individual SNPs

Extending M_1 to account for *cis*-genetic variation effects as mentioned above, we obtain the following model, $M_1^{G_{cis}}$, that has the effects of the best-associated *cis*-SNPs of each gene (regardless of significance level) in each condition as additional covariates:

$$M_1^{G_{cis}} : E(i,j) = \begin{cases} \beta_o(i) + \beta_{Af}^{NI}(i) \cdot Af(j) + \beta_G^{NI}(i) \cdot G^{NI}(i,j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^{NI}(j) + \varepsilon^{NI}(i,j) & \text{if} \quad Condition = NI \\ \beta_o(i) + \beta_L(i) + \beta_c^L(i) \cdot c_L(j) + \beta_{Af}^L(i) \cdot Af(j) + \beta_G^L(i) \cdot G^L(i,j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^L(j) + \varepsilon^L(i,j) & \text{if} \quad Condition = L \\ \beta_o(i) + \beta_S(i) + \beta_c^S(i) \cdot c_S(j) + \beta_{Af}^S(i) \cdot Af(j) + \beta_G^S(i) \cdot G^S(i,j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^S(j) + \varepsilon^S(i,j) & \text{if} \quad Condition = S \end{cases} \tag{21}$$

In this model, $G^{NI}(i,j)$, $G^L(i,j)$, and $G^S(i,j)$ represent the genotypes of the *cis*-SNPs with the highest association to the expression of gene *i* in individual *j* and take values in the set $\{0,1,2\}$ accounting for the copies of the less abundant allele. It is worth mentioning that $G^{NI}(i,j)$, $G^L(i,j)$, and $G^S(i,j)$ generally differ across genes and conditions. Furthermore, these SNPs are not necessarily the true eQTLs, as being the top association for a given gene-condition pair does not automatically imply that the SNP satisfies the FDR-threshold criteria of statistical significance to be an eQTL. The reason why we do not require SNPs to pass any FDR threshold for significance is that smaller effect eQTL may still contribute to population differences, with the most strongly associated variant still being the best candidate eQTL. Moreover, we note that most of these "best variants" actually do have reasonably strong evidence for eQTL association (e.g., 40% of best variants are identified at an FDR < 0.1, and 61% are identified at an FDR < 0.2), even if they do not pass our more stringent primary threshold (FDR < 0.01).

In order to compare the role of ethnic admixture in shaping gene expression levels before and after regulatory variants are introduced to the model, for each gene, let us define its reduced expression vector associated to a given condition *C* as the set of expression values within such condition from which the effects of all model covariates except ethnic admixture have been removed:

$$e_{M_1}^C(j) = \beta_{Af}^C \cdot Af(j) + \varepsilon^C(j) \tag{22}$$

where β_{Af}^C coefficients and $\varepsilon^C(j)$ come from M_1 fit. If we denote the gene mean of the reduced expression values across individuals for condition *C* as $\langle e \rangle_{M_1}^C$, then the total sum of square deviations from $\langle e \rangle_{M_1}^C$ for that gene is $SS_{tot(M_1)}^C = \sum_j (e_{M_1}^C(j) - \langle e \rangle_{M_1}^C)^2$; and can be expressed as the sum of two components: a between-population component, explained by the admixture effect in the regression model: $SS^{C}_{reg(M_1)} = \sum_{j} (\beta^{C}_{Af}.Af(j) - \langle e \rangle^{C}_{M_1})^2$; and a within-population, or unexplained component corresponding to the residuals: $SS^{C}_{res(M_1)} = \sum_{j} \varepsilon^{C}(j)^2$.

From these magnitudes, the P_{ST} statistics is build for each gene as follows:

$$Pst_{M_1}(C) = \frac{SS^{C}_{reg(M_1)}}{SS^{C}_{reg(M_1)} + SS^{C}_{res(M_1)}} \tag{23}$$

which measures the fraction of variance of the reduced expression that is explained by ethnicity. Defined this way, the P_{ST} indexes constitutes a phenotypic analog of the population genetics parameter F_{ST} (Leinonen et al., 2013), when the population's structure is not defined in a binary fashion but according to a continuous trait as the ethnic admixture. From a merely formal point of view, the P_{ST} statistics is just the coefficient of determination $\mathbf{R^2}$ associated to the regression model that defines the reduced expression vectors.

Likewise, we can obtain the reduced expression vectors from $M_1^{G_{cis}}$, which we denote by $e^{C}_{M_1^{G_{cis}}}$; and from these, the respective $Pst_{M_1^{G_{cis}}}(C)$. In order to assess the contribution of genetic information to reduce the fraction of variance that ethnicity explains with respect to the residuals, we can compare $Pst_{M_1}(C)$ against $Pst_{M_1^{G_{cis}}}(C)$ through the following relative variation index:

$$\Delta Pst^{cis}_{M_1}(C) = \frac{Pst_{M_1}(C) - Pst_{M_1^{G_{cis}}}(C)}{Pst_{M_1}(C)} \tag{24}$$

Analogously, we also build a version of M_1 including, instead of *cis*-SNPs, the best *trans*-SNP of each gene as an additional covariate, denoted as $M_1^{G_{trans}}$; from whose comparison to $\mathrm{M}_{\boldsymbol{1}}$ we ultimately derive $\Delta Pst^{trans}_{M_1}(C)$ in the same way.

