Detecting Signals of Coevolution within Immune Response Pathways in the African-American Ancestral Population

by

Catherine Marina Ross

A thesis submitted in conformity with the requirements
for the degree of Master of Science

Department of Molecular Genetics
University of Toronto

Detecting Signals of Coevolution within Immune Response Pathways in the African-American Ancestral Population

Catherine Marina Ross

Masters of Science

Department of Molecular Genetics
University of Toronto

2018

Abstract

The natural variation of human populations was shaped by mechanisms of evolutionary selection as our species began their worldwide migration thousands of years ago. Candidate gene and genome-wide association studies have revealed evidence for ancestry-specific positive selection, yet we lack insights regarding the broader biological context of population-driven selection. In this thesis, I describe my efforts to explore how interactions within biological pathways might impact evolution, using a comparative population-based analysis of pathway enrichment and inter-chromosomal allelic association. I compared genetic data from individuals of European-American and African-American ancestry, and identified enrichment for several biological processes, with prominent signals of genetic coevolution in two immune-associated pathways among African-Americans exclusively. Substantiated by considerable experimental-based literature, these findings suggest an effect of population variation on pathway-level selection, in which a global comparative analysis would further the ultimate goal of precision medicine.

Acknowledgments

I would first like to thank my thesis advisors Dr. Charlie Boone and Dr. Gary Bader at the University of Toronto. At any point where I had run into a problematic spot or had a question about my research, the doors to their offices were always open. This thesis would not have been possible without their shared scientific insight and guidance throughout my studies.

I also thank my supervisory committee members Dr. Brenda Andrews, Dr. Philip Awadalla, and Dr. Lucy Osborne whose encouragement and knowledgable feedback have been most valuable.

I must give a special thank you to Shraddha Pai for her continued patience and guidance during the course of my research and graduate school, in general—her knowledgable support was invaluable to the completion of this thesis (it is still hard for me to believe I had next to no experience in statistical programming before beginning my Masters). I also have to thank the entire Stack Overflow online community for their plentiful and immeasurably useful expertise in the universe of R programming and beyond.

Finally, I must express my upmost gratitude to my parents, sister, and a handful of close friends and loved ones for providing me with unwavering support and reassurance throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them all. Thank you.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

1KGP:	1000 Genomes Project
ASW:	African ancestry in Southwest USA
CEU:	Utah residents (CEPH) with Northern and Western European ancestry
dbPSHP:	Database of Positive Selection across Human Populations
ES:	Enrichment score
FDR:	False discovery rate
GSEA:	Gene Set Enrichment Analysis
GWAS:	Genome-wide association study
HM3:	International HapMap Project phase 3
LD:	Linkage disequilibrium
MAF:	Minor allele frequency
NES:	Normalized enrichment score
PNC:	Philadelphia Neurodevelopmental Cohort
SNP:	Single nucleotide polymorphism

1. Thesis rationale and structure

Several lines of evidence show that numerous regions of the human genome are under forces of adaptive evolutionary selection, which has enabled the analysis of mechanisms by which selection has shaped the natural variation of human populations. Early studies in the field primarily employed scans for positive selection upon the identification of trait-associated variants via candidate gene and genome-wide association approaches. Although this research provided compelling evidence for ancestry- or region-specific positive selection acting on various genomic loci, evidence remains lacking for selective pressures acting within biologically meaningful pathways associated with human ancestral background, particularly among the ethnically underrepresented. This gap in knowledge emphasizes the need to explore evolutionary mechanisms in a broad biological and phenotypic context. An opportunity to address this gap is presented by the plethora of ethnically diverse panels of single nucleotide polymorphism (SNP) genotyping data and pathway annotation resources currently available to the scientific community. In this thesis, I used computational and biostatistical methods to investigate evidence for the phenotypically-favourable maintenance of physically unlinked genetic interactions within pathways (i.e., SNP-SNP pairs annotated to genes on separate chromosomes within the same pathway) of biological relevance and importance to human ancestral background. I uncovered evidence for within-pathway epistatic coevolution based on differential signals of inter-chromosomal allelic association between individuals of European-American and African-American ancestry. A flowchart summarizing the computational pipeline I developed is shown in **Figure 1**.

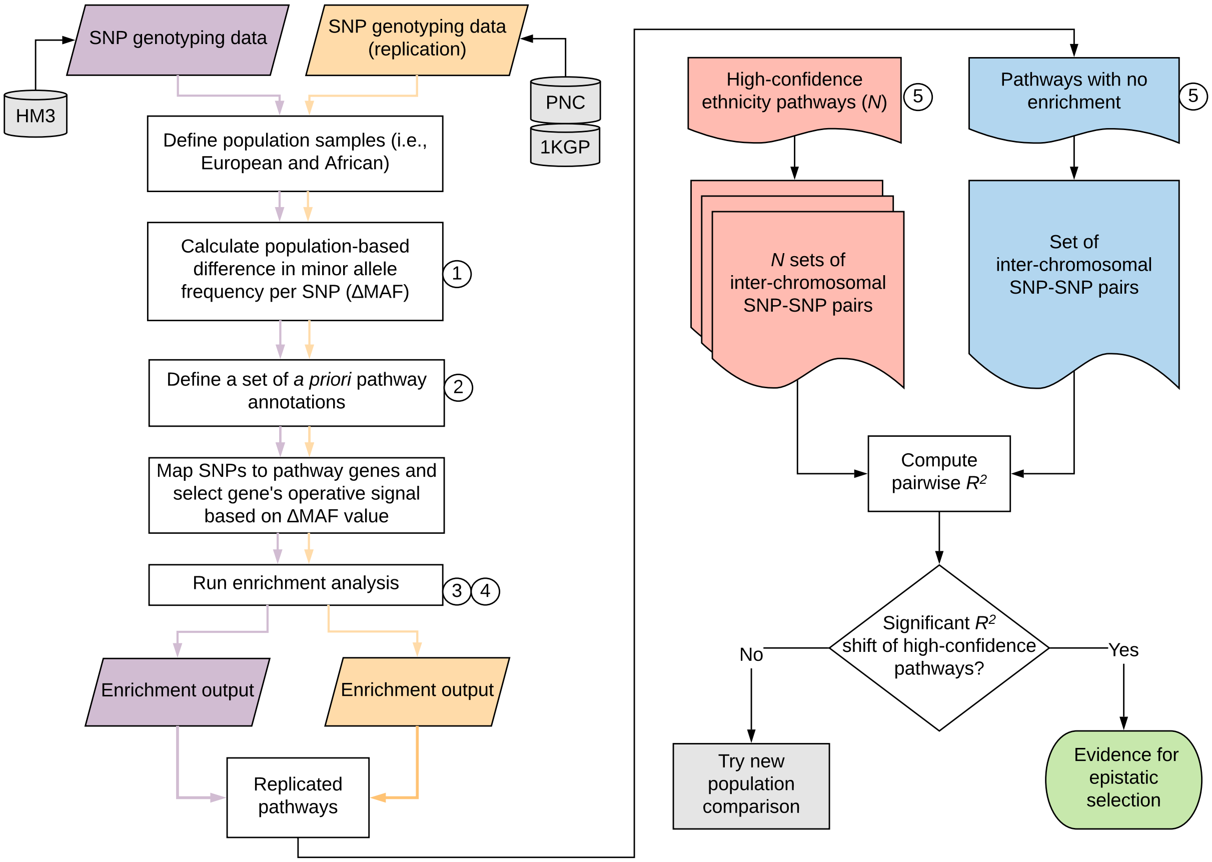

SNP genotyping data
SNP genotyping data (replication)
HM3
PNC
1KGP
Define population samples (i.e., European and African)
Calculate population-based difference in minor allele frequency per SNP (ΔMAF)
1
Define a set of *a priori* pathway annotations
2
Map SNPs to pathway genes and select gene's operative signal based on ΔMAF value
Run enrichment analysis
3
4
Enrichment output
Enrichment output
Replicated pathways
High-confidence ethnicity pathways (*N*)
5
Pathways with no enrichment
5
N sets of inter-chromosomal SNP-SNP pairs
Set of inter-chromosomal SNP-SNP pairs
Compute pairwise R^2
Significant R^2 shift of high-confidence pathways?
No
Yes
Try new population comparison
Evidence for epistatic selection

Figure 1. Flowchart illustrating the computational pipeline developed for discovering population-driven pathway enrichment and selective associations in human genomics data. Circled numbers adjacent to text boxes represent modifiable arguments at various stages of the pipeline that require predefinition. These arguments require the user to: (1) determine the method of calculating the SNP-level test statistic (i.e., directional or absolute value); (2) limit the minimum and maximum number of genes per pathway; (3) run the genotype- or phenotype-based permutation method; (4) set the number of permutation cycles to run; and (5) set the thresholds to define significant pathway enrichment and non-enrichment. **Abbreviations**: HM3; International HapMap Project phase 3; PNC, Philadelphia Neurodevelopmental Cohort; 1KGP, 1000 Genomes Project; SNP, single nucleotide polymorphism; GSEA, Gene Set Enrichment Analysis.

2. Introduction

2.1. The history of modern human adaptation

Our understanding of life on Earth was fundamentally shifted when British naturalists Alfred Wallace and Charles Darwin each described their discoveries of the theory of natural selection in the mid-19th century [1]. Darwin's primary work, *On the Origin of Species by Means of Natural Selection, or the Preservation of Favored Races in the Struggle for Life [2]*, provided the scientific community, along with the entire global community, insight into the mechanisms by which evolutionary change is mediated by natural selection. In the years following this seminal discovery, research in the field of population genetics had advanced extensively, paving the way to considerable progress in the elucidation of the various molecular and mechanistic factors behind the evolution of natural populations [3]. Furthermore, we now understand the important contribution of both random chance and natural selection as drivers of evolutionary change [4, 5], an understanding that has resulted in the evolution of the very concept of natural selection. Several modes of selection have thus been characterized, all of which ultimately stem from whether an allele is advantageous or deleterious within a certain population. Like Darwin and Wallace, I was primarily interested in exploring modes of adaptive evolution, which is controlled by positive evolutionary forces acting on phenotypically-beneficial loci [6].

The story of modern human adaptive evolution begins with the out-of-Africa migration that occurred roughly 200,000 years ago. Our species spread across the globe to inhabit a variety of novel habitats, from tropical to arctic climates, from high- to low-altitude terrains, and even from regions of high-fat to starch-rich nutritional diets. During this migration period, our early human ancestors encountered and interbred with archaic populations such as Neandertals and Denisovans, resulting in the introgression of archaic genomes into the modern human genomes of non-African, and possibly African, populations [7, 8]. In addition, several human populations underwent rapid growth in population size, primarily due to the transition from a primitive hunting-gathering lifestyle to the era of practicing agriculture and pastoralism. This shift in lifestyle resulted not only in rapid population growth and increased population densities, but also an overall increase in infectious diseases [9]. As a result of having to adapt to such newly diverse dietary practices and geographic environments, local positive selection pressures were driven towards population- or region-specific mutations that influence adaptive phenotypes, such as skin pigmentation, height, lactase persistence, hypoxia modulation, and endemic pathogenic response [9].

Interpreting the multifaceted nature of local human adaptation has proven to be particularly challenging, as it ultimately requires the "identification of the genomic regions under selection, the phenotypes that selection is acting upon, and ideally, the external conditions driving the selection" [9]. Candidate gene and genome-wide association studies (GWAS) represent two complementary approaches traditionally used to identify the genetic basis of common diseases [10], and have been widely used to draw connections between genetic variation and natural selection before the maturation of population genomics [11]. In this vein, several methods have used these complementary approaches to identify candidate trait-associated variants and subsequently employ genome-wide scans for selection [12, 13]. Although these single-loci methods have their inherent pitfalls, as I will discuss in further detail below, they have shed considerable light on the history of global human adaptation, assisting us in understanding how natural selection has shaped modern population variation. Notably, this work has provided compelling evidence for strong population-specific selective pressures acting upon numerous genomic loci, some of which include: *LCT* [14], granting the persistence of lactose tolerance throughout adulthood within various populations of northern Europe and Africa; *EPAS1* [15], allowing individuals of Tibetan ancestry to thrive in a heavily oxygen-depleted environment; and *G6PD* [16], *DARC* [17, 18], and *HBB* [19, 20], which, on their own, confer a reduced risk to malaria infection within populations of sub-Saharan Africa.

2.2. Pathway analysis—bridging the gap between complex genotype and adaptive phenotype

As of September 2016, numerous human GWAS publications have described more than 20,000 unique SNP (single nucleotide polymorphism) associations to a wide variety of diseases and traits [21], after which candidate gene approaches can be applied to gain an improved understanding of those genetic associations. However, due to the inherent circular nature of the candidate gene study, inferring new hypotheses about development and disease from GWAS-identified SNP associations has been a major ongoing challenge. A candidate gene study requires both the formation of an *a priori* hypothesis about the potential genes under selection, as well as an understanding of the underlying genotype–phenotype relationships, an understanding that is crucial to forming the initial hypothesis [22]. For a handful of traits with a well-defined, Mendelian-like architecture, such as lactase persistence and cystic fibrosis, the identification of the proper candidate genes to study can be accomplished due to the ease of distinguishing their adaptive phenotypes. However, this identification process becomes increasingly difficult for complex phenotypes with uncertain genetic architecture [6], which is the case for the majority of common traits and diseases. Several studies have thus attempted to focus on genes that frequently appear as targets of selection, such as those involved in the immune response [23], but could ultimately lead to a biased set of candidate loci. Also, the majority of significant trait-associated functional variants are located in regulatory regions far removed from genic loci [24], which poses a major issue when attempting to efficiently detect positive selection using a candidate gene approach [6]. Finally, those variants with genome-wide significance typically account for a small portion of the observed heritability of a given trait—a phenomenon commonly known as the problem of missing heritability [25].

Researchers have since explored the mystery of missing heritability, and evidence suggests that estimates of total heritability may be inflated by genetic interactions [26]. Unlike additive effects, genetic interactions describe unexpected phenotypes that result from combinations of two or more functionally-related genetic variants [26, 27]. For example, a synthetic lethal genetic interaction results when the combined mutation of two or more genes causes cell death, while the mutation of either single gene does not [28]. Recently, the first global genetic interaction network for any system was completed by mapping millions of double mutant genetic interactions in *S. cerevisiae* [29]. The network revealed the complex functional wiring of the cell, demonstrating that interactions tend to occur in particular network structures connecting across functionally-related biological mechanisms and pathways. Within the yeast genome, ~1000 genes are essential for haploid cell viability and may be considered as Mendelian-like genetic traits as the phenotype is manifest following mutation of a single gene [30, 31]. In contrast, ~500,000 negative digenic interactions exist in the global genetic interaction network, including ~10,000 synthetic lethal interactions between nonessential genes [29], highlighting the power of genetic interactions to compound phenotypes associated with single mutations and generate unexpected phenotypes. Previous studies have shown that for essential pathways and gene complexes (e.g., the proteasome), numerous within-pathway hypomorphic alleles can combine to generate a synthetic lethal phenotype [32, 33]. However, many more between-pathway synthetic lethal interactions have identified pathways that operate together to control essential cellular functions [34]. Between-pathway interactions in human genetics data have recently been explored using a method known as BridGE (Bridging Genes with Epistasis) [34], which I discuss in more detail later [**4.3 Research applications**].

A genetic interaction in which one mutation masks or suppresses the effects of an allele at another locus is referred to in classical genetics as epistasis, and may explain a significant component of missing heritability [25, 26], though the mechanistic details are difficult to infer [35]. It is clear from many genetic studies that epistatic interactions can affect heritability [36]. For instance, it is fairly common for the phenotypic effect of a gene knockout to be masked by a second knockout [37], or for an additional mutation to be required to elicit a phenotypic effect in a particular mutant background [38]. Epistasis is also of fundamental importance at the macroevolutionary scale. In terms of molecular evolution, advantageous changes in the genotype of one species are often deleterious in others, indicating that the genotypic effect on fitness is dependent upon the genetic background in which it is identified [39]. Below, I will return to a discussion of the role of epistasis in relation to missing heritability and evolutionary selection. But first, I will highlight a relatively novel methodological approach that has been devised to address the aforementioned issues regarding candidate gene and genome-wide association studies—pathway enrichment analysis.

Given our understanding that most common allele associations identified by GWAS have exhibited modest effect sizes and that genes typically function and interact within functionally associated pathways and networks [35], genome-wide datasets are increasingly being used as foundations for uncovering pathways and networks associated to phenotypes. A major motivation for this type of analysis is the vital importance of developing strategies for the diagnosis, treatment, and prevention of complex, polygenic diseases [36]. During the last several years, pathway-based methods have been increasingly used for secondary post-GWAS analysis [22]. Generally, pathway enrichment analysis tests for the significant relationship of biologically or functionally associated sets of genes with a phenotype, and can reveal hidden effects that would otherwise be missed from single loci analyses [22]. In cases where individual loci fail to exhibit significant genome-wide association with a trait, pathway-based studies have demonstrated that functionally related loci can collectively have a large impact on biological phenotypes and disease susceptibility, as has been shown in studies of breast cancer [37], Crohn's disease [38], and type 2 diabetes [39]. Pathway analysis thus represents an increasingly "powerful and biologically-oriented bridge between genotypes and phenotypes" [22], and will serve as the primary tool I utilize to draw functional associations between pathways and human ancestry.

2.3. The progressive diversification of the genomics study sample landscape

For my thesis work, I take advantage of the diverse catalogues of human genetic variation (i.e., SNP array-based genotyping data) that have become available to discover how biologically meaningful pathways are associated with ancestral background. Prior to 2009, however, this effort would have been challenging since the vast majority of GWAS participants were of European ancestry (96%) [40]. In an effort to avoid limiting the benefits of genomics research and medicine to "a privileged few" [41], the scientific community began investigating individuals from an increasingly broader panel of ancestral backgrounds, and as a result, increased the proportion of GWAS participants of non-European descent or admixed origin five-fold by 2016 to nearly 20% [42]. However, most non-European participants were of Asian ancestry, with individuals of African and Latin American ancestry, Hispanic people, and native or indigenous peoples together representing less than 4% of all samples analyzed in GWAS [42].

Thus, despite the US National Institutes of Health (NIH) mandating the inclusion of diverse participants in its funded biomedical research more than 20 years ago, NIH-funded genomics studies are still missing a large portion of global genetic variation [42]. The lack of diversity in study samples impairs our ability not only to accurately identify associations between genetic variants and disease in people of a variety of ancestral backgrounds, but also affects our ability to detect rare conditions in underrepresented populations [43]. A complex network of logistical, cultural, and historical factors likely contribute to this observed “European bias” [42] in GWAS, and the lack of sampling diversity remains an issue when considering the widespread application of genome-wide association findings to the greater global population. In the case of targeted drug therapies, ancestry-specific differences determine the frequency of metabolism-associated variants, and consequently have crucial implications in the consideration of drug efficacy and safety between individuals of various backgrounds [44]. Additionally, recent research has found that, compared with Europeans, individuals of African ancestry have a greater chance of being genetically misdiagnosed with a mutation that gravely increases the risk of hypertrophic cardiomyopathy [45], a misdiagnosis that could be prevented by including more ethnically diverse controls in candidate gene studies [43]. In fact, this misclassification has been primarily attributed to the fact that the original study control sample was mostly comprised of European individuals [46]. Since the implications of using undiversified ethnic sampling in genomics studies are significant, several integral efforts have since addressed the pervasive sampling bias in this line of research.

One of the first resources describing population genetic variation was released in 2001, providing a high-density SNP map of over 1.4 million genetic variants [47] of which ~2% were genotyped in three populations by the SNP Consortium [48]. Closely following this initiative was the International HapMap Project [49], an effort that provided a publicly available microarray-based variant catalogue and associated haplotype map of 3.8 million variants genotyped in several populations from Europe, Africa, Asia, and America, allowing researchers to begin studying natural variation at the population level. More recently, whole genome and whole exome sequence data are being used for genotyping studies as accessibility to these tools increases [42]. In 2008, the 1000 Genomes Project (1KGP) [50] was launched to establish a deep catalogue of human genetic variation, and by October 2015, the project had genotyped ~88 million variants across 26 populations worldwide. Prior to the recent release of the final 1KGP dataset, the previous data releases had inspired researchers to begin the fine mapping of GWAS-identified genomic loci [51], revealing further insights into the molecular basis of several complex diseases. Other studies have identified between-population differences in the genetic architecture of several traits and diseases using ancestry-specific variant panels derived from the

1KGP data [52]. Using these rich catalogues of ethnically diverse genetic variation provided by the HapMap and 1000 Genomes projects, I aimed to identify signals of ancestry-specific selection within the context of biologically meaningful pathways via statistical measures of population variation.

2.4. Detecting evidence for pathway-level epistatic coevolution via linkage disequilibrium

A key distinguishing feature of genetic variation among human populations is the nonrandom association between pairwise combinations of sites in the genome, otherwise known as linkage disequilibrium (LD) [53]. In statistical terms, LD is defined as the difference between the observed frequency of a particular combination of alleles at two loci and the frequency expected for random association, based on the assumption that recombination will result in an equilibrium distribution of alleles at each locus given sufficient evolutionary time [54]. Between proximally linked sites in the genome, disequilibria is predominantly the result of random genetic drift, and comparably, the common ancestry of unrecombined regions of a chromosome [55]. Such instances of proximally linked, or short-range, LD are of great practical interest as they can be used for the identification and/or localization of genes contributing to disease susceptibility [56]. Additionally, recent and ongoing selective sweeps can be identified using blocks of unrecombined chromosome regions [57, 58]. Though these instances of short-range LD have been well studied, LD between sites separated by considerably larger distances, including those residing on entirely separate chromosomes, have not been characterized to the same extent. Despite the term ‘linkage’ in linkage disequilibrium, alleles at separate physically unlinked loci may also exist in disequilibrium [53], and was originally devised to detect allelic association due to epistatic selection [59, 60]. Notably, such instances of allelic association could potentially reveal the presence of significant evolutionary pressures on the genome, given that random recombination events tend to rapidly break down disequilibria between sites separated by large physical distances, and since sites on separate chromosomes are not physically linked. In other words, the existence of allelic association between physically unlinked genomic loci suggests the possibility of counteracting evolutionary forces at work [55].

Several potential forces may explain the observation of LD between largely separated or physically unlinked genetic variants, including population admixture, random genetic drift, hitchhiking, chromosomal structural variation, and epistatic selection [55]. Population admixture has been proposed to explain unusual patterns of long-range LD found in the Southern African Lemba tribal population [61]. Random genetic drift may also contribute to long-range LD but, even in a population at demographic equilibrium, recombination between distant chromosomal sites will largely, but not completely, eradicate LD caused by random genetic drift. Further, disequilibria can be generated and amplified by recurrent bottlenecks [62] and/or changes in population demography [63], respectively, as these forces may have contributed to the disequilibria identified in some non-African populations [64]. Genetic hitchhiking could also explain patterns of long-range LD by generating large haplotype blocks with a positively selected mutation, resulting in widespread disequilibria over the spanning region [57]. In terms of structural variation, chromosomal inversions can effectively alter patterns of recombination and consequently cause LD to extend over unusually large regions of a chromosome [65]. Finally, and of most interest for my thesis work, epistasis can create and maintain LD indefinitely [59] by selecting for specific combinations of alleles at different loci. Notably, epistatic coevolution between two physically unlinked loci has been reported in the human genome [66]; however, test power was not high enough to reliably detect associations given the study sample size. Therefore, by testing for significant signals of inter-chromosomal pairwise linkage disequilibrium within pathways enriched for human ancestry, I aimed to increase the power of detecting selection-induced allelic associations and determine evidence for biologically meaningful epistasis.

As eloquently stated in Phillips' paper, epistasis has long been recognized to be fundamentally important to understanding both the structure and function of genetic pathways and the evolutionary dynamics of complex genetic systems [67, 68]. Since the favoured phenotype depends on particular epistatic interaction of alleles at multiple loci, the nonrandom association between those alleles is expected to increase over generations [69], and can thus be used to study the occurrence of species-specific evolutionary forces. In model organisms such as *S. cerevisiae [70, 71]*, *C. elegans* [72], and *D. melanogaster* [73] these interactions have been widely observed, but in humans, they are difficult to detect and explain in a functional context. This challenge may reflect several factors relating to human samples, such as diverse genetic background, low allele frequency, limited sample size, complexity of interactions, and insufficient effect size [74, 75]. Despite these drawbacks, several genome-wide interaction-based association studies have provided evidence for epistasis in a variety of complex traits and diseases in humans [76-82]. As an interesting example, a team of researchers recently found that the epistatic interaction between the risk alleles *DDX39B* (rs2523506) and *IL7R* (rs2523506A) increases the risk of multiple sclerosis considerably more than the independent effect of either variant [82]. Thus, identifying comparable interactions across the human genome may uncover a significant source of missing heritability for several complex traits and diseases [25, 26], and can shed additional light on evolutionary forces acting within functionally associated pathways (e.g., [83-86]) in an ancestry-specific manner.

In this thesis, I take advantage of existing databases of biological pathway annotations, ethnically diverse panels of genotyping data, and Gene Set Enrichment Analysis (GSEA) to identify pathways enriched for human ancestral background. I then perform a test for within-pathway epistatic coevolution by measuring the extent and strength of allelic association between physically unlinked genetic interactions (i.e., statistical association of SNP-SNP pairs on separate chromosomes within a pathway) among those identified pathways. By explicitly filtering for pairwise combinations of variants interacting on separate chromosomes, I can effectively distinguish between signals of coevolution generated by epistatic selection from those generated by intrinsic cellular factors such as recombination, chromosomal structural variations and/or large sweeps resulting from genetic hitchhiking. Using this approach, I identify significant signals of adaptive coevolution within pathways primarily involved in the immune response among individuals of African-American ancestry.

3. Results

3.1. Environmentally associated pathways are enriched for human ancestry

To begin to identify biological pathways that are enriched for human ancestral background, I performed GSEA on the basis of SNP-level variation in minor allele frequency (ΔMAF) between individuals of European-American and African-American ancestry. I used SNP genotyping data from two independent datasets, the International HapMap Project phase 3 (HM3) and the Philadelphia Neurodevelopmental Cohort (PNC) (**Table 1**). I gathered pathway annotation data from the publicly available Bader Laboratory database (http://download.baderlab.org/EM_Genesets/), which provided a set of 3,781 biological pathways to test after controlling for pathway size (see Materials and Methods for additional details). Prior to running GSEA, I linked the array-based SNP genotyping data to pathways by the nearest-gene mapping method [87], and selected the single SNP with the largest positive ΔMAF test statistic to represent the operative signal of each gene. Typically, a GWAS is performed to generate SNP-level *p* values or chi-squared values that are then used for downstream pathway enrichment analysis [88]. However, when I ran GSEA using either *p* values or chi-squared values obtained from GWAS, I saw widely insignificant pathway enrichment results. Therefore, for my analysis, I implemented a novel ΔMAF test statistic method (see Materials and Methods) due to its computational suitability for the type of array-based genomic data I analyze, and since it represents a straightforward indicator of population-based genomic differentiation. The distribution of the calculated ΔMAF statistic within the HM3 and PNC datasets is shown in **Figure 2**.

I first ran GSEA with the SNP genotyping data taken from the publicly available HM3 dataset. Of the ~3,700 tested biological pathways, I identified 76 as significantly enriched (FDR ≤ 0.1) for ancestral background (**Supplementary Table 1; Table 2**). Given the chosen FDR threshold, however, a 10% probability exists that the significantly observed pathway enrichments represent false positive findings. Thus, in order to uncover pathways that are confidently enriched for ancestral background, I repeated this analysis using the genotyping data taken from the independent PNC dataset. In this replication analysis, I identified 237 pathways with

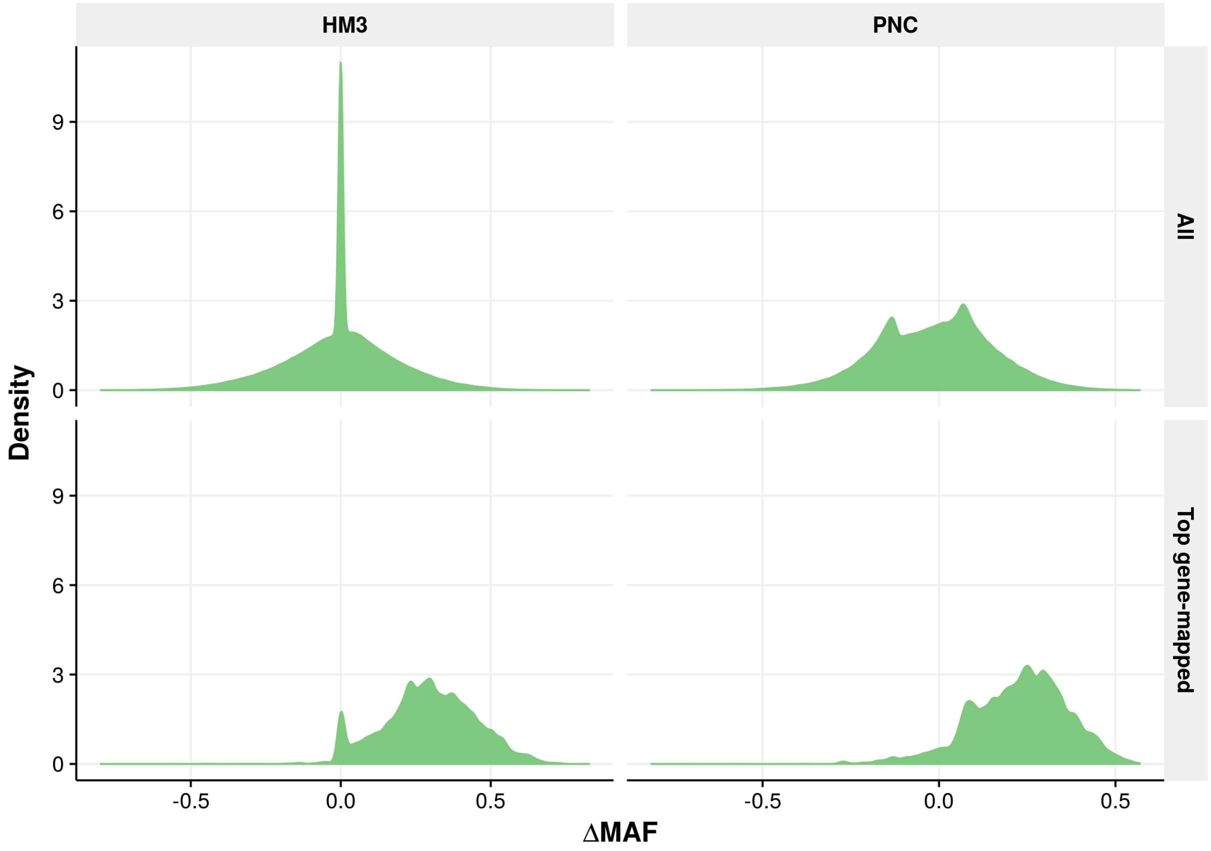


Figure 2. Distribution of the ΔMAF SNP-level test statistic. The population-based difference in minor allele frequency was measured per genotyped SNP from the HM3 (*N* = 1,594,675) and PNC (*N* = 3,730,475) datasets (top panel). The ΔMAF distribution of the single largest gene-mapped SNPs from the HM3 (*N* = 220,064) and PNC (*N* = 208,622) datasets is also shown (bottom panel). The relatively high proportion of 0-value ΔMAF SNPs in the complete HM3 dataset (top left) represents those variants in which a minor allele frequency was not assigned (i.e., NA) in at least one population. This is an artifact that can most likely be attributed to the low sample size in the dataset. The two peaks within the positive and negative hemispheres of the PNC distribution plot (top right) represent concentrated subsets of SNPs with larger ΔMAF values in the European-American and African-American populations, respectively.

Abbreviations: HM3, International HapMap Project; PNC, Philadelphia Neurodevelopmental Cohort; MAF, minor allele frequency.

Table 1. Sample size and number of genotyped SNPs from the HM3 and PNC datasets.

Dataset	Sample size (*N*)		SNPs genotyped (*N*)	Overlap (*N*)	Genotyping rate (%)
	European	African			
HM3	165	83	1,594,675	863,354	87.25
PNC	3,314	1,840	3,730,475		98.51

Abbreviations: HM3; HapMap Phase 3; PNC, Philadelphia Neurodevelopmental Cohort; SNPs, single nucleotide polymorphisms.

Table 2. Summary of GSEA input and results.

Dataset	Pathways (*N*)	Overlap (*N*)	Genes in pathways (*N*)	SNPs in pathways (*N*)	Top SNPs (*N*)	Significant pathways (*N*)	CE (*N*)	nE (*N*)
HM3	3,706	3,464	21,255	862,092	220,064	76	19	18
PNC	3,526		18,978	1,799,870	208,622	237		

Note—the FDR threshold used to determine pathway significance differed between each dataset, at FDR ≤ 0.1 and FDR ≤ 0.05 for the HM3 and PNC datasets, respectively, to account for the difference in sample size. The reported number of SNPs in pathways reflects the sum of unique variants mapped to all tested pathways after filtering out SNPs occurring > 10kb away from its mapped gene. The number of top SNPs reflects the number of unique variants mapped to pathways after selecting the single largest ΔMAF SNP to represent each gene's operative signal.
Abbreviations: HM3; HapMap Phase 3; PNC, Philadelphia Neurodevelopmental Cohort; SNPs, single nucleotide polymorphisms; CE, confidently enriched; nE, nonenriched.

significant enrichment (FDR ≤ 0.05) for ancestral background (**Supplementary Table 2; Table 2**). By explicitly searching for the pathways significantly enriched in both GSEA analyses, I identified a total of 19 biological pathways that demonstrate true enrichment for ancestral background. I provide a list of these pathways and their enrichment statistics in **Tables 3 and 4**, respectively, and I refer to them as the 'confidently enriched' or 'ancestry-enriched' pathways for the remainder of this thesis. To serve as a negative control for downstream analysis, I also defined a set of 18 pathways not enriched for ancestry, or 'nonenriched' pathways, (**Tables 3 and 5**) as a representative set of neutral pathways without enrichment for ancestral background. In other words, this set of pathways contain genes that are not largely differentiated in minor allele frequency between the two studied populations, potentially reflecting conserved biological functionality among European- and African-Americans, rather than conferring selective advantages. The marked variance of GSEA pathway enrichment statistics (particularly the normalized enrichment score (NES), nominal *p* value, and false discovery rate (FDR)) between the sets of enriched and nonenriched pathways is graphically represented in **Figure 3**. Encouragingly, the distribution of the ΔMAF SNP-level test statistic did not vary significantly (*p* = 0.434) across the pathway sets (**Figure 4**), providing evidence for non-artifactual enrichment of the confidently enriched pathways due to an overrepresentation of large ΔMAF SNPs mapped to those pathways.

I next generated a visual overlap-based network of the ancestry-enriched pathways using Enrichment Map [89], a tool that assists in prioritizing related pathways for further downstream exploration. Using this tool, I discovered prominent themes in the set of confidently enriched pathways that are associated with immunity, metabolism, and cell regulation (**Figure 5**). As an additional means of validation, I performed a simple cross-reference analysis of the confidently enriched pathway markers with all significantly identified GWAS SNP-trait associations [21]. Interestingly, of the pathway loci that overlapped with previously defined GWAS hits (~6%), a relatively high number were associated with various metabolic- and immune-related diseases and traits (**Figure 6**). In summary, the original HM3 ~1.5 million SNP dataset was filtered and characterized into a biologically meaningful set of 19 pathways significantly enriched for ancestral population based on SNP-level variation in minor allele frequency between European-American and African-American individuals. These pathways will serve to aid in the interpretation of potential within-pathway signals of epistatic coevolution by providing a common phenotypic context for the statistically interacting SNP-SNP pairs.

Table 3. List of confidently enriched and nonenriched pathways as determined by GSEA.

ID #	Pathway set	
	Confidently enriched	**Nonenriched**
1	Cellular respiration	Activation of gene expression by SREBP
2	Energy derivation by oxidation of organic compounds	Ameboidal-type cell migration
3	Eukaryotic translation termination	COPII mediated vesicle transport
4	Liposaccharide metabolic process	COPI-independent Golgi-to-ER retrograde traffic
5	Lymphocyte activation	Extension of telomeres
6	Lymphocyte differentiation	GABA$_B$ receptor II signaling
7	Macromolecular complex disassembly	Hemopoietic progenitor cell differentiation
8	Morphogenesis of an epithelium	Localization within membrane
9	Nucleoside monophosphate metabolic process	Mesoderm development
10	Nucleoside triphosphate metabolic process	Negative regulation of inflammatory response
11	Protein complex disassembly	Phospholipid transport
12	Regulation of bone mineralization	Platelet aggregation
13	Regulation of cAMP biosynthetic process	Positive regulation of NIK/NF-κB signaling
14	Regulation of cyclic nucleotide metabolic process	Positive regulation of telomere maintenance
15	Regulation of hemopoiesis	Regulation of cell killing
16	Rhythmic process	Regulation of DNA damage response, signal transduction by p53 class mediator
17	Toll-like receptor signaling pathway	Transcriptional regulation of pluripotent stem cells
18	TP53 regulates metabolic genes	Transport of the SLBP-dependant mature mRNA
19	Vacuole organization	

Table 4. Significantly replicated GSEA pathway enrichment statistics.

ID #	Size		ES		NES		NominalP		FDR	
	HM3	PNC	HM3	PNC	HM3	PNC	HM3	PNC	HM3	PNC
1	105	98	0.258	0.284	3.650	4.770	0.000	0.000	0.041	0.001
2	148	135	0.267	0.271	3.812	4.415	0.000	0.000	0.035	0.003
3	75	62	0.117	0.115	2.939	3.622	0.018	0.007	0.100	0.013
4	91	84	0.346	0.349	3.114	3.807	0.004	0.000	0.085	0.010
5	183	173	0.321	0.293	3.652	5.236	0.000	0.000	0.043	0.000
6	99	93	0.350	0.318	3.089	4.146	0.000	0.000	0.088	0.005
7	165	146	0.219	0.225	3.559	3.937	0.001	0.000	0.049	0.007
8	168	144	0.327	0.255	2.984	2.842	0.002	0.004	0.100	0.047
9	169	156	0.236	0.208	3.128	3.055	0.002	0.002	0.083	0.035
10	155	140	0.248	0.219	4.499	3.773	0.000	0.000	0.044	0.011
11	157	138	0.226	0.238	3.742	4.654	0.001	0.000	0.037	0.002
12	46	45	0.458	0.405	2.995	3.058	0.004	0.001	0.098	0.035
13	76	70	0.433	0.383	3.194	2.802	0.001	0.006	0.078	0.050
14	102	90	0.387	0.336	3.517	2.852	0.002	0.004	0.047	0.047
15	183	168	0.304	0.223	3.956	3.411	0.000	0.001	0.029	0.021
16	101	93	0.360	0.314	3.217	3.257	0.002	0.000	0.076	0.026
17	74	68	0.374	0.308	3.293	3.043	0.004	0.001	0.070	0.035
18	73	70	0.363	0.265	4.289	3.227	0.000	0.002	0.045	0.027
19	137	119	0.294	0.323	3.272	4.328	0.001	0.000	0.070	0.003

Note—Each ID number corresponds to the respective ‘Confidently enriched’ pathway ID number in Table 3 (left). FWER statistics were omitted from the table as they are not relevant to this thesis. **Abbreviations**: HM3; HapMap Phase 3; PNC, Philadelphia Neurodevelopmental Cohort; ES, enrichment score; NES, normalized enrichment score; NominalP, nominal *p* value; FDR, false discovery rate.

Table 5. Insignificantly replicated GSEA pathway enrichment statistics.

ID #	**Size**		**ES**		**NES**		**NominalP**		**FDR**	
	HM3	**PNC**	**HM3**	**PNC**	**HM3**	**PNC**	**HM3**	**PNC**	**HM3**	**PNC**
1	39	37	0.280	0.261	-0.022	0.005	0.507	0.470	0.742	0.643
2	92	86	0.293	0.245	0.065	-0.014	0.456	0.499	0.721	0.650
3	64	63	0.234	0.217	0.047	0.065	0.475	0.455	0.725	0.620
4	28	28	0.356	0.345	-0.066	-0.024	0.521	0.494	0.755	0.654
5	24	22	0.280	0.186	-0.048	0.043	0.524	0.627	0.750	0.630
6	33	32	0.385	0.291	-0.025	0.096	0.492	0.434	0.742	0.611
7	25	23	0.227	0.190	-0.022	-0.002	0.565	0.628	0.742	0.646
8	94	86	0.270	0.242	0.041	0.085	0.481	0.459	0.727	0.615
9	48	41	0.238	0.200	0.001	0.030	0.495	0.509	0.734	0.635
10	47	36	0.196	0.166	0.066	0.018	0.456	0.546	0.721	0.640
11	48	40	0.365	0.337	-0.006	-0.090	0.500	0.527	0.736	0.673
12	29	25	0.238	0.187	-0.047	-0.058	0.539	0.637	0.750	0.665
13	20	20	0.302	0.208	-0.093	0.032	0.531	0.604	0.761	0.634
14	42	37	0.249	0.271	0.012	0.026	0.470	0.474	0.730	0.636
15	45	42	0.251	0.205	0.019	-0.004	0.468	0.491	0.730	0.647
16	24	24	0.307	0.258	0.003	-0.007	0.482	0.501	0.734	0.647
17	31	25	0.204	0.205	0.062	0.014	0.533	0.596	0.721	0.641
18	28	26	0.244	0.317	-0.073	-0.065	0.562	0.516	0.756	0.667

Note—Each ID number corresponds to the respective 'Nonenriched' pathway ID number in Table 3 (right). FWER statistics were omitted from the table as they are not relevant to this thesis. **Abbreviations**: HM3; HapMap Phase 3; PNC, Philadelphia Neurodevelopmental Cohort; ES, enrichment score; NES, normalized enrichment score; NominalP, nominal *p* value; FDR, false discovery rate.