Control of Ancestry-Associated Differential Response to Infection by Individual SNPs

Similar to what proceeded, the following extension of M_3 is set up to incorporate the genetic variants that affect the gene-expression responses to infection:

$$M_3^{G_{cis}} : E(i,j) = \begin{cases} \beta_o(i,j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^{NI}(j) + \varepsilon^{NI}(i,j) & \text{if} \quad Condition = NI \\ \beta_o(i,j) + \beta_L(i) + \beta_c^L(i) \cdot c_L(j) + \beta_{Af}^L(i) \cdot Af(j) + \beta_G^L(i) \cdot \widehat{G}^{L}(i,j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^{L}(j) + \varepsilon^{L}(i,j) & \text{if} \quad Condition = L \\ \beta_o(i,j) + \beta_S(i) + \beta_c^S(i) \cdot c_S(j) + \beta_{Af}^S(i) \cdot Af(j) + \beta_G^S(i) \cdot \widehat{G}^{S}(i,j) + \sum_{k=2}^{4} \beta_{x_k}(i) \cdot x_k^{S}(j) + \varepsilon^{S}(i,j) & \text{if} \quad Condition = S \end{cases} \tag{25}$$

For the gene-*i* individual-*j* pair, $\widehat{G}^{L}(i,j)$ and $\widehat{G}^{L}(i,j)$ represent the genotypes of the *cis*-SNPs with the highest association to the gene-expression fold changes after infection with *Listeria* and *Salmonella*, respectively. When statistically significant, these top associations constitute response eQTLs, as their effects on gene expression differ across conditions.

In a similar fashion to the comparative analysis of M_1 and $M_1^{G_{cis}}$, so as to compare M_3 and $M_3^{G_{cis}}$, we start by defining the reduced response vectors as the expected fold change after infection with each pathogen, from which all the covariates but ethnicity have been removed using model M_3 estimates:

$$\begin{aligned} fc^{L}_{M_3}(j) &= \beta^{L}_{Af} \cdot Af(j) + \varepsilon^{L}_{fc}(j) \\ fc^{S}_{M_3}(j) &= \beta^{S}_{Af} \cdot Af(j) + \varepsilon^{S}_{fc}(j) \end{aligned} \tag{26}$$

where the fold change residuals are derived from the differences between infected and non-infected samples residuals for each individual j from which a valid sample of each condition was collected: $\varepsilon^{L}_{fc}(j) = (\varepsilon^{L} - \varepsilon^{NI})(j)$ and $\varepsilon^{S}_{fc}(j) = (\varepsilon^{S} - \varepsilon^{NI})(j)$.

From this point, the computation of P_{ST} statistics provides us with a measure of the proportion of variance that is explained by the interaction terms β^{Af}_{L} and β^{Af}_{S} at the reduced fold changes obtained from M_3 estimates. More precisely, for the infected condition C (*Listeria* or *Salmonella*), the total sum of square deviations from the average $\langle fc \rangle^{C}_{M_3}$ reads.

As $SS^{C}_{tot(M_3)} = \sum_{j} (fc^{C}_{M_3}(j) - \langle fc \rangle^{C}_{M_3})^2$; while the regression and residuals components are $SS^{C}_{reg(M_3)} = \sum_{j} (\beta^{C}_{Af}.Af(j) - \langle fc \rangle^{C}_{M_3})^2$ and $SS^{C}_{res(M_3)} = \sum_{j} \varepsilon^{C}(j)^2$, respectively. Therefore, the corresponding P_{ST} index reads as follows:

$$Pst_{M_3}(C) = \frac{SS^{C}_{reg(M_3)}}{SS^{C}_{reg(M_3)} + SS^{C}_{res(M_3)}} \tag{27}$$

Moreover, the *Pst* coefficients obtained from model M_3, can be also obtained from $M_3^{G_{cis}}$ within each infected condition to get the corresponding $Pst_{M_3^{G_{cis}}}(C)$ statistics. Finally, these values are compared to $Pst_{M_3}(C)$ through the following relative variation index:

$$\Delta Pst_{M_3}^{cis}(C) = \frac{Pst_{M_3}(C) - Pst_{M_3^{G_{cis}}}(C)}{Pst_{M_3}(C)} \tag{28}$$

Finally, the equivalent analysis is performed using *trans*-SNPs: from $M_3^{G_{trans}}$, we calculate the corresponding P_{ST} statistics $Pst_{M_3^{G_{trans}}}(C)$, and compare those against $Pst_{M_3}(C)$ through the corresponding relative variation indexes $\Delta Pst_{M_3^{G_{trans}}}(C)$.

Null Model

As a means of validating the significance of the comparisons conducted before and after introducing genetic information in our linear models, we consider a null model in which the top associated SNPs in $M_1^{G_{cis}}$ and $M_3^{G_{cis}}$ (or $M_1^{G_{trans}}$ and $M_3^{G_{trans}}$) are substituted by (*i*) randomly selected SNPs matched for the allele frequency of the lead *cis*-SNP, or *(ii)* lead *cis*-SNPs identified after permuting the genotype data.