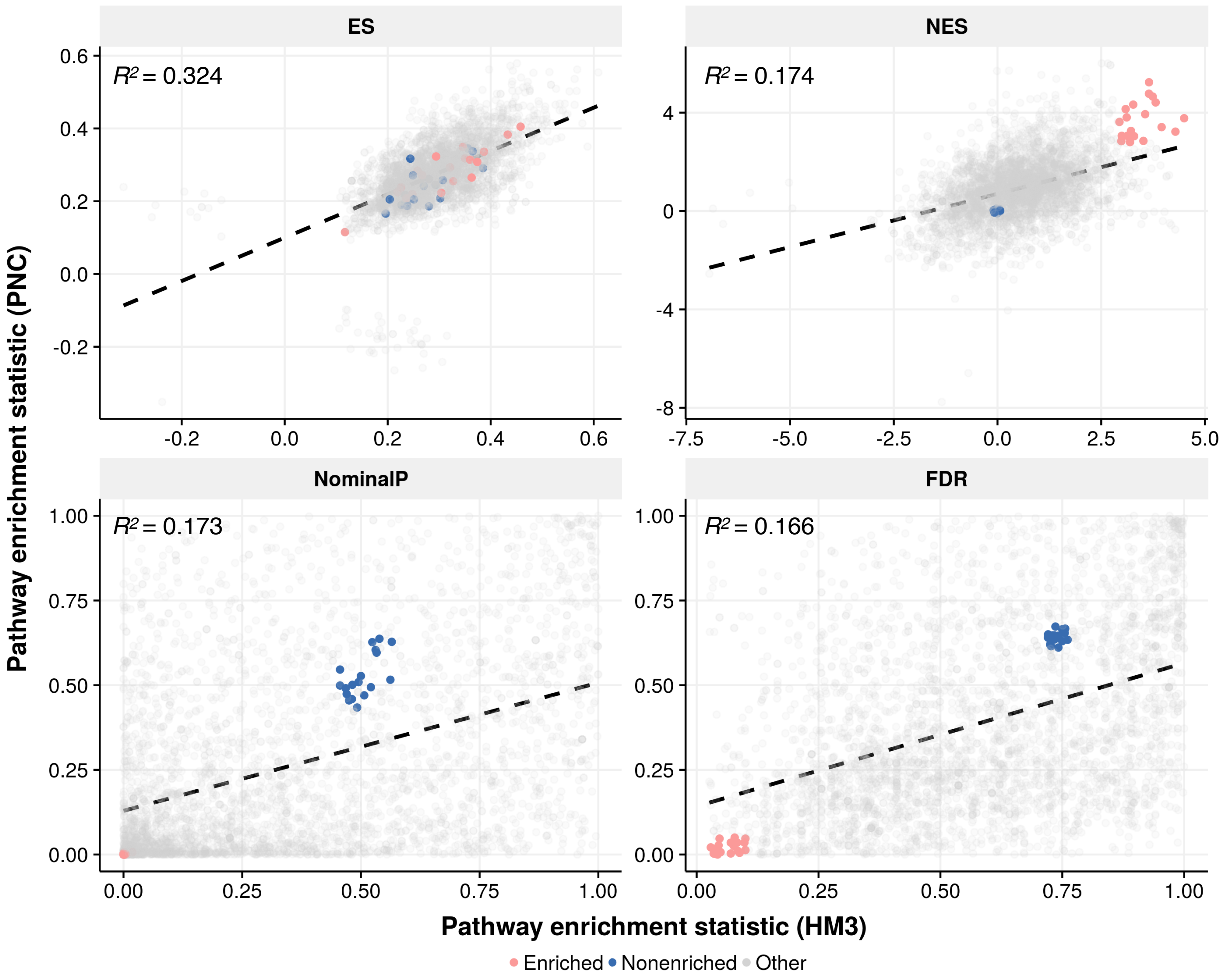


Figure 3. Replication of GSEA pathway enrichment statistics. Each scatterplot represents the concordance of the given enrichment statistic calculated per replicated pathway (*N* = 3,464 total points per plot). The black dashed line in each plot represents the correlation of the enrichment statistic between both dataset analyses. The strength of correlation is indicated by the R^2 regression coefficient. The confidently enriched pathways (*N* = 19; pink) are characterized by a largely positive NES, low nominal *p* value, and low FDR. Conversely, the nonenriched pathways (*N* = 18; blue) are characterized by a 0-value NES, and a relatively high nominal *p* value and FDR. The lack of correlation between the nominal *p* value and FDR statistic can be explained by the relative difference in sample size between the HM3 and PNC datasets. This could effectively generate a lower pathway FDR in one dataset and not the other. **Abbreviations**: HM3; HapMap Phase 3; PNC, Philadelphia Neurodevelopmental Cohort; ES, enrichment score; NES, normalized enrichment score; NominalP, nominal *P* value; FDR, false discovery rate.

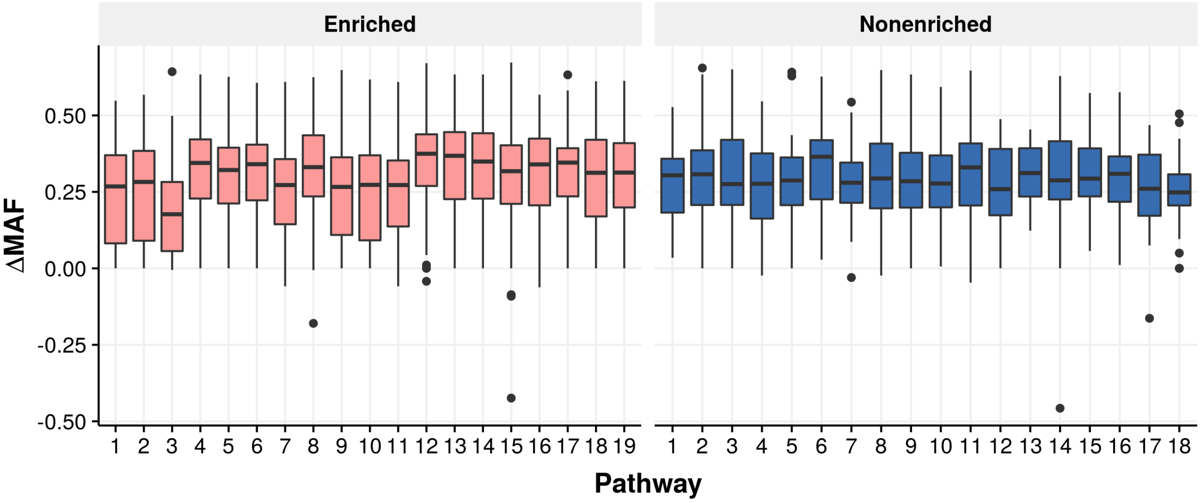


Figure 4. Distribution of the ΔMAF SNP-level test statistic per confidently enriched and nonenriched pathway. Each box plot represents the total ΔMAF distribution of all SNPs mapped to a pathway from the HM3 dataset. The mean ΔMAF of the complete set of confidently enriched (0.285 ± 0.157 s.d.) and nonenriched (0.290 ± 0.146 s.d.) pathways does not differ significantly (*p* = 0.434). The *p* value was calculated using the two-sided Student's t-Test. **Abbreviations**: MAF, minor allele frequency; s.d., standard deviation.

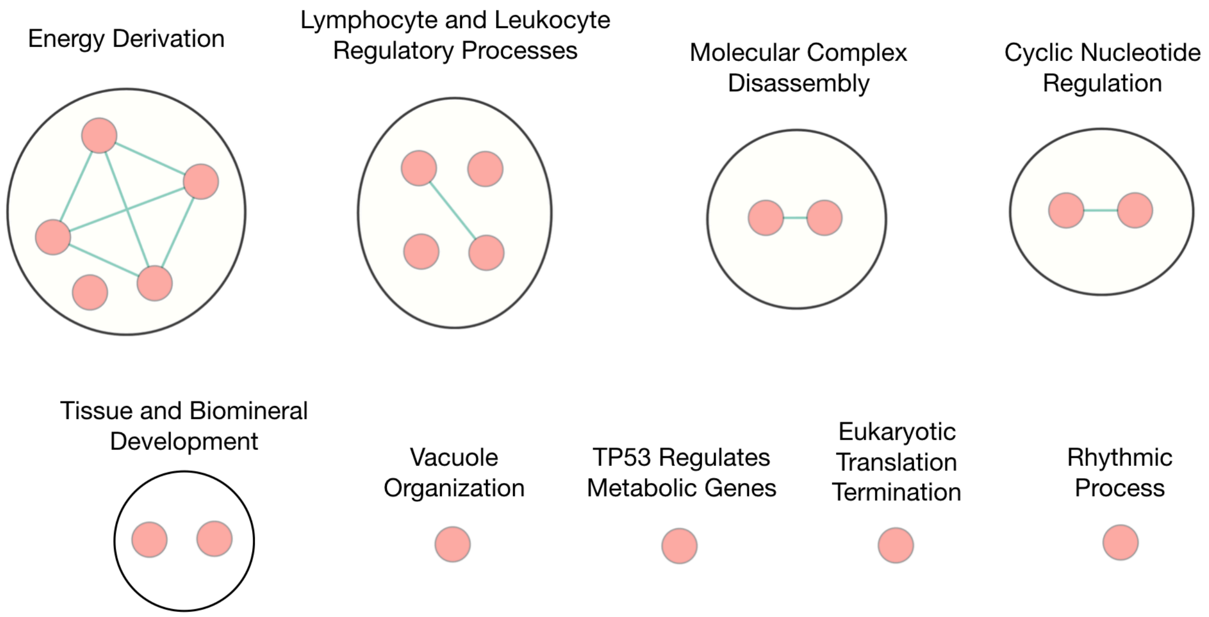

Energy Derivation
Lymphocyte and Leukocyte Regulatory Processes
Molecular Complex Disassembly
Cyclic Nucleotide Regulation
Tissue and Biomineral Development
Vacuole Organization
TP53 Regulates Metabolic Genes
Eukaryotic Translation Termination
Rhythmic Process

Figure 5. Pathway enrichment map of the ancestry-enriched pathways. The set of confidently enriched pathways (*N* = 19 pink nodes) are visually summarized within thematically associated subnetworks using an overlap-based network algorithm, in which pathway nodes are laid out using a force-directed layout. Subnetworks (annotated circles) are annotated according to the shared function of the pathways within that cluster. The size of each node corresponds to the size of the given pathway. Green edges connect pathways with at least 0.375 gene member overlap.

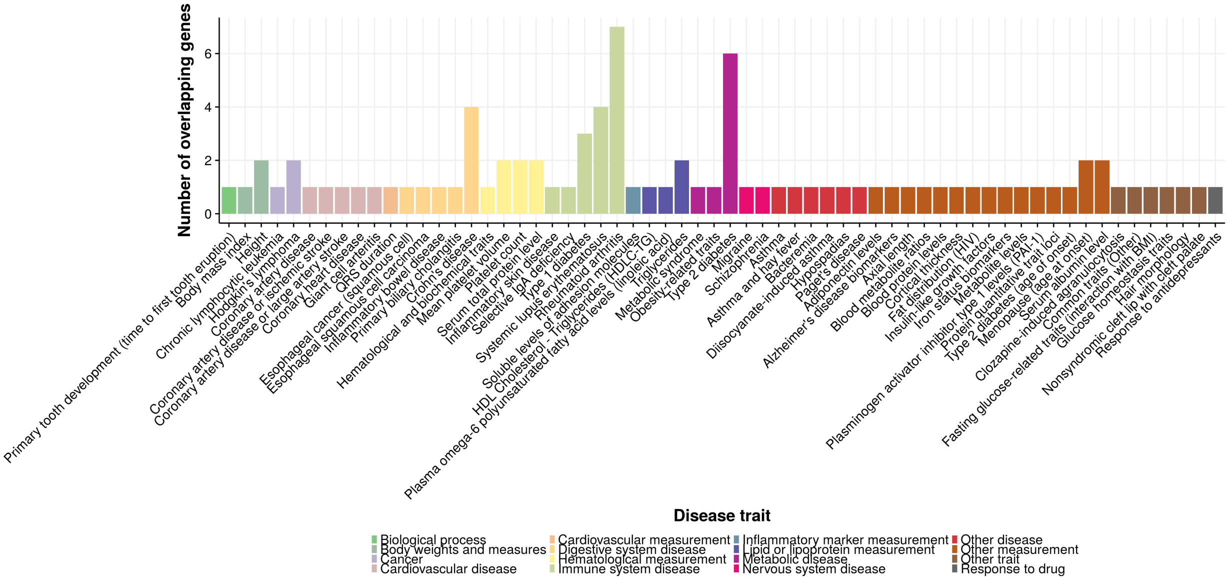


Figure 6. Cross-reference of the confidently enriched pathway markers with known GWAS SNP-trait associations. Each bar represents the number of overlapping ancestry-enriched SNPs (*N* = 89 total) per GWAS trait, and are thematically grouped to 16 different categories according to the NHGRI-EBI GWAS Catalog. The GWAS SNP-trait associations (*N* = 50,086) were taken from the NHGRI-EBI Catalog database, version 1.0.1 (available for download online from https://www.ebi.ac.uk/gwas/docs/file-downloads, accessed September 26, 2017). To note, multifactorial traits and diseases are only assigned to a single category. For example, Crohn's disease is classified as a disease of the digestive system but may also fall under the category of an immune system disease. Each ancestry-associated pathway contains at least one SNP with a GWAS-derived SNP-trait association.

3.2. Immune response pathways demonstrate significant signals of coevolution in African-Americans

I measured inter-chromosomal pairwise linkage disequilibrium between pathway-level SNP-SNP associations to infer signals of within-pathway epistatic coevolution. Though I use the term LD to describe these associations, I want to reiterate that they are not a result of linkage between the loci. Instances of allelic association between physically unlinked loci could potentially reflect other population genetic factors unrelated to epistatic coevolution; however, until further validated, I will refer to any identified signal as within-pathway coevolution. For example, one possible contributor to is the difference of genotypic structure between the

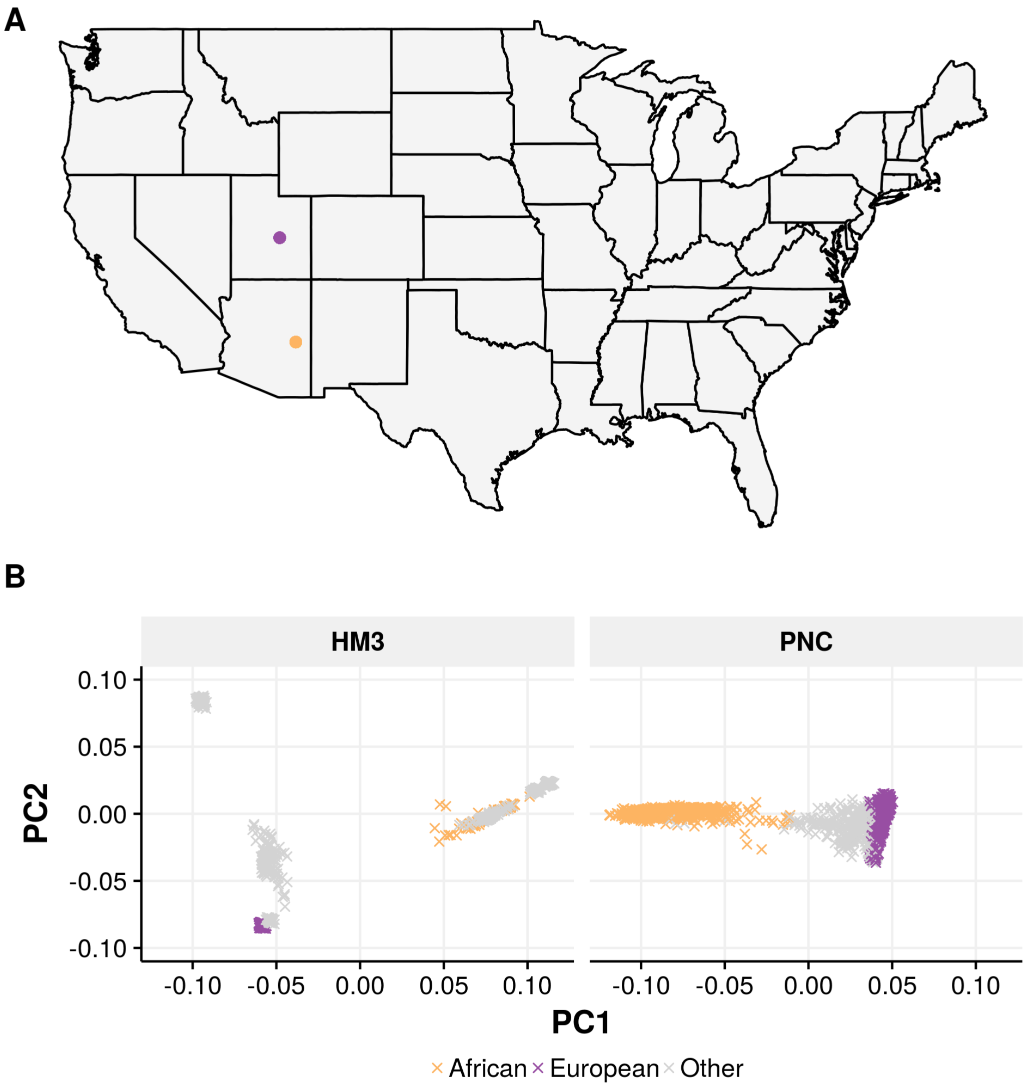


Figure 7. Geographic location and genotype stratification of the European-American and African-American study samples. Of the two ancestral samples studied (**A**), principal component analysis (**B**) of the HM3 and PNC data visually stratifies the genotypes of the 83 African-American ancestry individuals (yellow) and 165 European-American individuals (purple) from the HM3 dataset. Ethnicity of the PNC samples was defined as groups that lie within ±5 standard deviations of the PC1, PC2 centroid of the corresponding HM3 samples, resulting in 1,840 African-American and 3,314 and European-American samples. **Abbreviations**: HM3; HapMap Phase 3; PNC, Philadelphia Neurodevelopmental Cohort; PC, principal component.

European- and African-American populations (**Figure 7**). As mentioned previously in the Introduction, forces affecting population demography (e.g., bottlenecks, colonization) can effectively generate disequilibria. The particular African population (ASW) I studied was sampled from the United States, and as such, genetically resembles a widely diverse mixture of African populations. The majority of African Americans are descended from the ~600,000 Africans brought to British North America during the Atlantic slave trade [90, 91], and were deported primarily from particular geographic regions of western Africa, although more central and eastern locations have also contributed [92, 93]. The European population (CEU) was also sampled from the United States, but more specifically consists of individuals with northern and western European ancestry. This is a population known to have developed subsequent to the out-of-Africa migration [94], a bottleneck which effectively resulted in decreased genetic diversity within this population, corresponding to their increased distance from Africa [95]. Overall, there is much historical and epidemiological evidence to suggest substantial heterogeneity in the genetic composition between individuals of European-American and African-American ancestry, which could play a role in the generation of disequilibria within each respective population.

I first determined all intra- and inter-chromosomal SNP-SNP interaction pairs within the complete sets of confidently enriched (N = 150,630 total pairs) and nonenriched (N = 19,565 total pairs) pathways, filtered for the inter-chromosomal pairs among both sets of pathways (confidently enriched, N = 141,999; nonenriched, N =18,437; **Table 6**), and then measured the strength of association (R^2) per inter-chromosomal SNP-SNP pair. I determined the overall strength of coevolution within the complete set of confidently enriched pathways by measuring whether the cumulative distribution of R^2 correlation within the enriched pathways differed significantly from that of the nonenriched pathways (see Materials and Methods). I reasoned that rather than generating a null R^2 distribution of 'pseudo' pathways as my negative control, by randomly grouping genes together without any functional or biological association, it would be more informative to identify a significant shift in R^2 distribution between two functionally distinct pathway sets (i.e., enriched and not enriched for ancestry). This approach should facilitate discovery of confidently enriched pathways of true biological relevance to ancestral background, as opposed to determining the significance of pathway classification in general. Upon comparing the R^2 distribution shift between the cumulative sets of enriched and

Table 6. Total number of within-pathway pairwise SNP-SNP interactions per complete set of enriched and nonenriched pathways.