Gene Ontology Enrichment Analysis

Using ClueGO cytoscape module (Bindea et al., 2009), we interrogated for enrichments of ontology terms related to Biological processes in different target sets of DE genes with respect to a background composed by all genes analyzed. For this particular, for pop-DR, genes within the target sets were required to present an absolute fold change larger than 0.2 -positive or negative, depending on the direction of the effect- and a false discovery rate lower than 0.2 for the DE effect considered. For infection DE enrichments, characterized by much larger size effects, we conducted one analysis per pathogen, regardless direction of effects, using a more stringent cutoff (absolute fold change larger than 0.5 and FDR < 0.01). Regarding the program's parameters defining the test to consider, we configured them as follow:

- Statistical Test Used = Enrichment (Right-sided hypergeometric test).
- Fusion of related Parent-child terms activated.
- Correction Method Used = Benjamini-Hochberg.
- Min GO Level = 3.
- Max GO Level = 8.
- Minimum Number of Genes = 20.
- Min Percentage = 5.0.

For the graphical representation of the enrichment analysis among pop-DR genes, ClueGO clustering functionality was used (kappa threshold score for considering or rejecting term-to-term links set to 0.4 for *Salmonella* and 0.5 for *Listeria*, fraction for groups merging = "25%" in both cases). Only clusters with at least three GO terms were plotted. The detailed results of this analysis are presented in Table S3, where terms enriched at FDR < 0.1 are registered.

Signatures of Selection

F_{ST} values between the YRI and CEU individuals were obtained using data coming from 1000 Genomes data (phase 3 20130502). The phased data were downloaded from Impute reference data panel (https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html) and were filtered for biallelic SNPs found in either the CEU (n = 99) or the YRI (n = 108) samples. The phased genotypes were obtained using ShapeIT v2.r790 (Delaneau et al., 2013) with the '-output-vcf' option. The vcf files (one for each chromosome) were then converted to the PLINK format using vcftools v0.1.12b (https://vcftools.github.io/man_0112b.html) (Danecek et al., 2011). PLINK files were afterward converted to Arlequin format using the PGDSpider program (Lischer and Excoffier, 2012). Arlequin version 3.5.1.3 (http://cmpg.unibe.ch/software/arlequin35/) (Excoffier and Lischer, 2010) was then used to calculate F_{st} estimates derived from ANOVA. Integrated Haplotype Scores (iHS) (Voight et al., 2006) and cross-population Extended Haplotype Homozygosity (XP-EHH) (Sabeti et al., 2007) scores were calculated with the program selscan v1.1.0b (Szpiech and Hernandez, 2014) with default parameters. We defined high iHS values as those above the 99th percentile of genome-wide distribution in the CEU ($|\text{iHS}| > 2.70$) and the YRI ($|\text{iHS}| > 2.68$) populations. For XP-EHH, we used YRI as the reference set of haplotypes. Therefore, negative XP-EHH values correspond to longer haplotypes in the YRI population, and positive XP-EHH values correspond to longer haplotypes in the CEU population.

Coalescence Neutral Simulations for Evaluating Putatively Selected eQTL Sites

We identified 258 genes carrying a signature of recent positive selection ($|\text{iHS}| >$ 99th percentile of the genome-wide distribution) in either CEU or YRI samples. In order to provide additional support for adaptive evolution for each of these genes, we performed 500 replicates of neutral simulations matched to the known demographic histories of CEU and YRI populations (Gutenkunst et al., 2009), the observed allele frequency of the putatively selected variant (if several were present because of strong linkage disequilibrium, we chose the one with the highest iHS value), and the local recombination rate around that candidate eQTL. Simulations were performed with the program mssel, a modified version of ms (Hudson, 2002) that simulates the coalescent process conditional on a frequency trajectory.

For each site, and separately for each population, we simulated 500 allele frequency trajectories. These were simulated backward in time with the appropriate demographic history (Gutenkunst et al., 2009) from a fixed present-day allele frequency, which we take to be the observed allele frequency in the population of interest.

To determine the local recombination rate surrounding the putatively selected eQTL site, we took a 1Mbp window (500kb on each side) around the site and calculate the number of centimorgans based on the genetic map. For a region of length d centimorgans, we use the Haldane's Map Function to estimate the probability of recombination, $r = (1/2)(1 - e^{-2d/100})$. We then compute the population scaled recombination parameter as $\rho = 4N_e r$, where $N_e = 10000$ is taken to be the ancestral effective population size.

The population scaled mutation rate in mssel/ms is parameterized as $\theta = 4N_e L\mu$, where L = 1,000,000 is the length in bp of the region, $N_e = 10000$ is the effective population size, and $\mu = 10^{-9}$ is the mutation rate per site per generation.

Thirteen of our putatively selected variants were within 500kb of the edge of our genetic map, and hence, we could not estimate recombination rates, and 2 had such high recombination rates that coalescent simulations did not finish. These were excluded from the analyses.

Determining Significance of iHS Scores

We calculate iHS scores for all neutral simulations using selscan v1.1.0b, as we did for the real data. For each eQTL site in the real data, we took the unstandardized iHS score that was observed for a given population and normalize it (subtracting the mean and dividing by the standard deviation) with the iHS score of the frequency matched neutral simulation, giving 500+1 total scores in the normalization. As normalized iHS scores have a standard normal distribution (Voight et al., 2006), the scores can be treated as Z-scores. From these Z-scores, we calculated a p value for the observed scores based on the standard normal distribution.

Identification of Neanderthal-like Sites

Bi-allelic SNPs across five European population samples (CEU, FIN, GBR, IBS, TSI), three African population samples with low levels of Eurasian ancestry (ESN, MSL, YRI), and ancestral allele were extracted from the phase 3 release of the 1,000 Genomes Project. SNPs with alleles segregating in any of the three African samples were removed from the analysis.

Genotypes at the remaining SNPs were extracted from the high-coverage Altai Neanderthal genome after applying the minimum set of quality filters (map35_50 downloaded from https://bioinf.eva.mpg.de/altai_minimal_filters/). To summarize, Neanderthal-like sites were called as all bi-allelic SNPs for which the Altai genome carries a derived allele, the derived allele is segregating in the European sample, and the African samples are fixed for the ancestral allele.

DATA AND SOFTWARE AVAILABILITY

Software

All the Software packages and methods used in this study have been properly detailed and referenced under "QUANTIFICATION AND STATISTICAL ANALYSIS."

Data Resources

All data generated in this study are freely accessible via the ImmunPop QTL browser (http://www.immunpop.com). RNA sequencing data reported in this paper is available in GEO: GSE81046.

Supplemental Figures

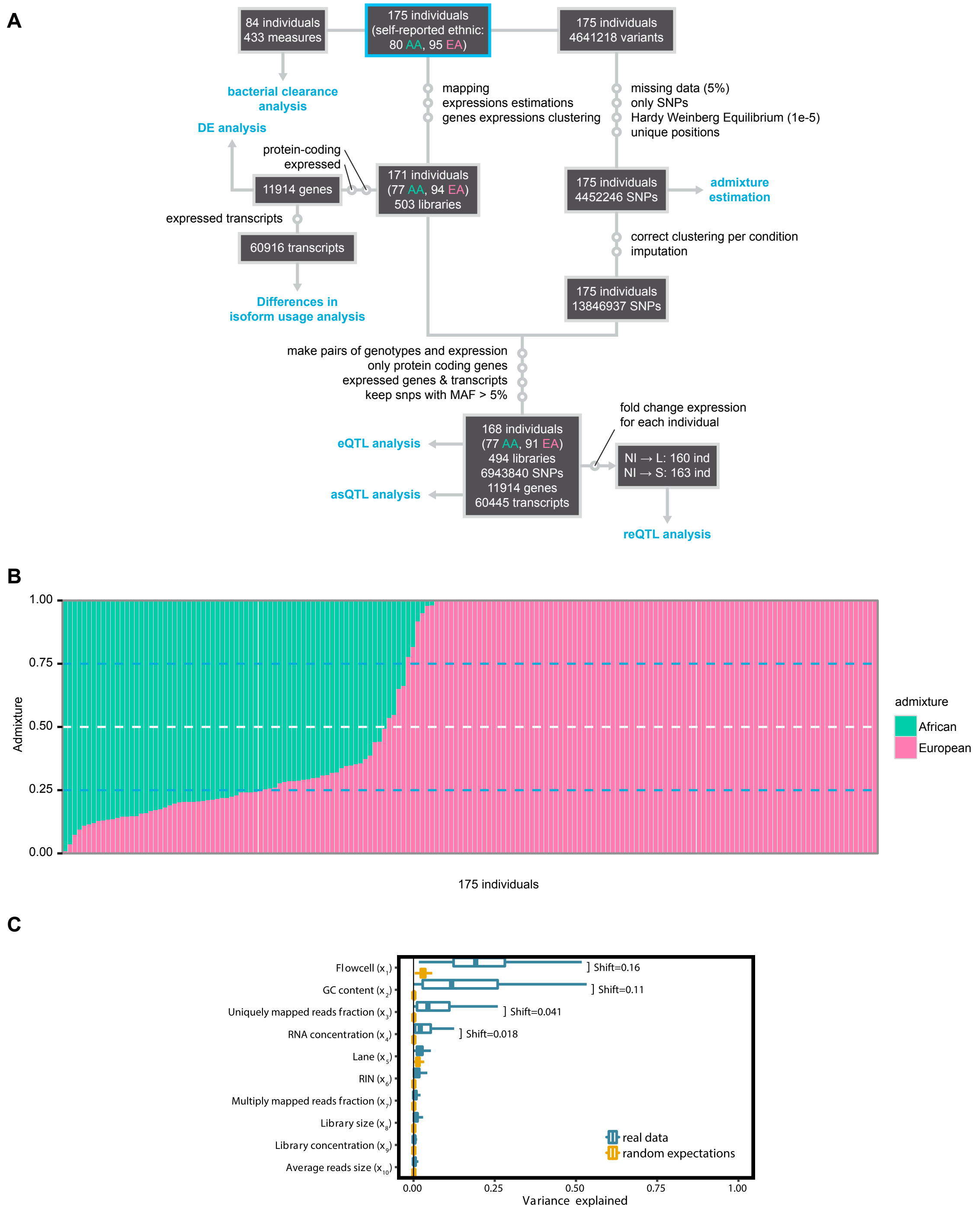


Figure S1. Study Design and Evaluation of Technical Confounders, Related to STAR Methods