Pathway set	**All samples**			**European only**			**African only**		
	Total (*N*)	**Inter (*N*)**	**Prop (%)**	**Total (*N*)**	**Inter (*N*)**	**Prop (%)**	**Total (*N*)**	**Inter (*N*)**	**Prop (%)**
Enriched	150,630	141,999	94.27	144,044	135,780	94.26	146,973	138,557	94.27
Nonenriched	19,565	18,437	94.23	18,745	17,673	94.28	19,291	18,175	94.21

Abbreviations: Inter, inter-chromosomal; Prop, proportion.

nonenriched pathways, I did not determine any significant evidence for coevolution ($p = 0.54$) within the entire set of confidently enriched pathways (**Figure 8**). It is highly plausible, however, that the cumulative R^2 distribution could be negatively skewed by pathways containing variants with weak inter-chromosomal pairwise LD association.

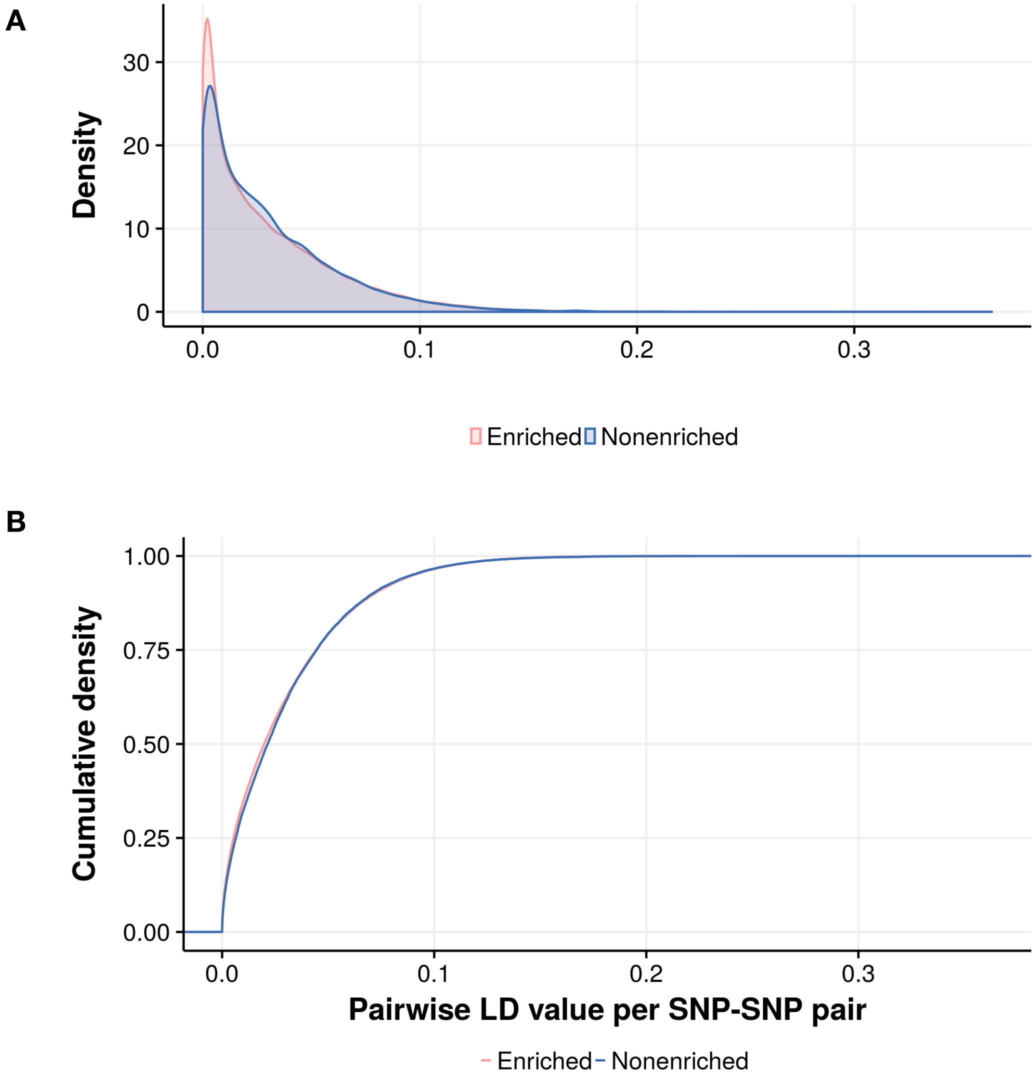


Figure 8. Degree of coevolution signal within the confidently enriched pathways. Density (**A**) and empirical cumulative distribution function (eCDF; **B**) plots graphically depict the strength of association (R^2) between all inter-chromosomal pairwise SNP-SNP interactions within the confidently enriched (pink; $N = 141{,}999$ interactions) and nonenriched (blue; $N = 18{,}437$ interactions) pathways. The strength of coevolution signal within the entire set of confidently enriched pathways is insignificant ($p = 0.54$) at this level. The p value was calculated using the two-sample one-sided Kolmogorov-Smirnov (KS) test. **Abbreviations**: LD, linkage disequilibrium; SNP, single nucleotide polymorphism.

Table 7. Number of inter-chromosomal pairwise interactions per confidently enriched pathway.

Pathway name	**All samples (*N*)**	**European (*N*)**	**African (*N*)**
Cellular respiration	4,766	4,577	4,677
Energy derivation by oxidation of organic compounds	9,581	9,177	9,320
Eukaryotic translation termination	2,438	2,134	2,185
Liposaccharide metabolic process	3,811	3,719	3,726
Lymphocyte activation	15,087	14,417	14,754
Lymphocyte differentiation	4,433	4,257	4,428
Macromolecular complex disassembly	12,579	11,954	12,430
Morphogenesis of an epithelium	13,285	12,806	13,132
Nucleoside monophosphate metabolic process	12,602	11,991	12,168
Nucleoside triphosphate metabolic process	10,388	9,677	9,998
Protein complex disassembly	11,368	10,775	11,228
Regulation of bone mineralization	932	932	932
Regulation of cAMP biosynthetic process	2,644	2,364	2,645
Regulation of cyclic nucleotide metabolic process	4,700	4,328	4,700
Regulation of hemopoiesis	15,168	15,004	14,655
Rhythmic process	4,541	4,449	4,361
Toll-like receptor signaling pathway	2,549	2,548	2,408
TP53 regulates metabolic genes	2,420	2,351	2,353
Vacuole organization	8,707	8,320	8,456

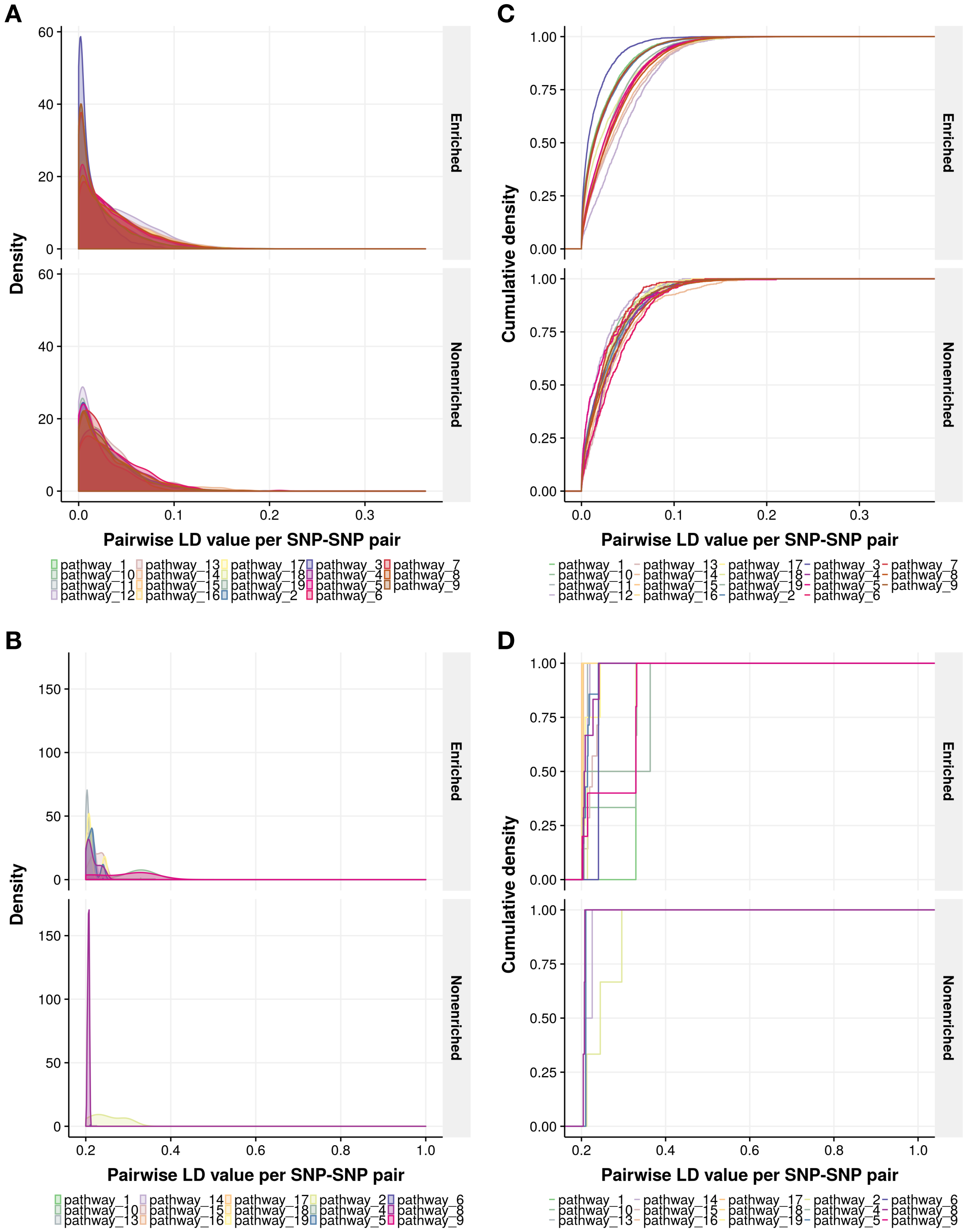


Figure 9. Degree of within-pathway signals of coevolution. Density (**A**, **B**) and eCDF (**C**, **D**) plots graphically depict the strength of association (R^2) between all inter-chromosomal pairwise SNP-SNP interactions within each confidently enriched and nonenriched pathway. The x-axes of plots **B** and **D** (representative of plots **A** and **C**, respectively) are truncated at $R^2 \geq 0.2$ to

demonstrate the pathways with longer tails of inter-chromosomal pairwise SNP-SNP association.

Abbreviations: LD, linkage disequilibrium; SNP, single nucleotide polymorphism.

Table 8. Significance of within-pathway coevolution signal per confidently enriched pathway.

Pathway name	***p* value**
Cellular respiration	1
Energy derivation by oxidation of organic compounds	1
Eukaryotic translation termination	1
Liposaccharide metabolic process	7.21×10^{-18}
Lymphocyte activation	4.02×10^{-15}
Lymphocyte differentiation	3.47×10^{-21}
Macromolecular complex disassembly	1
Morphogenesis of an epithelium	1.14×10^{-53}
Nucleoside monophosphate metabolic process	0.999
Nucleoside triphosphate metabolic process	1
Protein complex disassembly	1
Regulation of bone mineralization	1.74×10^{-42}
Regulation of cAMP biosynthetic process	3.73×10^{-43}
Regulation of cyclic nucleotide metabolic process	7.23×10^{-49}
Regulation of hemopoiesis	9.59×10^{-24}
Rhythmic process	5.65×10^{-19}
Toll-like receptor signaling pathway	3.48×10^{-13}
TP53 regulates metabolic genes	0.067
Vacuole organization	2.09×10^{-9}

Note—*p* values were calculated using the two-sample one-sided KS test. Nominal significance was determined as $p < 0.003$ (Bonferroni corrected).

Given these observations, I next assessed the strength of coevolution within each confidently enriched pathway independently. To achieve this for a given pathway, I first measured the strength of association per inter-chromosomal pairwise SNP-SNP interaction (**Table 7 and Figure 9**), followed by the statistical comparison of the cumulative R^2 correlation distribution against the entire set of nonenriched pathways (see Materials and Methods). Using this approach, I identified 11 out of the 19 ancestry-enriched pathways as having a nominally significant ($p < 0.003$) signal of within-pathway coevolution (**Table 8**). These pathways are: *Liposaccharide metabolic process* ($p = 7.21\times10^{-18}$), *Lymphocyte activation* ($p = 4.02\times10^{-15}$), *Lymphocyte differentiation* ($p = 3.47\times10^{-21}$), *Morphogenesis of an epithelium* ($p = 1.14\times10^{-53}$), *Regulation of bone mineralization* ($p = 1.74\times10^{-42}$), *Regulation of cAMP biosynthetic process* ($p = 3.73\times10^{-43}$), *Regulation of cyclic nucleotide metabolic process* ($p = 7.23\times10^{-49}$), *Regulation of hemopoiesis* ($p = 9.59\times10^{-24}$), *Rhythmic process* ($p = 5.65\times10^{-19}$), *Toll-like receptor signaling pathway* ($p = 3.48\times10^{-13}$), and *Vacuole organization* ($p = 2.09\times10^{-9}$). Examples of the top fifteen inter-chromosomal pairwise interactions among all tested pathways are shown in **Table 9**.

Overall, the primary goal of my thesis project was to ask if there were significant signals of evolutionary selection acting within biological pathways in a population-specific manner. As a result, I repeated the same analysis within the 11 pathways identified above among the 165 European-American and 83 African-American individuals independently (**Figure 10**). This was accomplished by first categorizing SNP genotypes based on ancestral group and then measuring pathway-level inter-chromosomal pairwise SNP-SNP association among each set of population-specific genotypes separately. Interestingly, only the *Regulation of hemopoiesis* ($p = 4.74\times10^{-4}$) and *Toll-like receptor signaling pathway* ($p = 5.30\times10^{-4}$) pathways remained nominally significant ($p < 0.005$) in the African-American ancestry group after population-based stratification (**Table 10**). These two pathways are distinct in function, with only 3% overall gene member similarity (**Table 11**), but each have marked roles in the human immune response. As described in AmiGO v.2.4.26, the *Regulation of hemopoiesis* pathway broadly constitutes "any process that modulates the frequency, rate or extent of hemopoiesis" [96], a process that is primarily involved with the development of the immune system. Hemopoiesis (or hematopoiesis) is defined as the ongoing production of blood cells and platelets throughout adulthood, starting from embryonic development, to maintain the circulatory system [97]. By using model organisms such as zebrafish [98], mouse [99], as well as humans [100], researchers have begun to elucidate the mechanisms underlying hematopoietic stem cell development, an understanding that has important implications in the field of regenerative medicine [97]. Additionally, the *Toll-like receptor signaling pathway* pathway is involved with the innate immune activation response, broadly constituting "any series of molecular signals generated as a consequence of binding to a toll-like receptor (TLR), [which] directly bind pattern motifs from a variety of microbial sources to initiate innate immune response" [96]. In contrast to the adaptive immune system, the innate immune system represents the non-specific primary defense mechanisms used to fight against pathogenic infection, and is phylogenetically conserved across almost all multicellular organisms [101]. In *Drosophila*, the Toll signaling pathway is crucial for the response to fungal infection [102] whereas, in mammals, TLR-originated pathways are mostly responsible for host defense against microorganisms [103] and thus have important roles in various immune disorders and infectious diseases [104]. In the Discussion, I discuss in more depth the relevance and implications of these pathway findings in regards to adaptive evolution among the studied populations, as well as the implications of discovering population-driven coevolving genetic interactions within those pathways.

In order to provide a visual representation of the significant pathway-level genetic crosstalk identified via the allelic LD association analysis, **Figures 11 and 12** graphically depict the genome-wide distribution of the within-pathway inter-chromosomal pairwise SNP-SNP interactions among gene members of the *Regulation of hemopoiesis* and *Toll-like receptor signaling pathway* pathways, respectively, in 2D (generated via the RCircos package [105]). Additionally, examples of the top fifteen inter-chromosomal interactions among all tested pathways are shown in **Table 12**, which effectively depict the variation in the strength of SNP-SNP inter-chromosomal association between individuals of European- and African-American ancestry.

Table 9. Pairwise R^2 correlation value of the top within-pathway inter-chromosomal interactions.

Genetic interaction pair						$\boldsymbol{R^2}$	**ID #**
Chr	**SNP**	**Gene**	**Chr**	**SNP**	**Gene**		
4	rs28476740	*PIGY*	1	rs34676516	*PIGV*	0.363	4
17	rs34165301	*ALDOC*	1	rs12135218	*UQCRH*	0.331	9
17	rs34165301	*ALDOC*	1	rs12135218	*UQCRH*	0.331	10
16	rs2967157	*UQCRC2*	10	rs12259919	*NDUFB8*	0.329	1
16	rs2967157	*UQCRC2*	10	rs12259919	*NDUFB8*	0.329	2
10	rs12259919	*NDUFB8*	3	rs11706052	*IMPDH2*	0.329	9
16	rs2967157	*UQCRC2*	10	rs12259919	*NDUFB8*	0.329	9
16	rs2967157	*UQCRC2*	10	rs12259919	*NDUFB8*	0.329	10
5	rs40680	*SEMA5A*	2	rs11691947	*SEMA4F*	0.296	2*
3	rs2713616	*PLXND1*	2	rs11691947	*SEMA4F*	0.245	2*
11	rs1794072	*SYT7*	7	rs2699803	*SNX10*	0.244	19
14	rs2301113	*HIF1A*	3	rs1126478	*LTF*	0.243	15
8	rs2409805	*GATA4*	5	rs6580257	*FGF1*	0.242	8
5	rs2227282	*IL4*	1	rs1057079	*MTOR*	0.241	15
16	rs12444401	*PLCG2*	13	rs12428172	*FLT3*	0.240	5

Note—Each ID number corresponds to the respective ‘Confidently enriched’ pathway ID in Table 3. As a result of the inherent similarity between a number of the ancestry-enriched metabolic pathways, for example the *Nucleoside monophosphate metabolic process* (#9) and *Nucleoside triphosphate metabolic process* (#10) pathways, the interacting loci are identical. This pathway similarity is captured by the enrichment map shown in **Figure 5**. The pathway marked by an asterisk (*) represents the nonenriched *Ameboidal-type cell migration* pathway (#2). **Abbreviations**: Chr, chromosome; SNP; single nucleotide polymorphism.