(A) Schematic representation of the different steps used to process the RNA-sequencing and the genotyping data. The figure also depicts the number of samples and SNPs included in each of the analyses reported in the manuscript. (B) Population structure analysis of all samples based on autosomal SNPs. Each individual is represented as a vertical line, with population origins indicated below the lines. Cluster membership proportions are depicted in green (inferred proportion of African ancestry) and pink (inferred proportion of European ancestry). (C) Fraction of variance explained in the RNA-seq data by potential technical confounders. For each confounder, the shift between the distribution of variance explained by the real data and by random (when shuffling the real data) was estimated using a non-parametric Mann-Whitney U test. Confounders with shifts larger than 1% variance (shown in the figure) were corrected for in downstream analysis.

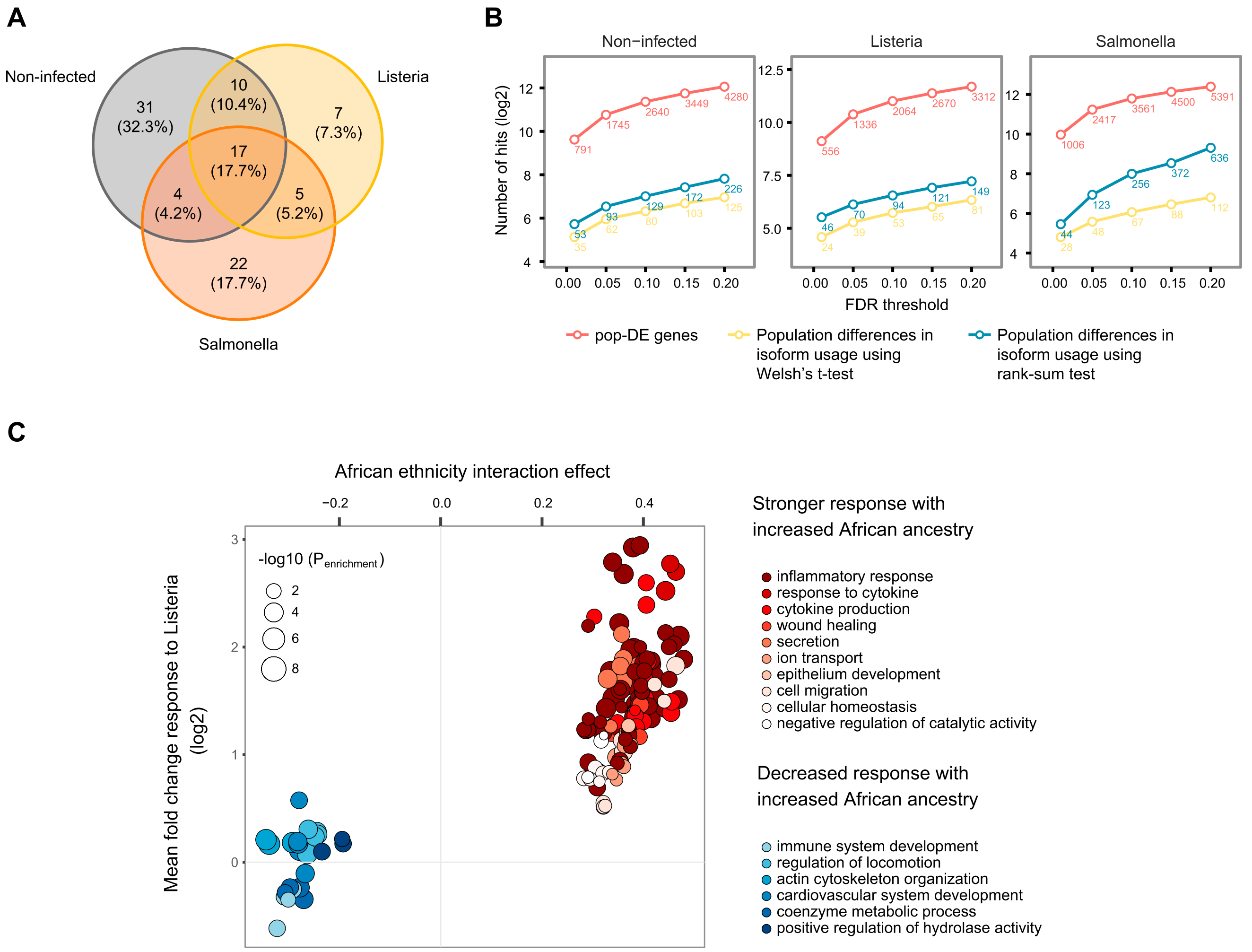


Figure S2. Population Differences in Gene Expression and Isoform Usage, Related to Figure 1