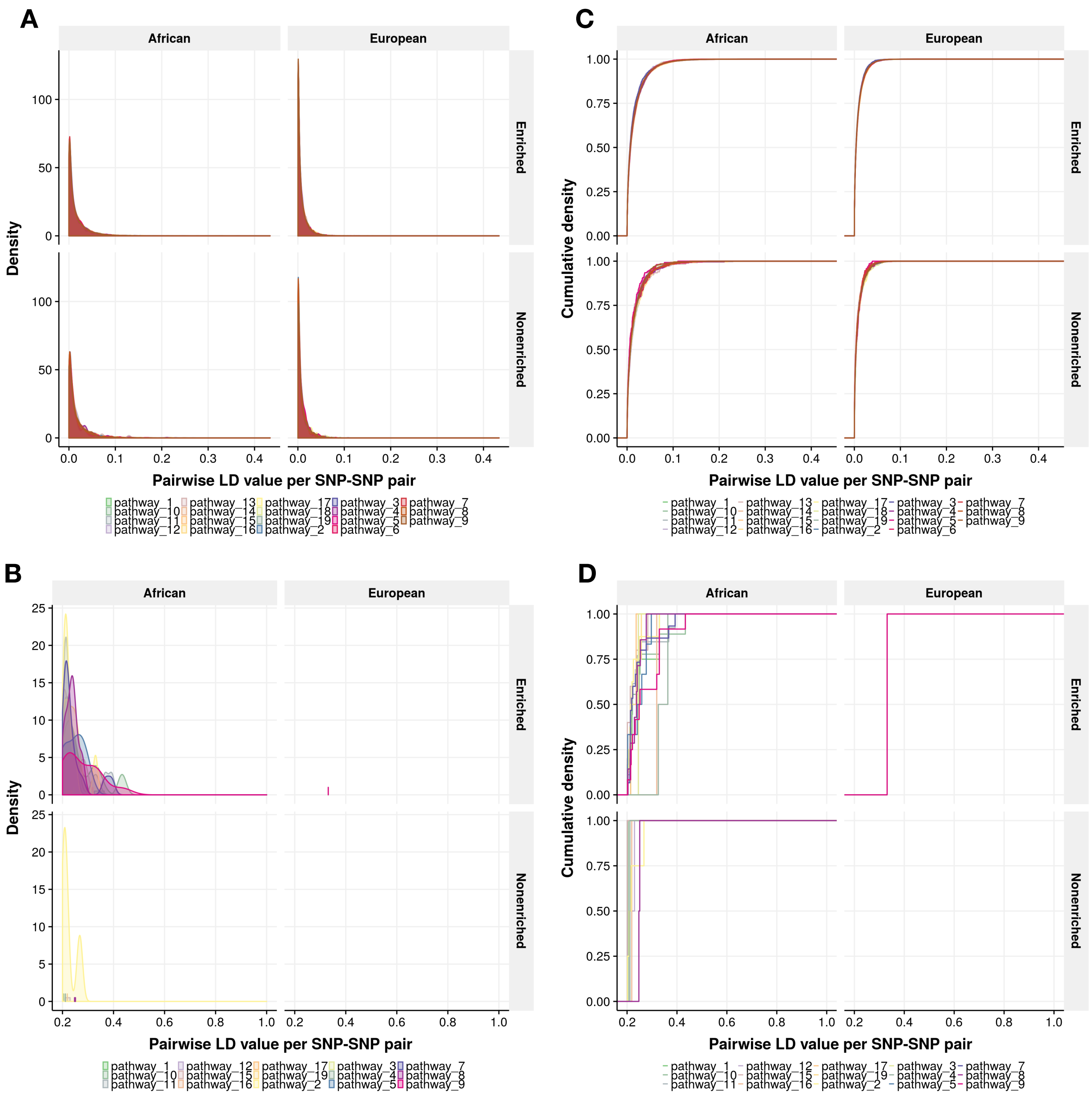


Figure 10. Degree of ancestry-specific within-pathway signals of coevolution. Density (**A**, **B**) and eCDF (**C**, **D**) plots graphically depict the strength of association (R^2) between all inter-chromosomal pairwise SNP-SNP interactions within each confidently enriched and nonenriched pathway across the European- and African-Americans independently. The x-axes of plots **B** and **D** (representative of plots **A** and **C**, respectively) are truncated at $R^2 \geq 0.2$ to demonstrate the pathways with longer tails of inter-chromosomal pairwise SNP-SNP association. **Abbreviations**: LD, linkage disequilibrium; SNP, single nucleotide polymorphism.

Table 10. Significance of ancestry-specific within-pathway coevolution signal per previously identified coevolving pathway.

Pathway name	***p* value**	
	European	**African**
Liposaccharide metabolic process	0.077	0.009
Lymphocyte activation	0.078	0.007
Lymphocyte differentiation	0.469	0.122
Morphogenesis of an epithelium	0.406	0.251
Regulation of bone mineralization	0.280	0.987
Regulation of cAMP biosynthetic process	0.106	0.493
Regulation of cyclic nucleotide metabolic process	0.316	0.452
Regulation of hemopoiesis	0.203	4.74×10^{-4}
Rhythmic process	0.738	0.650
Toll-like receptor signaling pathway	0.051	5.30×10^{-4}
Vacuole organization	0.307	0.019

Note—*p* values were calculated using the two-sample one-sided KS test. Nominal significance was determined as $p < 0.005$ (Bonferroni corrected).

Table 11. Gene members of the *Regulation of hemopoiesis* and *Toll-like receptor signaling pathway* pathways.

Pathway name	Gene members
Regulation of hemopoiesis	HIF1A*, MTOR, CDK6, LTF*, PRKCZ, IL4*, PRDM16*, RUNX1, RC3H1, FOXO3*, DCSTAMP, ZBTB16, FNIP1*, KIAA0922, FOXP1*, PRKCA*, RARA, VNN1*, NLRP3, C1QC, PRMT1, HES1, GPR55, GLI2*, IL34, IL7, MEF2C, LILRB3, GAS6, SPI1*, CAMK4*, TAL1, NOTCH1, MEIS1, FLCN*, TMEM178A, CTNNBIP1, HIST1H4D*, TESPA1, TESC, GLI3*, MYC, MEIS2, AXL, ACE*, ACVR2A*, IHH, SPINK5, ACVR1B, LEF1, KMT2E, IL17A*, **TRAF6**, **CASP8**, TGFBR2, IRF4*, CDKN2A*, CSF1, BMP4, IFNA2, CD86*, GATA3, FOXC1, FSHR, PTPN2, CD80, ZAP70*, **TLR4**, PTK2B*, THPO, APCS, HES5, MYB, LYN*, LGALS3, TNFSF4, CARTPT*, SYK, ZC3H8, HIST1H4H, ZFPM1, FSTL3, IL12B, ANXA1, **LGALS9**, ARNT, ETS1, PURB, L3MBTL1, ZNF16, HMGB1, N4BP2L2, PDCD2, SCIN, TNFSF11*, ZNF675, MAPK14*, CD2, LDB1, CLPTM1, MMP14, SFRP1, NCKAP1L, **TLR3**, SOCS5, IL23R, FAM213A, SART1, FES, JAK3, IL12RB1, IRF1*, HLA-B, **FADD**, LIF*, ZBTB1, CD46, CCR1*, IL3*, SHH, HIST1H4C*, LEO1*, TNFSF9, CDC73*, CTLA4, IFNG*, LILRB1, OCSTAMP, CIB1, HCLS1*, CALCA, CTR9*, XBP1, EIF2AK2, HAX1, HIST1H4A*, HIST1H4B*, **RIPK1***, MAPK11, HIST1H4F*, IL27, WDR61, CSF3, IKZF3, LILRB2*, MAFB, AGER, HOXA7, CCL3, LILRB4, HLA-G*, CCL19, ACIN1, HIST1H4J*, FOXJ1, IFNB1, INHBA*, LGALS1, PNP, INHA, ADIPOQ, TNF*, MED1, SOD1*, GAS2L1, MIXL1, HIST4H4, HOXB8, IL23A, HMGB2, PF4*, IFNL1, ADAM8*, HIST1H4E*, HIST1H4L*, HIST1H4K*, PRELID1, ANKRD54, **IRF**7, CEBPB, PPP2R3C, PAF1*, HOXA9
Toll-like receptor signaling pathway	TRAF3, TLR1*, TLR10, UBE2D2, PIK3R4, UBC, IRF3, UNC93B1, CD180*, TAB2, PRKCE, TNIP1*, PIK3C3, PIK3AP1, CNPY3, MAP3K1, TBK1, CD36*, **TRAF6**, TLR6, **CASP8**, UBA52, NR1H4*, RFTN1*, LGMN, TRIL, TANK*, TNIP3, IRAK3, IKBKE, **TLR4**, CHUK*, TNIP2, UBE2D1, UBE2D3, CTSK, COLEC12, CTSS*, ITGB2, **LGALS9**, CTSL, IRAK4*, RIPK2, SCARA3, HSP90B1*, TICAM1*, S100A14, CTSB*, IRAK2*, **TLR3**, TLR5*, ITGAM*, **FADD**, LBP, TAB1, IKBKB*, CD14, HSPD1*, TLR2, MAP3K7, **RIPK1***, LY96, REG3G, BIRC3, TIRAP, MAPKAPK2, MAPKAPK3*, TLR9, BCL10, BIRC2, MYD88, RPS27A, UBB, **IRF7**

Note—Bolded genes represent those annotated to both pathways. Genes followed by an asterisk (*) represent those with evidence for recent positive selection (see **Supplementary Tables 4 and 5** for additional details).

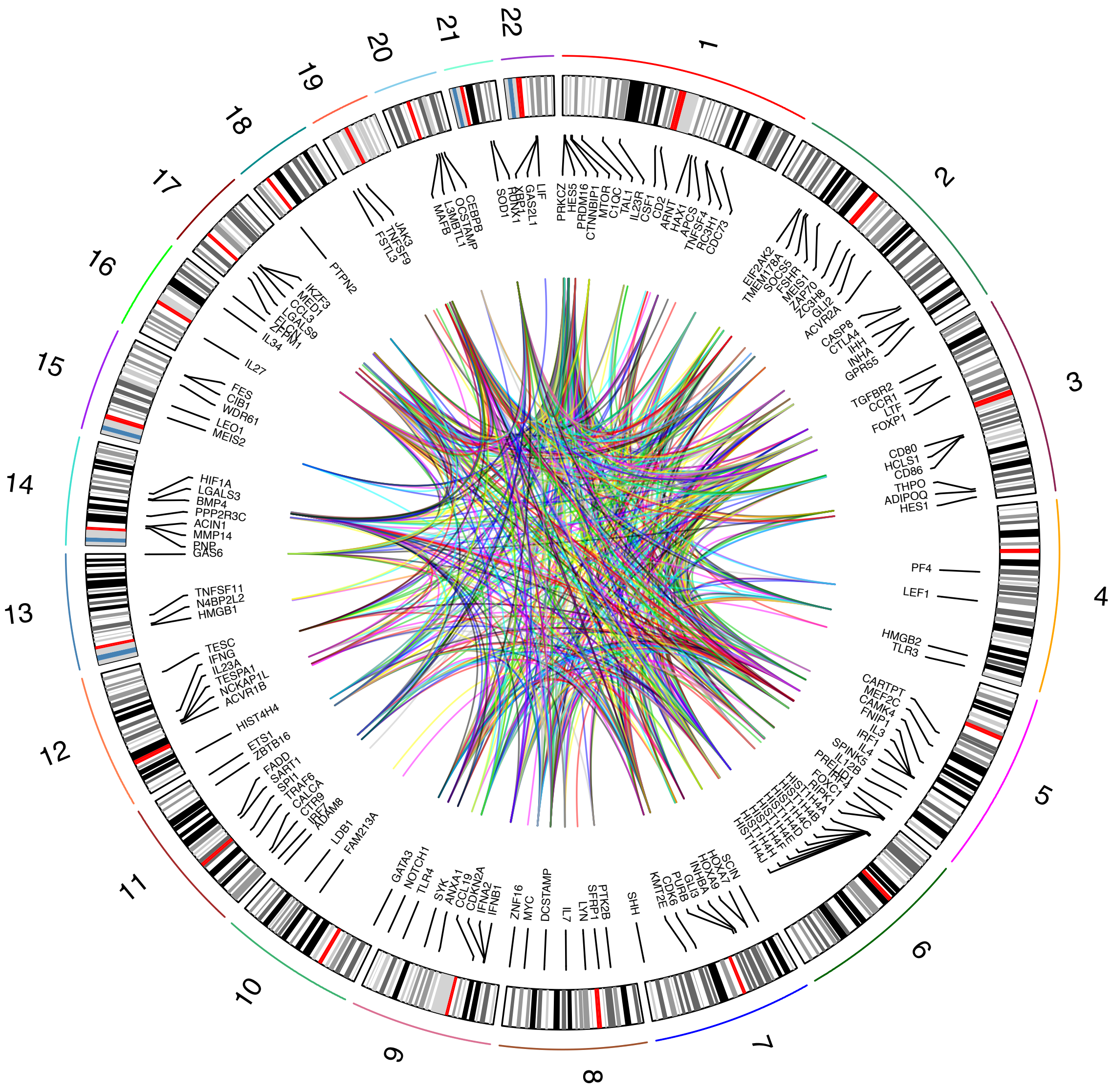


Figure 11. Graphical representation of inter-chromosomal pairwise genetic interactions within the *Regulation of hemopoiesis* pathway. The linked circos plot depicts the genome-wide spread of inter-chromosomal pairwise interactions found between the gene members ($N = 183$) of the *Regulation of hemopoiesis* pathway. Each coloured line represents a statistical interaction between a pair of physically unlinked loci. Only those interactions with a linkage disequilibrium correlation coefficient of $R^2 \geq 0.1$ are shown ($N = 747$ from 15,168 total interactions) to highlight the stronger inter-chromosomal interaction pairs within the pathway.

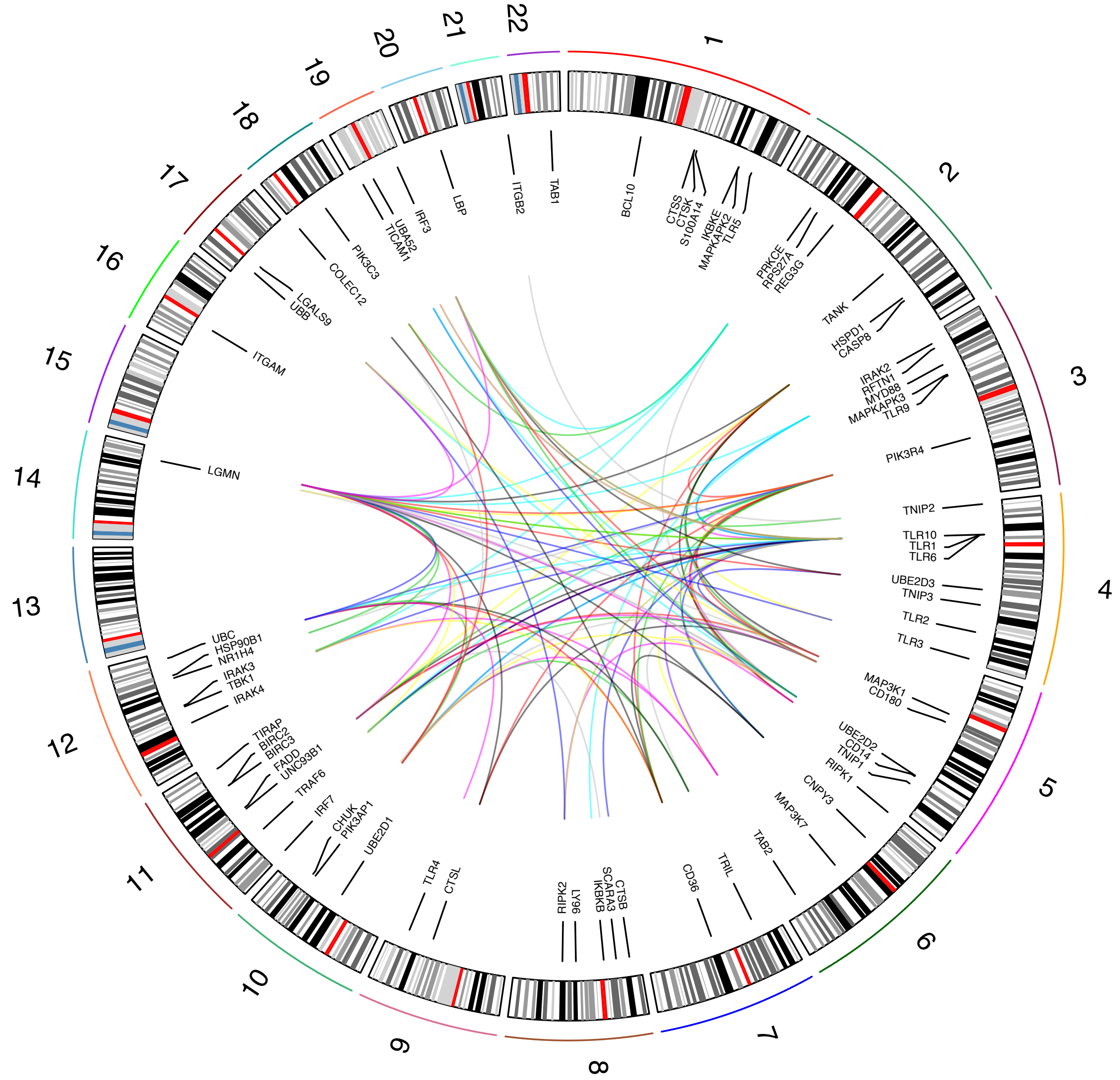


Figure 12. Graphical representation of inter-chromosomal pairwise genetic interactions within the *Toll-like receptor signaling pathway* pathway. The linked circos plot depicts the genome-wide spread of inter-chromosomal pairwise interactions found between the genes members (N = 74) of the *Toll-like receptor signaling pathway*. Each coloured line represents a statistical interaction between a pair of physically unlinked loci. Only those interactions with a linkage disequilibrium correlation coefficient of $R^2 \geq 0.1$ are shown (N = 103 from 2,549 total interactions) to highlight the stronger inter-chromosomal interaction pairs within the pathway.

Table 12. Population-stratified pairwise R^2 correlation value of the top within-pathway inter-chromosomal interactions.

Genetic interaction pair						R^2		**ID #**
Chr	**SNP**	**Gene**	**Chr**	**SNP**	**Gene**	**European**	**African**	
20	rs6142206	*PIGU*	1	rs6427184	*DPM3*	5.89×10^{-5}	0.325	4
14	rs7156293	*TGFB3*	12	rs10774604	*PPP1CC*	4.10×10^{-5}	0.319	16
9	rs2310312	*ANXA1*	1	rs6671710	*CD244*	0.004	0.297	5
7	rs1005346	*CDK6*	6	rs1035798	*AGER*	0.004	0.283	15
3	rs2338577	*CCDC39*	1	rs3855955	*DVL1*	0.006	0.277	8
10	rs1043003	*KLF6*	6	rs4343924	*ULBP2*	0.002	0.277	5
3	rs5016648	*RNF168*	2	rs2140148	*CD28*	2.84×10^{-5}	0.259	5
11	rs628957	*UVRAG*	8	rs10099610	*CLVS1*	0.002	0.257	19
7	rs3735035	*PODXL*	1	rs11161581	*BCL10*	0.003	0.253	8
14	rs7156293	*TGFB3*	12	rs8176345	*CYP27B1*	0.007	0.245	12
7	rs1636874	*SHH*	3	rs2338577	*CCDC39*	0.011	0.241	8
15	rs502720	*EIF2AK4*	1	rs35021967	*LCK*	1.25×10^{-4}	0.241	5
19	rs7248036	*ATG4D*	17	rs4789814	*WDR45B*	0.004	0.240	19
17	rs1042678	*SOX9*	5	rs216136	*CSF1R*	0.018	0.238	8
19	rs11667267	*TICAM1*	4	rs223340	*UBE2D3*	0.030	0.236	17

Note—Each ID number corresponds to the respective ‘Confidently enriched’ pathway ID in Table 3. **Abbreviations**: Chr, chromosome; SNP; single nucleotide polymorphism.

3.3. Epistatic selection as a biologically plausible explanation for signals of within-pathway coevolution

To validate the attribution of signals of within-pathway coevolution (i.e., significant association between physically unlinked SNP-SNP pairs within a pathway) to forces of epistatic selection, I next evaluated the degree of overlap between the ancestry-enriched pathways against an integrated map of 722 positively selected chromosome regions compiled by Akey in 2009 [6]. This map was compiled from nine genome-wide scans performed within the HapMap [49] and Perlegen Biosciences [106] datasets. These scans utilized datasets containing SNP genotyping data from individuals of European-American and African-American ancestral background, and I decided to use this resource for my analysis. For each confidently enriched ($N = 19$) and nonenriched ($N = 18$) pathway, I identified all gene members that overlapped with a region that appeared in Akey's compilation of genome-wide scans (see Materials and Methods for additional details). To achieve this, I first compiled the genomic locations of all genes annotated to each pathway and then determined the relative degree of overlap with Akey's selection regions on a pathway by pathway basis. This method accounts for every possible variant mapped to a gene as a result, rather than restricting a gene's location to its single most significant SNP. The degree of overlap between the confidently enriched pathways and genomic targets of positive selection is summarized in **Table 13** and **Figure 13**.

When comparing the degree of overlap between the entire sets of confidently enriched and nonenriched pathways, I did not detect statistical significance ($p = 0.926$, OR = 0.555, 95% CI [0.198, ∞]) in favour of the enriched pathways. Notably, the pathways with the largest relative number of genes overlapping regions of positive selection included the *Nucleoside monophosphate metabolic process*, *Regulation of cAMP biosynthetic process*, and *Regulation of cyclic nucleotide metabolic process* pathways. As a result of the inherent similarity between the latter two pathways, as characterized by the pathway enrichment map (**Figure 5**), the identified overlapping genes are identical. However, no genes within the *Toll-like receptor signaling pathway* pathway overlapped with a selection region and only a single gene within the *Regulation of hemopoiesis* pathway was found to overlap. Since these pathways represent those of primary interest to this study, I did not conduct further downstream analysis to validate my evidence for ancestry-specific within-pathway coevolution via this methodology.