(A) Venn diagram for the overlap of genes showing significant differences in isoform usage in the different experimental conditions. Non-infected, *Listeria*-infected, and *Salmonella*-infected genes with ancestry-associated differential isoform usage at FDR < 0.05 are illustrated in gray, yellow, and orange, respectively. (B) Number of genes identified as pop-DE and with ancestry-associated changes in isoform usage at different FDR cutoffs. The number of significant genes at the different cutoffs (x axis) is reported in log2 scale (y axis). For differences in isoform usage, we report results obtained using the Welsh's t test (yellow) and the rank-sum test (blue). (C) Gene ontology enrichment analysis for genes showing a significant interaction between ancestry and response to *Listeria*. Enrichments were performed separately for genes showing a significantly stronger and a significantly weaker response to Listeria (pop-DR) with increasing African ancestry (i.e., positive and negative interaction effects, respectively, as illustrated on the x axis). Only GO-terms with an enrichment at FDR < 0.1 are displayed, where GO terms are grouped into clusters and colored accordingly based on the overlap among gene sets (also obtained from ClueGO's clustering functionality). For each GO-term (each circle), the average interaction effect is plotted on the x axis, against the mean log2 fold change in gene expression levels in response to infection for that term (y axis).

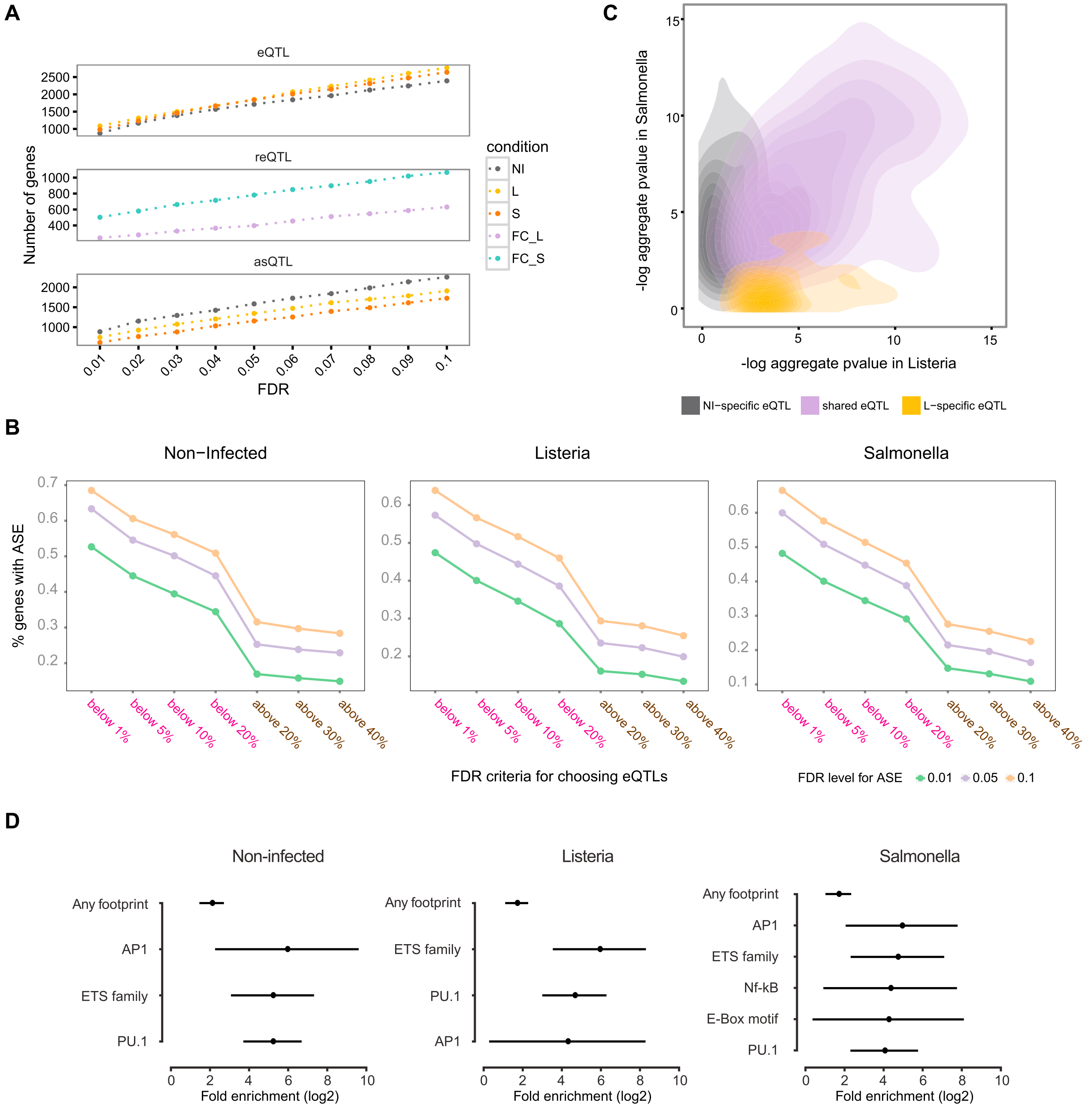


Figure S3. eQTL Are Enriched for Binding Sites of Actively Regulated TF Binding Sites, Related to Figure 3

(A) Comparison of the number of genes associated with an eQTL (top), reQTL (middle), and asQTL (bottom) at increasingly higher FDR cutoffs (x axis). (B) Percentage of genes showing significant ASE based on different FDR cutoffs adopted for ASE and cis-eQTL mapping. The plots depict the percentage of genes showing significant ASE (y axis) out of the total number of genes with cis-eQTL that pass FDR thresholds shown on the x axis. Three different FDR cutoffs are studied for ASE statistical significance, while the eQTL FDR thresholds on the x axis cover a wide range from extremely significant (to the left of the x axis) to extremely non-significant (to the right of the x axis). In particular, FDR criteria for selecting significant and non-significant *cis* regulatory variants are illustrated on the x axis in pink and brown, respectively. (C) Plot contrasting the evidence for ASE (-log10 *P values*) in non-infected macrophages (y axis) and in macrophages infected with *Listeria* (x axis), for genes where we identified cis-eQTL in both conditions (purple), genes for which cis-eQTL were only found in non-infected macrophages (gray), and genes for which cis-eQTL were only found in Listeria-infected macrophages (yellow). (D) ATAC-seq eQTL enrichments (x axis) in actively-regulated TF binding sites annotated by ATAC-seq footprinting. Error bars show 95% confidence intervals. Only significant enrichments are shown. Binding sites were grouped into functionally-overlapping “TF clusters” using sequence similarity and co-localization in the genome.

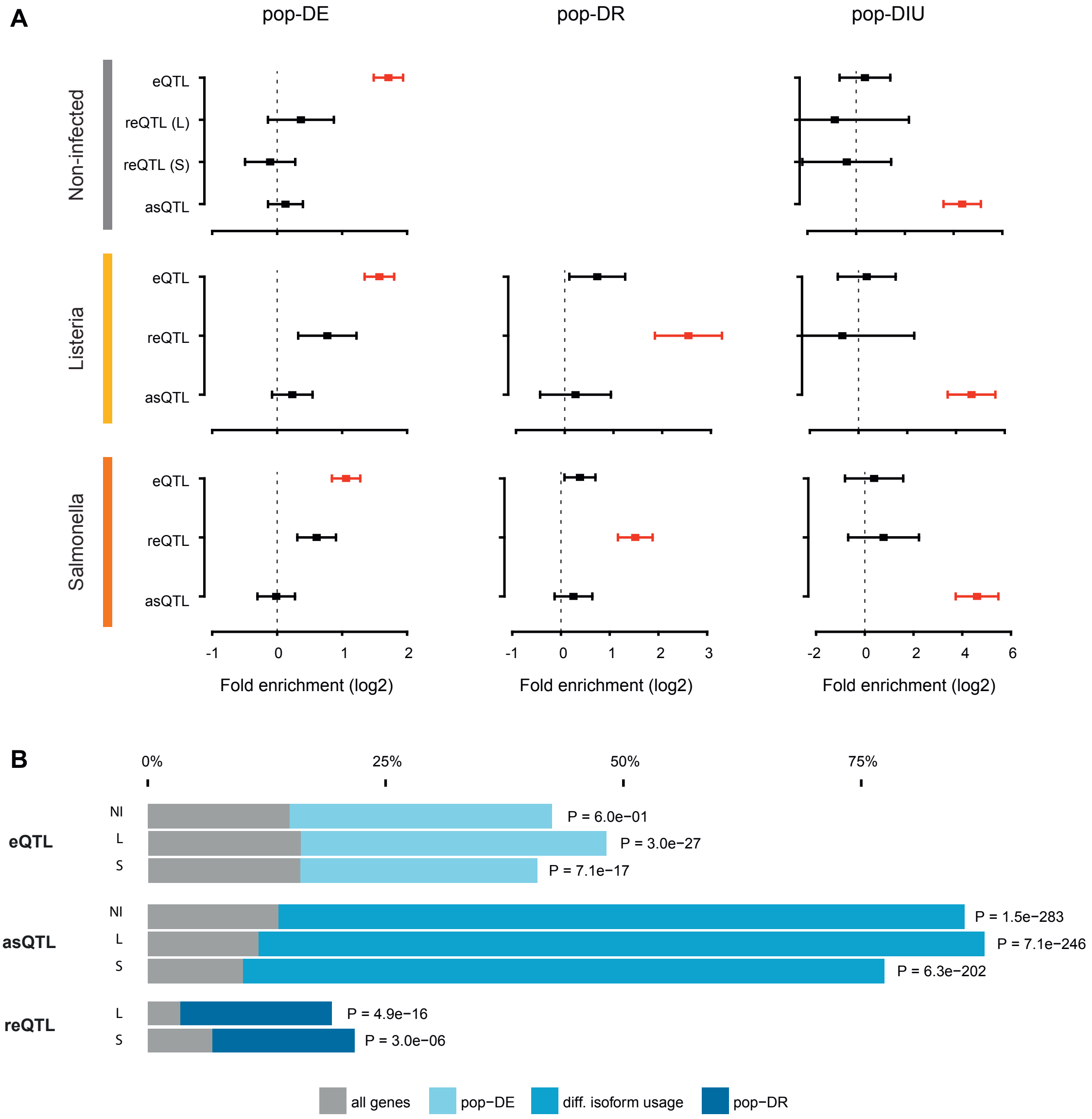


Figure S4. Contribution of *cis* Genetic Variation to Ancestry-Associated Transcriptional Variation, Related to Figure 4