In addition to implementing the selection region overlap method originally proposed by Koch *et al.* [55] to validate the within-pathway signals of coevolution, I also performed a search for enriched pathway genes with evidence of recent positive selection using the recently curated database of Positive Selection across Human Populations (dbPSHP) [107]. As previously discussed, the International HapMap Project and 1000 Genomes Project produced high quality genotyping data across individuals of diverse ancestral backgrounds, enabling systematic detection of signals of natural selection on a genome-wide scale. Wang *et al.* have thus developed dbPSHP, a comprehensive web resource primarily focused on compiling evidence for positive selection across human populations. The database consists of > 15,000 recent signals of positive selection and related information in various human populations from the HM3 and 1KGP genotyping datasets, and additionally contains 15 statistical measures of positive selection for each SNP site from both datasets (see Materials and Methods for additional details). Using this resource, I identified a total of 57 and 21 unique genes with evidence for positive selection within the *Regulation of hemopoiesis* and *Toll-like receptor signaling pathway* pathways, respectively (**Supplementary Tables 4 and 5**). However, upon controlling for pathway sizes between both sets of pathways, I did not detect statistical significance (p = 0.731, OR = 0.946, 95% CI [0.802, ∞]) in favour of the ancestry-enriched pathways when comparing the degree of positive selection between the two sets (**Figure 14**). Nonetheless, the identification of numerous positively selected loci within my two primary pathways of interest remains a promising discovery for subsequent downstream functional analysis.

Table 13. Confidently enriched pathway genes overlapping with positively selected genomic regions.

Pathway name	**Overlapping gene(s)**
Cellular respiration	-
Energy derivation by oxidation of organic compounds	*PRKAG2*
Eukaryotic translation termination	-
Liposaccharide metabolic process	-
Lymphocyte activation	*NOTCH2*
Lymphocyte differentiation	*NOTCH2*
Macromolecular complex disassembly	*NRG1*
Morphogenesis of an epithelium	-
Nucleoside monophosphate metabolic process	*MAGI3, ADK, DLG2, PRKAG2*
Nucleoside triphosphate metabolic process	*PRKAG2*
Protein complex disassembly	*NRG1*
Regulation of bone mineralization	*FBN2*
Regulation of cAMP biosynthetic process	*RAF1, GRM8, ADCY8*
Regulation of cyclic nucleotide metabolic process	*RAF1, GRM8, ADCY8*
Regulation of hemopoiesis	*CDK6*
Rhythmic process	-
Toll-like receptor signaling pathway	-
TP53 regulates metabolic genes	*TNRC6B, PRKAG2*
Vacuole organization	-

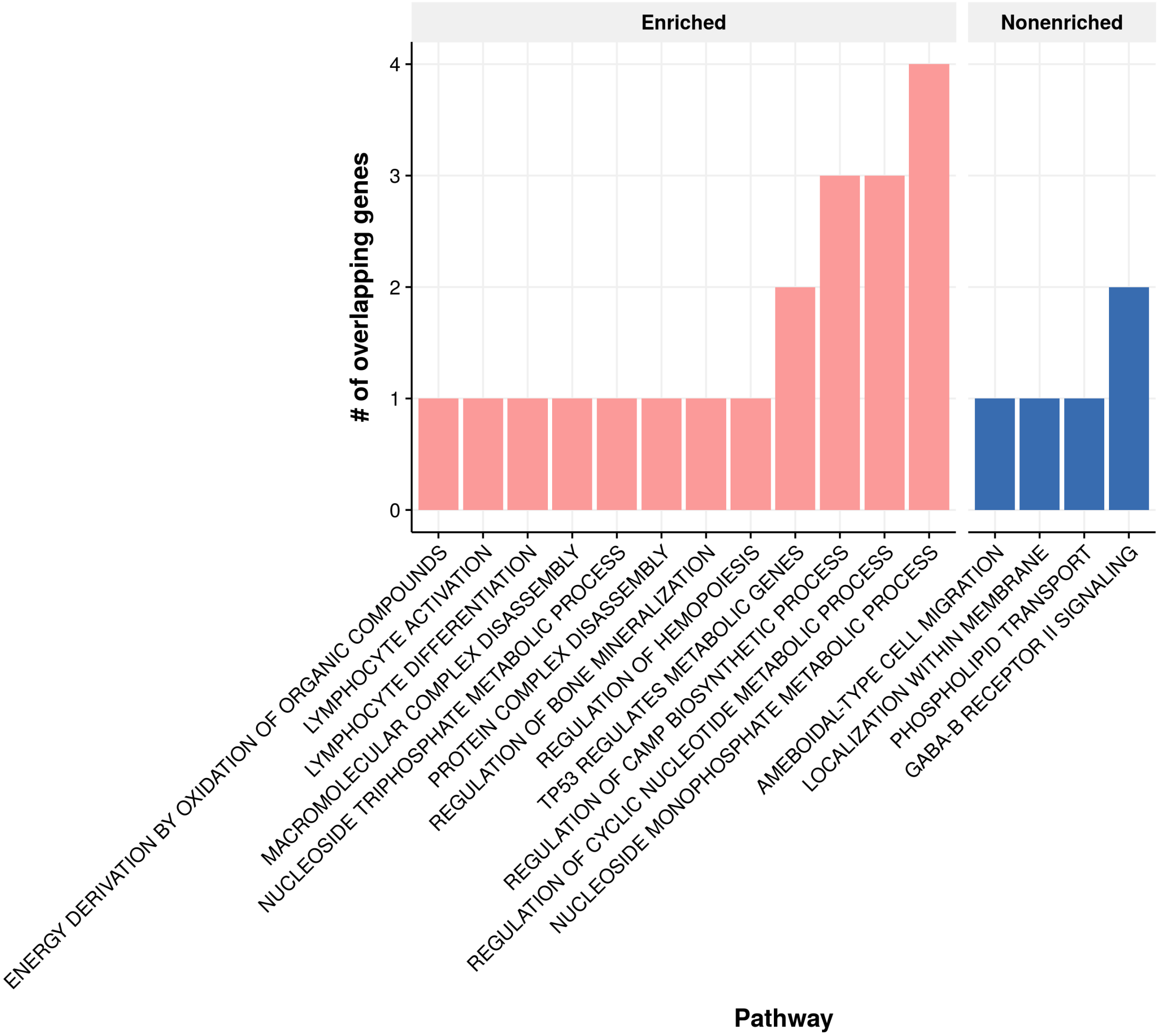


Figure 13. Pathway-level overlap with genomic targets of positive selection. Each bar represents the number of genes within a confidently enriched (pink) or nonenriched (blue) pathway that physically overlaps with a previously identified region of positive selection in the genome. Only pathways with at least one gene overlapping a selection region are shown. The entire set of confidently enriched pathways do not demonstrate a significantly greater (p = 0.926, OR = 0.555, 95% CI [0.198, ∞]) degree of overlap compared with the nonenriched pathways upon controlling for pathway size. The p value was calculated using the one-sided Fisher's exact test.

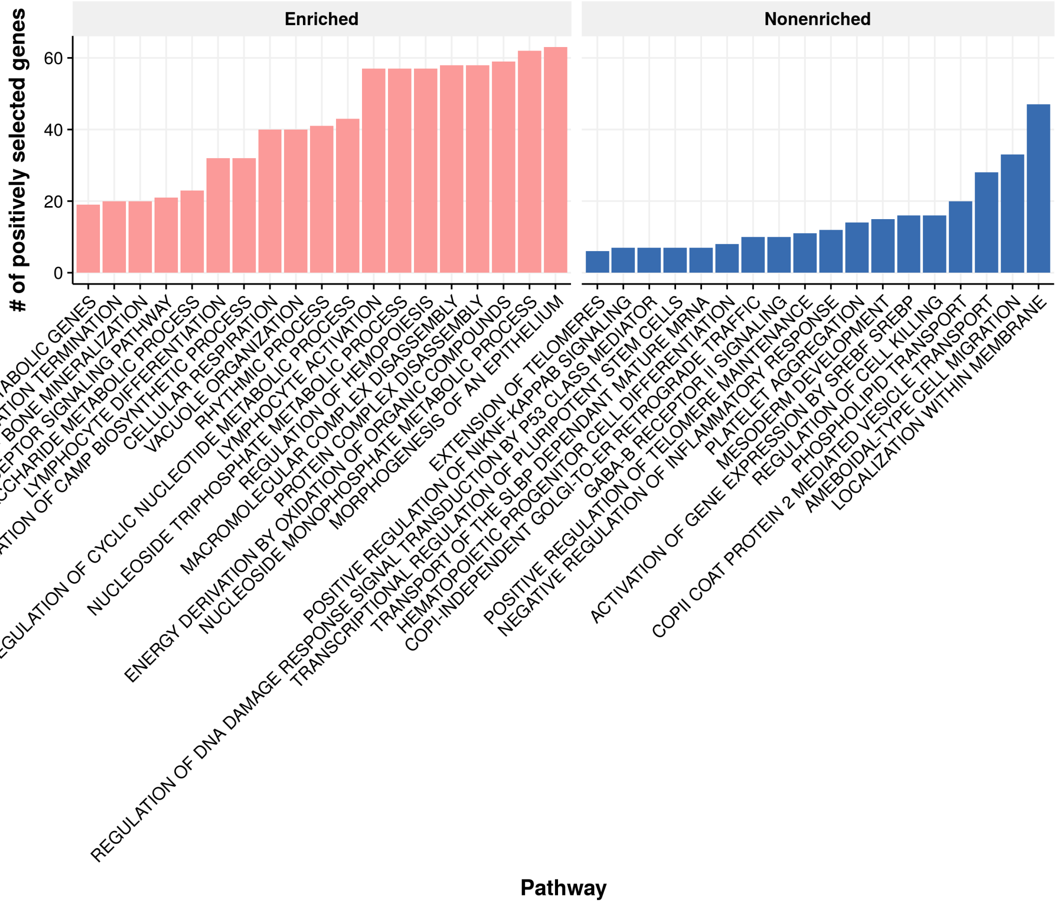


Figure 14. Pathway genes with evidence for recent positive selection. Each bar represents the number of genes within a confidently enriched (pink) or nonenriched (blue) pathway that contains statistically significant evidence for recent positive selection according to dbPSHP. The cumulative set of confidently enriched pathways do not demonstrate a significantly greater (p = 0.731, OR = 0.946, 95% CI [0.802, ∞]) degree of overlap compared with the nonenriched pathways upon controlling for pathway size. The *p* value was calculated using the one-sided Fisher's exact test.

4. Discussion

4.1. Summary and implications of research

In this thesis, I investigate evidence for within-pathway epistatic coevolution based on differential signals of SNP-level inter-chromosomal pairwise SNP-SNP association between individuals of European-American and African-American ancestry. As discussed in the Introduction, the human species encountered a highly diverse set of climatic, nutritional, and pathogenic conditions during their migration across the globe. Phenotypic traits increasing their chances of survival and reproduction in such environments were largely due to variation in genetic make-up and, as such, were transmitted across successive generations. Over the years, candidate gene and genome-wide association approaches have been used to identify the common variants associated with adaptable traits, allowing us to explore how past demographic events and natural selection have shaped the genetic diversity of human populations [95]. However, given the non-independent involvement of genes within functionally associated pathways, I implemented the GSEA pathway enrichment analysis method to identify sets of genetic variants associated with ancestral background, and identified 19 pathways enriched for various metabolic and immunological functions. Such observation of pathway enrichment might be expected due to the marked environmental heterogeneities experienced by European- and African-American populations. It is well known that the genetic make-up of the Americas has been significantly shaped by the Colonial Era and the Atlantic slave trade, which has long suggested historical and epidemiological evidence in support of substantial compositional heterogeneity between populations of European-American and African-American ancestry [108].

Exploration of the human genome for evidence of evolutionary events in the human genome has been crucial for the identification of genes underlying the broad morphological and physiological diversity observed across populations [9, 109]. Earlier in this thesis, I discussed the existence of genetic associations between physically unlinked loci as a fundamental characteristic of population-based variation in humans, as these associations could ultimately indicate the presence of epistatically-driven evolutionary forces at work [55]. Epistasis, specifically, has long been considered fundamentally important to understanding both the structure and function of genetic pathways and the evolutionary dynamics of complex genetic systems [68]. In this vein, I sought to uncover evidence for evolutionary adaptation acting within pathways enriched for human ancestral background via mechanisms of epistasis. Subsequent to the identification of 19 pathways confidently enriched for ancestry, I tested for signals of within-pathway epistatic coevolution by measuring the cumulative strength of association between pairwise combinations of inter-chromosomal variants within the enriched pathways. I then sought to validate my results by cross-referencing those significantly coevolving pathway loci with known positively selected regions of the genome, as well as by explicitly searching for coevolving pathway genes with statistically significant evidence for recent positive selection. Exclusively among African-Americans, I identified significant signals of coevolution within the *Regulation of hemopoiesis* and *Toll-like receptor signaling pathway* pathways, two pathways involved in the immune system. Although I was unable to validate the biological plausibility of epistatic selection as the primary driver for the identified signals, substantial evidence exists for positive selection acting on immune-associated traits across diverse human populations, particularly among individuals of European- and African-American ancestry.

Over the last decade, genomic scans of natural selection have identified genes and functions relating to immunity and host defense as targets of adaptive evolution [95]. Genes undergoing adaptive evolution through positive selection provide evidence of the functional variability that is beneficial to particular human populations, with complementary studies of the effects of selection on the diversity of immune loci increasing our understanding of the biological relevance of the functions concerned [110]. The contribution of genetic variants to variation in immune-associated traits is widely documented by GWAS [111-113]. However, as is a common issue with genome-wide approaches, the multiple variants found to be associated with those immune phenotypes tend to have small individual effect sizes and, as a result, the identification of causal functional variants has been challenging [25]. In an effort to understand the relationship between genetic variation and diversity of immune phenotypes, as well as the nature of the immunological mechanisms under selection, researchers have recently been employing analyses of expression quantitative trait loci (eQTLs) [114]. These are regulatory variants in the genome that influence gene expression, and are utilized as a tool to establish missing links between gene expression and immune response.

The analysis of eQTLs on lymphoblastoid cell lines from different human population cohorts has revealed that genetic variation accounts for differences in gene expression among individuals of various ancestral backgrounds [115-117]. One study measured gene expression in primary monocytes and T lymphocytes, demonstrating that cis-eQTLs are largely shared across populations, with only a small number of them showing population specificity [117]. Additionally, two recent studies determined the degree and underlying genetic mechanisms by which the response to immune stimulation is affected by population variation [118, 119]. Both studies explored the variance in transcriptional response to infection between European- and African-Americans through RNA sequencing, and mapped eQTLs in monocytes exposed to toll-like receptor ligands and influenza A virus [119], and in macrophages exposed to *Listeria monocytogenes* and *Salmonella Typhimurium* [118]. Despite significant differences in the experimental settings, both identified variation in gene regulatory regions that displayed strong ancestry-specific variation in response to immune stimulation, with the regulatory variants involved presenting different allele frequencies between individuals of European and African ancestry. Specifically, one team demonstrated that African ancestry is associated with a stronger inflammatory response in comparison with Europeans by measuring the rate of bacterial clearance post-infection, and provided further evidence that natural selection contributed to the ancestry-associated differences in gene regulation [118]. Although these findings were demonstrated at the single-gene level, they provide substantial support for the ancestry-specific coevolution signal that I identified within the immune-associated *Toll-like receptor signaling pathway* pathway. Regarding the *Regulation of hemopoiesis* pathway, site-specific correction of the sickle mutation in mouse hemopoietic stem and progenitor cells has been recently discovered as an effective therapeutic intervention for sickle cell diseases [120], the most common inherited class of blood disorders in African Americans [121]. Although evidence has yet to be found in humans, it provides an interesting topic for future research in regards to elucidating functional ancestry-specific epistatic interactions within hemopoietic pathways for targeted therapy.

On a somewhat separate note, the quantification of the epigenetic and environmental factors affecting the diversity of immune responses across human populations is crucially important to understand, as its variation cannot be entirely attributed to genetic factors [118, 119, 122-124]. A recent study has shown that a number of nongenetic factors (e.g., age and gender) and environmental variables (e.g., annual seasonality) have a major impact on the production of inflammatory cytokines [125]. This study highlights the need to consider not only gene–gene but also gene–environment interactions. The immune system is a rich environment to observe epistasis due to the marked complexity of genotype and phenotype, and evidence for functionally important epistatic interactions has been observed in several autoimmune diseases, such as rheumatic arthritis, systemic lupus erythematosus, and multiple sclerosis [82, 126]. For example, the interaction of two multiple sclerosis risk alleles in *DDX39B* (rs2523506) and *IL7R* (rs2523506A) increases the risk of disease considerably more than either variant independently [82]. Interestingly, TLR genes and associated pathways, particularly the toll-like receptor 4 pathway, have roles in multiple sclerosis [127-129]. Understanding these interactions at a functional and genetic level will be key elucidating mechanisms of susceptibility to a wide range of complex diseases in individuals of diverse ancestral background [130].

Ultimately, the detection of epistatic interactions at the level of pathways will provide crucial insights into the biological mechanisms that underlie disease pathophysiology, as well as improve our understanding of trait heritability and disease genetic architecture across individuals of diverse ancestral backgrounds. Researchers and clinicians have long known that different people demonstrate different levels of susceptibility to disease, medication response, and standardized clinical test results [131]. Epistatic interactions, such as those seen between *DDX39B* and *IL7R* in multiple sclerosis, are likely widespread within and between pathways and may represent a significant source of missing heritability associated with human traits and diseases. Thus, understanding functional pathway-level epistasis across diverse human populations could inform precision medicine by providing information about how specific variant associations operate in different biological contexts, for example among associated loci within a toll-like receptor or hemopoietic pathway, as well as by identifying subsets of individuals in which those associations may have a significant impact. The consideration of epistasis and genetic interaction networks in treatment thus remains a promising avenue for improving disease treatment, such as the prediction of drug response in tumours [132] and guidance of antibiotic drug-resistance [133].

4.2. Current limitations and future directions

The detection and quantification of long-range and unlinked linkage disequilibrium from high density genomic data challenging for population and evolutionary geneticists alike, with methods in the literature proposing widely variable *ad hoc* approaches [55]. Aside from the analysis by Koch *et al.* that I referenced earlier [55], I could find one other analysis of long-range LD associations in humans in the literature by Sved in 2011 [134]. In this study, correlations in heterozygosity between chromosome blocks in the HapMap dataset were explored and weak correlations between blocks on different chromosomes were discovered. However, neither the statistical significance of the correlations, nor integration of the findings within a biological pathway context, were reported. In this thesis, I measure the extent and strength of associations between inter-chromosomal SNP-SNP pairs to identify signals of epistatically-driven coevolution within functionally associated pathways—this is relatively novel territory. By and large, the approach I implement to detect these signals is a foundational, “first principles” method. I search for associations explicitly between physically unlinked loci to determine the level of coevolution signal within a pathway; consequentially, however, this fails to capture all potential SNP-SNP interactions within a given pathway. As previously discussed, I decided to employ this particular method as it allowed me to eliminate non-epistatic LD-generating effects that could cause confounding signals of LD association between genetic loci, such as recombination and genetic hitchhiking. Among genome-wide scans for signatures of positive selection, it is implicitly assumed that population demographic history is a genome-wide force affecting all loci independently and equally, whereas selection is a force acting only on a subset of loci [135]. Consequentially, I do not expect all members of a pathway to display significant association among all physically unlinked SNP-SNP pairs; nonetheless, it is highly promising that I identified two ancestry-enriched pathways with significantly stronger association compared to a set of pathways without that biological enrichment.

Aside from selection, however, it is important to consider alternative processes that may be responsible for associations between physically unlinked alleles. One hypothesis is that the sample includes some sort of structure, such as population admixture. Though this was not detected in the original analysis of the HapMap dataset [136], a more recent study confirmed that the HapMap African-American population is a genetically recent admixture with an average of 19.2% of European (CEU) and 80.8% of continental African (YRI; Yoruba in Ibadan, Nigeria) ancestry [137]. Additionally, African populations present the highest levels of diversity worldwide, with diversity of non-African populations decreasing with increasing distance from Africa, attesting to the occurrence of bottlenecks and founder events during their migrations across the globe [95]. Further analyses of within-pathway signals of coevolution will need to be applied to additional human population samples to test this hypothesis. A second hypothesis is that inter-chromosomal SNP-SNP association is the result of random genetic drift, perhaps amplified by variations in the demography of the studied European-American and African-American populations. This represents the most plausible alternative explanation, but would be difficult to test given two factors: i) simulating datasets comparable to those I have analyzed is computationally challenging [138], and ii) these datasets would have to account for the complex demographic histories of the tested human populations [55].

Several follow-up analyses can be potentially employed at this stage. For example, the genomes of modern human could be compared with those of Neandertals and Denisovans, as such studies have demonstrated evidence of adaptive archaic haplotypes in genes related to both metabolism and innate immune response in ethnically diverse modern human populations [7]. Future studies of ancient DNA will be informative for reconstructing the origin of functional variants and inferring the strength of selection based on direct observation of changes in allele frequencies over time [139]. However, our current understanding of ancient adaptation events is limited by sparse ancient DNA data in certain temporal and geographical regions, most notably in Africa, and methods for studying ancient genotype variation tend to focus on ascertained variants in European populations [7]. To pursue the method of detecting within-pathway epistatic coevolution via linkage disequilibrium, allelic association can be investigated among pairwise SNPs on the same chromosome flanking recombination regions, which would ultimately account for a greater proportion of SNP-SNP interactions within a pathway, though additional experimental models to validate the findings would still be required. One such experiment could be modelled after the human epistatic interaction analysis performed by Galarza-Muñoz *et al.* [82]. Using logistic regression modelling, this team demonstrated the significantly increased joint genotypic effect of two risk alleles, *DDX39B* (rs2523506) and *IL7R* (rs6897932), on the risk of multiple sclerosis. To test the interaction of those alleles, they implemented a functional model that depleted HeLa cells of *DDX39B* and used *IL7R* splicing reporters carrying either the risk allele or protective allele. Subsequent RT (reverse transcription)-PCR analysis of exon 6 splicing revealed higher instances of exon 6 skipping, a known driver of increased multiple sclerosis risk, when levels of *DDX39B* were reduced in the context of the risk allele compared with the protective allele. Notably, the *DDX39B* and *IL7R* risk variants reside on separate chromosomes, providing further evidence that inter-chromosomal genetic interactions represent biologically compelling models for disease genetics. As a result, downstream analyses could implement assay-based tests for functional epistasis between candidate allelic interactions upon testing for significant within-pathway signals of coevolution, in which the overall change in the pathway-associated phenotype can be assessed. Given the large number of ancestry-enriched pathway genes identified with evidence for positive selection (**Figure 14**), this analysis has promising functional implications.

In this regard, **Figures 11 and 12** illustrate potential candidates for elucidating epistatic interactions within the *Regulation of hemopoiesis* and *Toll-like receptor signaling pathway* pathways, respectively, by searching for frequently interacting genetic loci. For example, in the *Toll-like receptor pathway* pathway, we can observe several interactions occurring with the *LGMN* locus as well as the *TLR10-TLR1-TLR6* cluster. Mammalian *LGMN* is a newly identified, well-conserved lysosomal cysteine protease that processes the self and foreign antigens expressed by antigen presenting cells and proteolytically activates toll-like receptors. A recent study has implicated *LGMN* in the role of tumour development, and, as such, both diagnostic and therapeutic markers targeting these loci could potentially be developed [140]. Moreover, the *TLR1-TLR6-TLR10* cluster has been identified as a target of recent selection among individuals of European ancestry [141, 142], which could have interesting implications when interacting with *LGMN*. Considering the bigger picture, I used Reactome (version 3.5 database release 62) [143, 144] to observe the significant overrepresentation of the *Regulation of hemopoiesis* and *Toll-like receptor signaling pathway* pathways across an integrated network of biological pathways and processes. Here I observed a statistically significant overrepresentation (FDR ≤ 0.05) of the *Regulation of hemopoiesis* pathway within processes related to gene expression (transcription) and the immune system, specifically adaptive immunity and cytokine signaling. The *Toll-like receptor signaling pathway* pathway is also significantly overrepresented within the immune system, as well as within processes relating to external stimuli cellular responses and programmed cell death. Interestingly, this pathway is also significantly overrepresented among diseases associated with the TLR signaling cascade, a family of diseases primarily involved with the deficiency of various classes of TLR proteins (e.g., IRAK4, MyD88).

Given the marked role of toll-like receptors in human immunological diseases, it may be interesting to explore how evolutionary processes and selection pressures from pathogens have shaped the spread of the TLR polymorphisms, and associated pathways, across ethnically diverse human populations. In the human innate immune system, TLRs are positioned directly at the host-environment interface, representing the first line of defense against pathogenic infection [145]. Due to potential coevolutionary dynamics with pathogenic molecules, certain TLR polymorphisms are thought to result from frequent protective evolutionary pressures [145], possibly explaining variation in disease susceptibilities and clinical manifestations of immune diseases among individuals of diverse ancestral backgrounds [146]. For instance, in the African population, a prevalent *TLR4* variant adaptively protects against mortality caused by malaria infection [147]. And in the European population, as previously mentioned, the *TLR1-TLR6-TLR10* cluster is under evolutionary pressure [141, 142]. Furthermore, positive selection of certain *TLR1* alleles also occurs in Europeans, which likely reflects an attenuated inflammatory response and beneficial effects in sepsis [141], and is associated with a large network of genes displaying decreased levels of expression in response to immune activation [119]. Ongoing studies are needed to further clarify the role of genetic variation and disease susceptibility in this important class of innate receptors, and to provide important clues for therapeutic targeting of TLR pathways for the treatment of various immunological diseases [146]. Overall, future comprehensive analyses of large human populations are needed to address this hypothesis. Additionally, as previously mentioned, the elucidation of functional ancestry-specific epistasis within regulatory hemopoietic pathways would be an interesting line of research to pursue in regards to targeted sickle cell disease therapy.

Earlier, I highlighted the importance of applying the computational pipeline I have developed to a wide array of human population backgrounds in order to draw a complete picture of ancestry-enriched pathway coevolution. In this regard, I plan on analyzing a wider range of non-admixed population samples from the HM3 dataset, such as the Yoruba, Japanese, and Toscani ancestral populations. Currently, I have completed pathway enrichment analysis using whole-genome sequencing (WGS) SNP genotyping data from the 1KGP phase 2 dataset (https://www.cog-genomics.org/plink/1.9/resources, accessed May 24, 2017), which includes ~36 million SNPs genotyped across individuals from various ancestral backgrounds (see Introduction). WGS data undoubtedly holds greater information content in comparison with array-based data, providing the opportunity to determine intra- and inter-chromosomal LD structures at maximal resolution. Furthermore, WGS genotyping data lacks ascertainment bias that could be encountered with the use of array-based marker selection, a bias that has hampered the evolutionary analyses of SNP genotyping data [6]. In my thesis work, analysis was restricted to those SNPs that were segregating (polymorphic) in both populations to mitigate the impact of any ascertainment bias [148, 149]. Using ancestry-specific genotyping data from both the HM3 and 1KGP datasets, I have performed pathway enrichment analyses of two additional population comparisons: i) European-American (CEU) versus Han Chinese (CHB), and ii) African-American (ASW) versus Han Chinese (CHB). A summary of these findings is presented in **Figure 15 and Table 14**. My results thus far suggest that population variation has an effect on signals of epistatically-driven selective associations within biologically meaningful pathways, and that empirically evaluating these patterns, along with other lifestyle and environmental factors, across all ancestral populations of interest will be imperative towards understanding the background population-driven biases underlying pathway-level interactions and evolution.

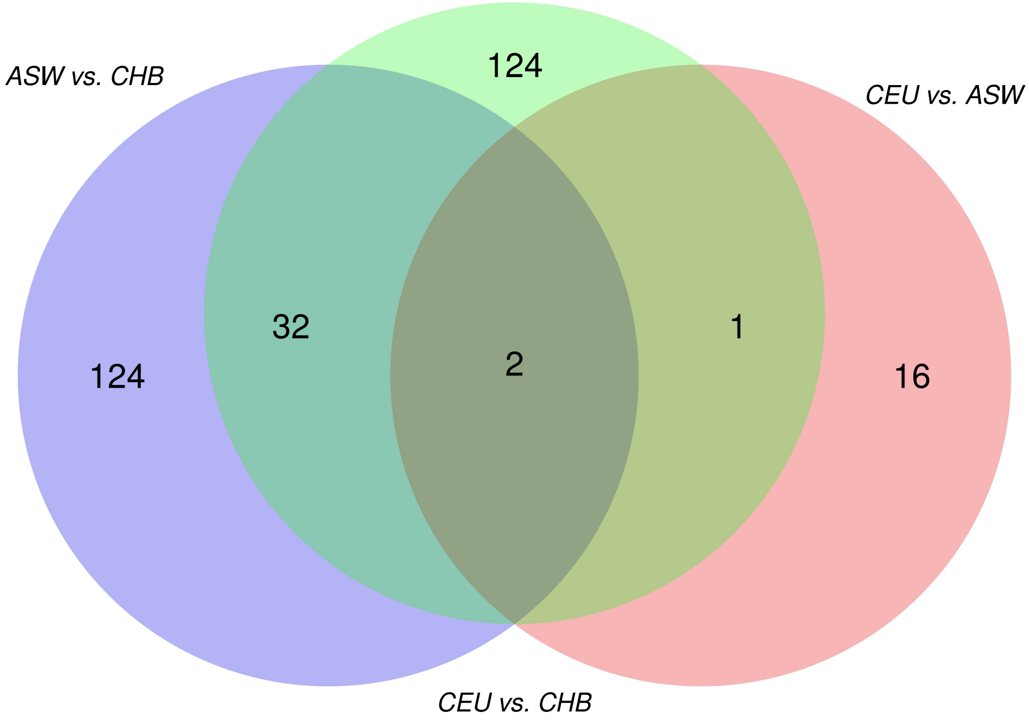


Figure 15. Venn diagram of significantly replicated ancestry-enriched pathways across three ancestry comparisons. Each coloured circle represents the total number of significantly replicated pathways that were identified from two independent GSEA analyses in the respective ancestry comparison. The overlap between the coloured circles represents the mutual ancestry-enriched pathways between the ancestry comparisons. The European vs. African comparison (pink) was completed using the HM3 and PNC array-based SNP genotyping datasets (i.e., the confidently enriched pathways presented in this thesis), whereas both the African vs. Chinese (blue) and European vs. Chinese (green) comparisons were completed using the HM3 and 1KGP datasets. The greater SNP coverage of the 1KGP whole-genome sequencing dataset most plausibly resulted in the higher number of significantly replicated pathways in the latter two analyses. **Abbreviations**: CEU, European-American; ASW, African-American; CHB, Han Chinese.

Table 14. List of significantly replicated ancestry-enriched pathways across three ancestry comparisons.

Pathway names		
CEU vs. CHB ∪ ASW vs. CHB	**CEU vs. ASW ∪ CEU vs. CHB**	**CEU vs. ASW ∪ CEU vs. CHB ∪ ASW vs. CHB**
Ameboidal-type cell migration	TP53 regulates metabolic genes	Eukaryotic translation termination
Calcium ion import		Morphogenesis of an epithelium
Cellular hormone metabolic process		
Cellular lipid catabolic process		
Cellular response to tumor necrosis factor		
Eukaryotic translation elongation		
Fatty acid metabolic process		
Isoprenoid metabolic process		
Lipid catabolic process		
Mesenchymal cell development		
Mesenchymal cell differentiation		
Mesenchyme development		
Negative regulation of cell growth		
Negative regulation of cellular component movement		
Negative regulation of growth		
Negative regulation of WNT signaling pathway		

Nonsense mediated decay (NMD) independent of the exon junction complex (EJC)
Positive regulation of defense response to virus by host
Positive regulation of vasculature development
Regulation of cellular response to growth factor stimulus
Regulation of cellular response to transforming growth factor beta stimulus
Regulation of defense response to virus by host
Regulation of protein binding
Regulation of transforming growth factor beta receptor signaling pathway
Regulation of transmembrane receptor protein serine/ threonine kinase signaling pathway
Response to tumor necrosis factor
Retinoid metabolic process
Stem cell development
Stem cell differentiation
Steroid metabolic process
Visual phototransduction
Xenophagy

Abbreviations: CEU, European-American; ASW, African-American; CHB, Han Chinese.

4.3. Research applications

The original motivation for my thesis project was to gain an improved understanding of pathway-level genetic variation across individuals of diverse ancestral backgrounds. In an effort to gain such an understanding, researchers have increasingly been analyzing GWAS data at the functional pathway level (e.g., [150-154]). For example, a recent high-profile GWA study [152] identified 65 new breast cancer risk loci among individuals of European and East Asian ancestry, in which pathway enrichment analysis was performed in order to understand the broader biological context of the newly identified risk loci. Enrichment of several growth/development and cancer related pathways (e.g., interferon signaling and cell-cycle pathways) were discovered, in addition to other pathways not previously found in earlier breast cancer GWAS.

Recently, a method known as BridGE (Bridging Genes with Epistasis) [34] has been developed by the Myers Lab group based in the University of Minnesota, a group with strong collaborative ties to the Boone and Andrews Labs (visit https://www.bridgegenomics.com for additional details). From GWAS cohort data, BridGE identifies pathway-level genetic interactions in human populations based on the between-pathway framework of genetic interactions [155]. Specifically, this framework refers to the clustering of interactions into coherent groups that connect across two functionally-distinct pathways, in contrast to the within-pathway framework that concerns the clustering of interactions within the same functional pathway. Using the BridGE method, significant pathway-level interactions were identified in several breast cancer cohorts representing many different ancestral backgrounds, in which the identified pathways included many relevant and newly-identified risk modifying variants associated with the disease, additionally demonstrating how genetic interactions differ across different human populations [154].

In an additional approach to understand pathway-level population variation, members of the Bader Lab are currently developing a machine learning-based patient classification framework called netDx [153], a novel R-based tool that serves as a mechanism for identifying biological pathways and biomarkers important for clinical treatment response prediction using pathway-level gene expression data. Based on quantitative patient information present in electronic medical record databases, the goal for netDx is to provide clinical researchers with the means to tailor treatment plans to a patient, ultimately providing a “complete framework for precision medicine” [153]. In the realm of post-GWAS analysis, netDx has widespread and intriguing biomedical and clinical applications.

However, these post-GWAS studies share a common methodological limitation—the inability, or lack of power, to identify pathway-level variation and genetic interactions across multiple populations in a single analysis. This limitation can be largely attributed to the lack of understanding background pathway-level biases present within human ancestral genomes, and thus the appropriate corrections cannot be made to address those biases. By implementing a comparative analysis of pathway enrichment and unlinked selective associations to characterize biologically meaningful within-pathway genetic interactions between pairs of diverse ancestral cohorts, we can begin to address this particular limitation. In this work, I have elucidated evidence for selection-induced genetic interactions within the *Regulation of hemopoiesis* and *Toll-like receptor signaling pathway* pathways specific to African-Americans as compared to European-Americans, demonstrating the potential of epistasis to promote associations between favourable allelic combinations among pathways that influence population fitness. A similar type of analysis can also be applied on a between-pathway basis in the future to discover the potential of adaptively compensatory pathways, ultimately advancing the goal of precision-based medicine via a global characterization within- and between-pathway genetic interactions across human populations.

5. Materials and Methods

The complete pipeline can be accessed online at https://github.com/rosscm/PopulationPathways (currently a privately owned repository). I developed this code primarily in the R statistical programming language [156].

5.1. SNP genotyping data and preprocessing steps

Genome-wide single nucleotide polymorphism (SNP) genotyping data were taken from the publicly available International HapMap Project phase 3 (HM3) (https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html) and the Philadelphia Neurodevelopmental Cohort (PNC) via dbGaP (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000607.v1.p1). A tabular summary of these data can be found in **Table 1.**

5.2. International HapMap Project phase 3

Data from the HM3 cohort consist of 1,594,675 non-imputed polymorphic SNPs (hg19 build) genotyped (total rate = 87.25%) in 165 individuals of genetically European-American ancestry (CEU; Utah residents with Northern and Western European ancestry) and 83 individuals of genetically African-American ancestry (ASW; African ancestry in Southwest USA). The markers were genotyped in several other ancestral samples [49] that were not analyzed in this thesis. Principal components analysis (PCA) was performed via PLINK v1.9 [157, 158] to visually summarize the genotyping data in two dimensions (**Figure 7**). PCA effectively demonstrates the distinct genotypic makeup of the studied European- and African-American individuals.

5.3. Philadelphia Neurodevelopment Cohort

Data from the PNC cohort consist of 3,730,475 imputed polymorphic SNPs (hg19 build) genotyped (total rate = 98.51%) in 9,498 subjects with medical, psychiatric, neurocognitive, and genomic data [159]. Since this data were previously available at the beginning of my thesis, they were used here for replication purposes when running pathway enrichment analysis. Imputation was completed by Shraddha Pai, a post-doc currently working in the Bader Lab, using tools provided by the Ritchie Lab (https://ritchielab.psu.edu; accessed in Summer 2016). Only imputed

SNPs with an info score > 0.8 were considered, a threshold chosen based on the pathway analysis performed by the Pathway Genomics Consortium working group [160]. Additional quality control measures included: retaining SNPs with geno score > 0.99, mind score > 0.99, and the exclusion of variants with one or more multi-character allele codes (performed via PLINK v1.9 [157, 158]). Final quality control measures were performed to extract ancestry-specific samples, which included: excluding SNPs with mind score < 0.05 and geno score < 0.05, excluding samples that failed IBD threshold (139 samples), and excluding symmetric SNPs and SNPs from high LD regions (683,023 SNPs). Samples imputed on the Axiom genotyping array were removed as the imputation rate was skewed for this platform, removing 6 and 633 CEU and ASW samples, respectively. The final PNC dataset comprised genotyping data from 3,314 individuals of genetically European-American ancestry (CEU) and 1,840 individuals of genetically African-American ancestry (ASW). Ancestry for these samples was defined as groups that were located within ±5 standard deviations of the PCA PC1, PC2 centroid in the reference HM3 population (**Figure 7**).

5.4. Pathway annotations

I downloaded a set of human biological pathway annotations from the Bader Lab database (http://download.baderlab.org/EM_Genesets/April_24_2016/Human/symbol/Human_GOBP_AllPathways_no_GO_iea_April_24_2016_symbol.gmt, accessed April 24, 2016). This database contains a repository of regularly updated pathway annotations that is compiled from manually and electronically curated pathway databases, including Reactome, BioCarta, GO (Gene Ontology) biological process, NCI Pathway Interaction Database, HumanCyc, MSigdb, NetPath, and Panther. GO terms were restricted solely to the 'biological process' class of pathways, thereby excluding those from the 'molecular function' and 'cellular component' classes. This restriction was applied since the large, hierarchal structure of GO annotations result in a high level of pathway overlap that could potentially obscure the true source of an association signal [22]. The chosen pathway set also excludes annotations that were inferred from electronic annotation and/or reviewed computational analysis, as well as annotations without availability of biological data. Pathways were restricted to a minimum of 20 genes and a maximum of 200 genes in order to control for potential inadvertent associations, as small pathways can exhibit false positive associations due to large single-gene or single-SNP effects [161], and large pathways are more likely to show association by chance alone [162]. Final pathway size was determined by the total number of pathway genes that were included in the HM3 and PNC datasets upon SNP-to-gene mapping. These inclusion measures rendered a final set of 3,781 human biological pathways from the original set of 15,835 pathways that were tested. To note, of the 14,432 unique genes annotated to the pathways within the original annotation file, 13,957 (97%) are protein-coding according to HGNC (ftp://ftp.ebi.ac.uk/pub/databases/genenames/new/tsv/locus_types/gene_with_protein_product.txt, downloaded January 24, 2018). Similarly, ~97% of the 12,416 unique genes annotated to the size-reduced pathway annotation file are protein-coding. Additional pathway size thresholds were tested, but no significant differences were found in the results.

5.5. Test for pathways enriched for ancestry

The pathway enrichment-based analysis method known as Gene Set Enrichment Analysis (GSEA) was used to identify biological pathways significantly enriched for human ancestral background. I performed GSEA using the perl GenGen v.1.0.1 package [88] (available from https://github.com/WGLab/GenGen/releases, downloaded May 24, 2017). In brief, the goal of GSEA is to determine whether the members of a pathway *S* are randomly distributed throughout the entire reference gene list *L* or are found primarily at the top or bottom of *L* (see **Supplementary Figures 1 and 2**), and is relatively robust to noise and outliers in the data. The initial step of pathway enrichment analysis requires the integration of genomic data into pathways by mapping such data to genes. For SNP-based genotype arrays, this is not a straightforward process since many SNPs are not located in known coding or regulatory regions. At this stage, I assigned each unmapped marker to its nearest gene via a simple distance-based SNP-to-gene mapping method [87]. Of the HM3 and PNC markers, 88% and 79% were mapped to protein-coding genes, respectively. GSEA utilizes one association signal per gene; however, SNP genotyping arrays include multiple, and possibly correlated, signals per gene. The operative signal of each gene was thus represented by the single largest positive (i.e., most significant) SNP-level test statistic of all SNPs mapped to that gene. For the purposes of my thesis, the SNP-level test statistic was defined as the difference in minor allele frequency (ΔMAF) per SNP between the tested populations (**Figure 2**). Population-stratified MAF was calculated using PLINK v.1.9 [157, 163], in which each SNP's corresponding test statistic was defined as,

$$\Delta MAF = MAF_{popA} - MAF_{popB}$$

where popA and popB correspond to the European-American and African-American populations, respectively.