(A) Enrichment of regulatory variants among pop-DE, pop-DR and genes that exhibit ancestry-associated differential isoform usage (pop-DIU). The enrichment factors are shown on the x axis in a log2 scale. The bars around the estimated enrichments reflect the 95% confidence intervals around the estimates. For pop-DE genes, enrichments were obtained from a logistic regression model aimed at testing if pop-DE (FDR < 0.05) (as compared to non-pop-DE genes; FDR > 0.05) were enriched among cis-eQTL, cis-reQTL or cis-asQTL. The same was done for pop-DR and genes showing differences in isoform usage between populations. (B) Proportion of pop-DE, pop-DR and genes that exhibit ancestry-associated differential isoform usage that are associated with a cis-eQTL, cis-reQTL or cis-asQTL, respectively (FDR < 0.05). Null expectations (based on the genome-wide proportion of genes associated with each QTL class) are shown in gray.

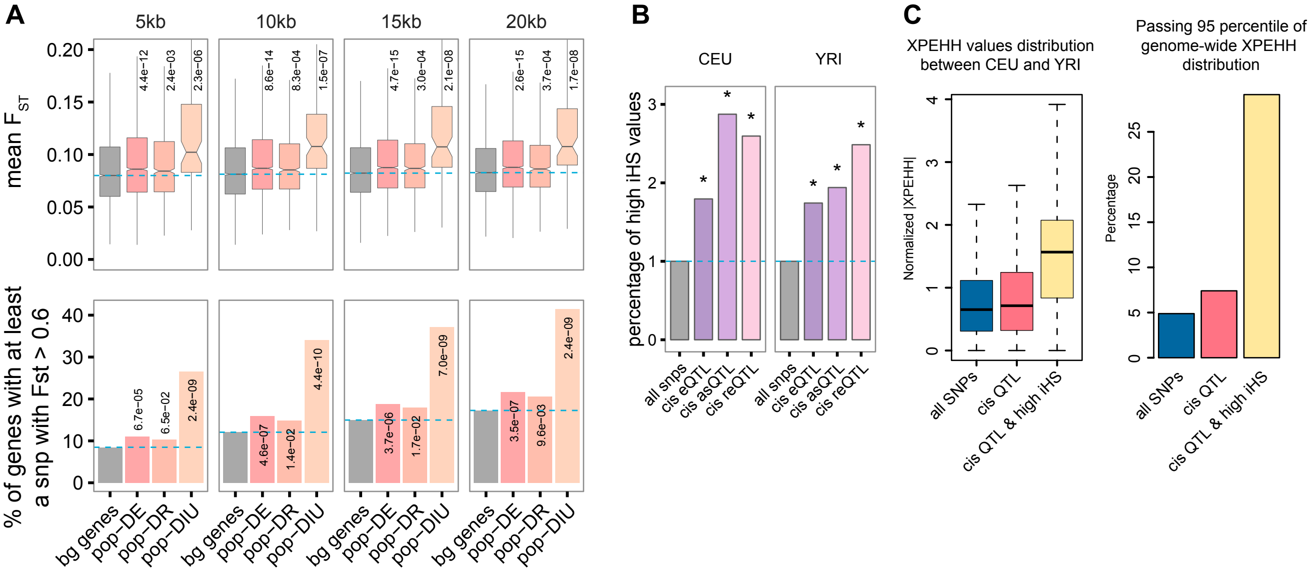


Figure S5. Natural Selection Contributes to Ancestry-Associated Regulatory Differences, Related to Figure 5

(A) The top panel shows boxplots of mean Fst values in a window of different sizes (mentioned above the plots) around the TSS of all genes, pop-DE, pop-DR and genes showing differences in isoform usage between populations. The bottom panel shows that proportion of genes in each of the above-mentioned categories that have at least one SNP in the window with an Fst value above 0.6 (the 99th percentile of the genome-wide distribution). (B) Proportion of all SNPs, cis-eQTL, cis-reQTL, and cis-asQTL identified at an FDR < 0.05 with an iHS value above the 99th percentile of the genome-wide distribution in the CEU (|iHS| > 2.70) and the YRI (|iHS| > 2.68) populations. (C) Boxplot showing the distribution of absolute XP-EHH values (x axis) among all *cis* SNPs tested (blue), the group of SNPs impacting one or more transcriptional phenotypes (i.e., cis-eQTL, cis-reQTL or cis-asQTL; pink), and SNPs impacting one or more transcriptional phenotypes that show an elevated iHS values (yellow). The right panel shows the proportion of SNPs (y axis) belonging to each of the groups described above that have an XP-EHH value above the 95th percentile of the genome-wide distribution. For all comparisons, QTL with an elevated iHS values show significantly higher XP-EHH values ($p < 1x10^{-10}$) as compared to all SNPs or cis-regulatory variants.