As a final filtration step before performing pathway enrichment analysis, SNPs located more than 10kb away from its mapped gene were excluded, leaving approximately 54% and 48% of the total variants genotyped in the HM3 and PNC datasets, respectively, to be analyzed within pathways (**Table 2**). The single-SNP-mapped genes are then grouped together as a biologically related set of genes, or pathways, based on a set of *a priori* pathway annotations (see **Subsection 4.4**). Each pathway gene is ranked from highest to lowest according to its test statistic (ΔMAF), producing an overall ranked list of genes. The significance of each pathway is then judged based on overrepresentation of pathway genes toward the top of the overall ranked gene list. An enrichment score (ES) is produced using a rank-based Kolmogorov–Smirnov-like statistic, measuring the deviation of the association statistics in a given pathway in comparison with a set of randomly selected genes (of the same size to the tested pathway). The genes that appear in the ranked list *L* at or before the point at which the running sum reaches its maximum deviation from zero is the leading-edge subset, and represents the core genes that account for a pathway's enrichment signal (**Supplementary Table 3**). However, since larger genes will harbour SNPs with higher test statistics by chance, a normalized enrichment score (NES) is also produced. The NES adjusts for gene size through permutation-based ($N = 1{,}000$ cycles) label swapping, which consequently allows for cross-comparison of GSEA results across all tested pathways. An empirical *p* value for each individual pathway is generated via the permutation approach, alongside a false discovery rate (FDR) calculated using NES scores from all permuted values in all pathways examined in a single experiment (see reference [88] for additional details).

5.6. Determining true pathway enrichment

Pathways were determined to be confidently enriched for ancestry upon replication in two independent SNP genotyping cohorts. Significant pathway enrichment was defined as pathways with FDR ≤ 0.1 and FDR ≤ 0.05 in the HM3 and PNC datasets, respectively. FDR thresholds were chosen based on the relative difference in sample size between the two datasets (**Table 1**). The replicated pathways that passed the respective significance threshold in each dataset were referred to as the 'confidently enriched' or 'ancestry-enriched' pathways. The 19 ancestry-enriched pathways contain a total of 2,307 single-gene-annotated SNPs (1,501 unique). Conversely, pathways without ancestral enrichment (i.e., nonenriched) were defined as those with a NES ≃ 0 (specifically within the range of [-0.1 ≥ NES ≤ 0.1]) in each dataset. Replicated pathways within this range were referred to as the 'nonenriched' pathways. The nonenriched pathways contain a total of 761 single-gene-annotated SNPs (672 unique). Given the nature of pathway enrichment analysis, it is possible that the association measure of a pathway can be erroneously inflated if a single SNP is correlated and consequently assigned to multiple genes in the same pathway (e.g., the HLA cluster on chromosome 6) [22]. However, since the primary aim of this thesis is to determine evidence for within-pathway epistatic selection by measuring linkage disequilibrium between variants residing specifically on separate chromosomes, I do not consider this as a potential confounder. An enrichment map was created to thematically summarize the confidently enriched pathways using the Enrichment Map version 3.0 app [89] in Cytoscape version 3.5.1 [164]. Related pathway nodes were clustered and labelled as themes using the AutoAnnotate Cytoscape app [165].

5.7. Test for within-pathway signals of epistatically-driven coevolution

In the Introduction, I characterized three patterns of linkage disequilibrium that can exist between loci throughout the human genome: (i) short-range on the same chromosome, (ii) long-range on the same chromosome, and (iii) long-range on separate chromosomes (i.e., inter-chromosomal). The first pattern of LD described is not of relevance to this thesis, as its primary use is concerned with disease association mapping. Between measuring the second and third patterns of LD, in the latter, I can effectively eliminate single-chromosome LD-generating factors such as genetic hitchhiking and chromosomal structural variations, and focus specifically on the determination of LD generated by epistatic selection. I thus measured LD association explicitly between inter-chromosomal pairwise variants within each ancestry-enriched pathway as a proxy to test for signals of within-pathway epistatic selection. However, until subsequent validation of true epistatic effects, any significantly identified signal via this method was referred to as a coevolving pathway.

To measure linkage disequilibrium, the observed and expected frequencies for a biallelic pair of SNPs are compared, in which the difference between these two values constitutes the deviation or D for that particular combination. If two loci are in linkage equilibrium, then $D = 0$; conversely, if two loci are in linkage disequilibrium, then $D \neq 0$. Thus, for a single inter-chromosomal pairwise interaction between genes A and B, with alleles A/a and B/b

$$D = p_{AB} - p_A p_B$$

where p_{AB} is the probability of seeing the marker allele pair AB, p_{A} is the observed probability of allele A, and p_{B} is the observed probability of allele B. I utilize the ld function from the R snpStats package [166] for this calculation, which reports LD association via the [-1, 1] scaled D' statistic and the corresponding Pearson's R^2 correlation coefficient. I report the strength of each inter-chromosomal pairwise SNP-SNP interaction by the R^2 correlation coefficient, which is calculated as

$$R^2 = \frac{D^2}{p_A p_a p_B p_b}$$

where p_{a} $(1 - p_{\mathrm{A}})$ is the observed probability of allele a and p_{b} $(1 - p_{\mathrm{B}})$ is the observed probability of allele B. In this equation, $R^2 = 0$ indicates two loci existing in complete linkage equilibrium and $R^2 = 1$ indicates two loci existing in complete linkage disequilibrium. For each pairwise SNP combination, phased allele pair frequencies are estimated by maximum likelihood using the method described by Clayton and Leung [167]. R^2 arrays representing the strength of each inter-chromosomal pairwise SNP-SNP interaction are then tabulated for each tested pathway.

5.8. Determining significance of within-pathway coevolution signal

Significance per within-pathway signal of coevolution was determined for the confidently ancestry-enriched pathways using the nonparametric two-sample one-sided Kolmogorov–Smirnov (KS) test (less). The R^2 distribution of the inter-chromosomal pairwise interactions per confidently enriched pathway (X_n) was compared to the total R^2 distribution of the inter-chromosomal pairwise interactions within the complete set of nonenriched pathways (Y). Each reported p value calculated by the KS test implemented the 'less' alternative hypothesis, which specifies that the true distribution function of X_n is lesser than the distribution function of Y. The KS test is a comparison of cumulative distribution functions in which the test statistic (D^-) is the maximum difference in value, and is calculated as

$$D^- = max[f_Y(u) - f_{X_n}(u)]$$

when implementing the 'less' alternative. Thus, in the two-sample case, this test statistic includes distributions for which X_n is stochastically greater than Y (i.e., the cumulative distribution function of X_n lies below and hence to the right of that for Y) [156]. After applying a Bonferroni correction to adjust for multiple testing, nominal significance was determined as $p < 0.003$ (0.05/19). Consequently, when assessing significance per ancestry-specific within-pathway signal of coevolution, nominal significance was determined as $p < 0.005$ (0.05/11). The denominator adjustment reflects the total number of pathways in which significant coevolution signal was previously identified from the total set of ancestry-enriched pathways.

5.9. Validation of pathways as genomic targets of population-driven epistatic selection

In order to validate whether the confidently enriched pathways are associated with previously identified targets of positive selection in the genome, and thereby as targets of epistatic selection, I evaluated the degree of overlap between the given pathway genes and an integrated genomic map of 722 positively selected chromosome regions compiled by Akey [6] (a method previously implemented by Koch *et al.* [55]). These regions contain a total of 2,465 genes that span 245Mb of the genome (~8%), and were compiled from nine genome-wide scans performed in the HapMap [49] and Perlegen Biosciences [106] datasets. For the purposes of my thesis, the coordinates of each selection region were lifted from the UCSC hg18 genomic build to the hg19 build (corresponding to the genome build of the HM3 dataset), which resulted in four dropped coordinates. The nine genome-wide scans employed tests based on within-species polymorphism to identify targets of positive selection in humans (Supplemental Table 1 in [6]). These methods fall into three general categories: (i) site frequency-based methods (Tajima's D and Fay and Wu's F), (ii) linkage disequilibrium-based methods (e.g., extended haplotype homozygosity test and integrated haplotype score), and (iii) population differentiation-based methods (F_{ST} fixation index) [109]. Studies have highlighted the importance of utilizing multiple measures to investigate evidence for selection [168], as each tool employs different patterns of genetic variation dependent on the nature and time scale in which the selection occurred [11]. For each confidently enriched pathway (N = 1,501 total unique genes), I identified all gene members that overlapped with a region appearing in Akey's genome-wide scan. To determine if the number of these occurrences is larger than expected for any given non-associated pathway, the same test was conducted for the set of nonenriched pathways (N = 672 total unique genes). The genomic location and corresponding HGNC ID symbol for each pathway gene was downloaded in the hg19 assembly from https://genome.ucsc.edu/cgi-bin/hgTables (accessed October 10, 2017). Nominal significance ($p < 0.05$) was calculated using the one-sided Fisher's exact test (greater).

As an additional method of within-pathway epistatic selection validation, I used the recently curated database of Positive Selection across Human Populations (dbPSHP) [107] in an effort to identify genes within the ancestry-enriched (N = 1,501 total unique genes) and nonenriched (N = 672 total unique genes) pathways with evidence for recent positive selection. The integrated literature-based database contains 15,472 manually collected loci physically located within approximately 8,000 genes (hg19 build) from studies that have detected positive selection in both a specific function-related manner (101 publications) and on a genome-wide scale (31 publications). For each genotyped SNP from the HM3 and 1KGP databases, 15 statistical terms measuring the degree of positive selection is provided according to the population in which the evidence for selection was determined. According to the dbPSHP web browser, these statistical terms include measures of variant allele frequency, variant heterozygosity, within population diversity, haplotype homozygosity, long-range haplotypes, pairwise population differentiation, and evolutionary conservation. Each empirical score employed a defined cutoff based on those frequently used for estimations of positive selection in current evolutionary studies in order to correct for false positive signals (Supplemental Table in 6 [107]). The curated loci were further evaluated for reliability and accuracy by validating the statistical scores generated for two popular cases of strong population-specific selection, *LCT* and *SLC24A5*, and comparing the 15-score distribution for the observed selective loci against the background (i.e., random selection of genomic loci). The curated evidence file I use for reference was downloaded from ftp://jjwanglab.org/dbPSHP/curation/dbPSHP_20131001.tab (accessed January 4, 2018). Nominal significance ($p < 0.05$) was calculated using the one-sided Fisher's exact test (greater).

6. Bibliography

1. Darwin CR, Wallace AR. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. Journal of the Proceedings of the Linnean Society of London Zoology. 1858;3:46-50.

2. Darwin CR. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. 1st ed. London, UK: John Murray; 1859.

3. Provine WB. The origins of theoretical population genetics. Chicago, IL: University of Chicago Press; 1971.

4. Kimura M. Evolutionary rate at the molecular level. Nature. 1968;217(5129):624-6.

5. King JL, Jukes TH. Non-Darwinian evolution. Science. 1969;164(3881):788-98.

6. Akey JM. Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res. 2009;19(5):711-22.

7. Racimo F, Sankararaman S, Nielsen R, Huerta-Sanchez E. Evidence for archaic adaptive introgression in humans. Nature reviews Genetics. 2015;16(6):359-71.

8. Hsieh P, Woerner AE, Wall JD, Lachance J, Tishkoff SA, Gutenkunst RN, et al. Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies. Genome Res. 2016;26(3):291-300.

9. Fan S, Hansen ME, Lo Y, Tishkoff SA. Going global by adapting local: A review of recent human adaptation. Science. 2016;354(6308):54-9.

10. Chang CQ, Yesupriya A, Rowell JL, Pimentel CB, Clyne M, Gwinn M, et al. A systematic review of cancer GWAS and candidate gene meta-analyses reveals limited overlap but similar effect sizes. European journal of human genetics : EJHG. 2014;22(3):402-8.

11. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, et al. Positive natural selection in the human lineage. Science. 2006;312(5780):1614-20.

12. Turchin MC, Chiang CW, Palmer CD, Sankararaman S, Reich D, Genetic Investigation of ATC, et al. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nature genetics. 2012;44(9):1015-9.

13. Berg JJ, Coop G. A population genetic signal of polygenic adaptation. PLoS genetics. 2014;10(8):e1004412.

14. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, et al. Genetic signatures of strong recent positive selection at the lactase gene. American journal of human genetics. 2004;74(6):1111-20.

15. Huerta-Sanchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature. 2014;512(7513):194-7.

16. Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, et al. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science. 2001;293(5529):455-62.

17. Hamblin MT, Di Rienzo A. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. American journal of human genetics. 2000;66(5):1669-79.

18. Hamblin MT, Thompson EE, Di Rienzo A. Complex signatures of natural selection at the Duffy blood group locus. American journal of human genetics. 2002;70(2):369-83.

19. Friedman MJ. Erythrocytic mechanism of sickle cell resistance to malaria. Proceedings of the National Academy of Sciences of the United States of America. 1978;75(4):1994-7.

20. Currat M, Trabuchet G, Rees D, Perrin P, Harding RM, Clegg JB, et al. Molecular analysis of the beta-globin gene cluster in the Niokholo Mandenka population reveals a recent origin of the beta(S) Senegal mutation. American journal of human genetics. 2002;70(1):207-23.

21. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic acids research. 2017;45(D1):D896-D901.

22. Ramanan VK, Shen L, Moore JH, Saykin AJ. Pathway analysis of genomic data: concepts, methods, and prospects for future development. Trends in genetics : TIG. 2012;28(7): 323-32.

23. Bamshad M, Wooding SP. Signatures of natural selection in the human genome. Nature reviews Genetics. 2003;4(2):99-111.

24. Mooney SD, Krishnan VG, Evani US. Bioinformatic tools for identifying disease gene and SNP candidates. Methods Mol Biol. 2010;628:307-19.

25. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53.

26. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences of the United States of America. 2012;109(4):1193-8.

27. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nature reviews Genetics. 2009;10(6):392-404.

28. Boone C, Bussey H, Andrews BJ. Exploring genetic interactions and networks with yeast. Nature reviews Genetics. 2007;8(6):437-49.

29. Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan G, et al. A global genetic interaction network maps a wiring diagram of cellular function. Science. 2016;353(6306).

30. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, et al. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science. 1999;285(5429):901-6.

31. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, et al. Functional profiling of the Saccharomyces cerevisiae genome. Nature. 2002;418(6896):387-91.

32. Krause SA, Gray JV. The functional relationships underlying a synthetic genetic network. Commun Integr Biol. 2009;2(1):4-6.

33. Blomen VA, Majek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, Staring J, et al. Gene essentiality and synthetic lethality in haploid human cells. Science. 2015;350(6264):1092-6.

34. Fang G, Wang W, Paunic V, Heydari H, Costanzo M, Liu X, et al. Discovering genetic interactions bridging pathways in genome-wide association studies. bioRxiv. 2017.

35. Schadt EE. Molecular networks as sensors and drivers of common human diseases. Nature. 2009;461(7261):218-23.

36. Hirschhorn JN. Genomewide association studies--illuminating biologic pathways. The New England journal of medicine. 2009;360(17):1699-701.

37. Menashe I, Maeder D, Garcia-Closas M, Figueroa JD, Bhattacharjee S, Rotunno M, et al. Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade. Cancer Res. 2010;70(11):4453-9.

38. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nature reviews Genetics. 2010;11(12):843-54.

39. Zhong H, Yang X, Kaplan LM, Molony C, Schadt EE. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. American journal of human genetics. 2010;86(4):581-91.

40. Need AC, Attix DK, McEvoy JM, Cirulli ET, Linney KL, Hunt P, et al. A genome-wide study of common SNPs and CNVs in cognitive performance in the CANTAB. Human molecular genetics. 2009;18(23):4650-61.

41. Bustamante CD, Burchard EG, De la Vega FM. Genomics for the world. Nature. 2011;475(7355):163-5.

42. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016;538(7624): 161-4.

43. Levenson D. Non-European populations still underrepresented in genomic testing samples: Dearth of African-American, Latino, and other non-European groups contributes to healthcare, research disparities. Am J Med Genet A. 2017;173(2):296-7.

44. Desta Z, Ward BA, Soukhova NV, Flockhart DA. Comprehensive evaluation of tamoxifen sequential biotransformation by the human cytochrome P450 system in vitro: prominent roles for CYP3A and CYP2D6. J Pharmacol Exp Ther. 2004;310(3):1062-75.

45. Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, et al. Genetic Misdiagnoses and the Potential for Health Disparities. The New England journal of medicine. 2016;375(7):655-65.

46. Richard P, Charron P, Carrier L, Ledeuil C, Cheav T, Pichereau C, et al. Hypertrophic cardiomyopathy: distribution of disease genes, spectrum of mutations, and implications for a molecular diagnosis strategy. Circulation. 2003;107(17):2227-32.

47. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001;409(6822):928-33.

48. Thorisson GA, Stein LD. The SNP Consortium website: past, present and future. Nucleic acids research. 2003;31(1):124-7.

49. International HapMap Consortium. A haplotype map of the human genome. Nature. 2005;437(7063):1299-320.

50. The 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.

51. Zheng-Bradley X, Flicek P. Applications of the 1000 Genomes Project resources. Brief Funct Genomics. 2017;16(3):163-70.

52. Hara K, Fujita H, Johnson TA, Yamauchi T, Yasuda K, Horikoshi M, et al. Genome-wide association study identifies three novel loci for type 2 diabetes. Human molecular genetics. 2014;23(1):239-46.

53. Slatkin M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nature reviews Genetics. 2008;9(6):477-85.

54. Robinson MA. Linkage Disequilibrium. Encyclopedia of Immunology. 2 ed1998. p. 1586–8.

55. Koch E, Ristroph M, Kirkpatrick M. Long range linkage disequilibrium across the human genome. PloS one. 2013;8(12):e80754.

56. Gibson G, Muse S. A primer of genome science. Sunderland, MA: Sinauer Associates; 2009.

57. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006;4(3):e72.

58. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007;449(7164):913-8.

59. Lewontin RC, Kojima, K. The Evolutionary Dynamics of Complex Polymorphisms. Evolution. 1960;14(4):458-72.

60. Lewontin RC. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics. 1964;49(1):49-67.

61. Wilson JF, Goldstein DB. Consistent long-range linkage disequilibrium generated by admixture in a Bantu-Semitic hybrid population. American journal of human genetics. 2000;67(4):926-35.

62. Schaper E, Eriksson A, Rafajlovic M, Sagitov S, Mehlig B. Linkage disequilibrium under recurrent bottlenecks. Genetics. 2012;190(1):217-29.

63. Slatkin M. Linkage disequilibrium in growing and stable populations. Genetics. 1994;137(1):331-6.

64. Schmegner C, Hoegel J, Vogel W, Assum G. Genetic variability in a genomic region with long-range linkage disequilibrium reveals traces of a bottleneck in the history of the European population. Human genetics. 2005;118(2):276-86.

65. Caceres A, Sindi SS, Raphael BJ, Caceres M, Gonzalez JR. Identification of polymorphic inversions from genotypes. BMC Bioinformatics. 2012;13:28.

66. Rohlfs RV, Swanson WJ, Weir BS. Detecting coevolution through allelic association between physically unlinked loci. American journal of human genetics. 2010;86(5):674-85.

67. Hartman JLt, Garvik B, Hartwell L. Principles for the buffering of genetic variation. Science. 2001;291(5506):1001-4.

68. Phillips PC. Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. Nature reviews Genetics. 2008;9(11):855-67.

69. Beissinger TM, Gholami M, Erbe M, Weigend S, Weigend A, de Leon N, et al. Using the variability of linkage disequilibrium between subpopulations to infer sweeps and epistatic selection in a diverse panel of chickens. Heredity (Edinb). 2016;116(2):158-66.

70. Musso G, Costanzo M, Huangfu M, Smith AM, Paw J, San Luis BJ, et al. The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. Genome Res. 2008;18(7):1092-9.

71. Bloom JS, Ehrenreich IM, Loo WT, Lite TL, Kruglyak L. Finding the sources of missing heritability in a yeast cross. Nature. 2013;494(7436):234-7.

72. Lehner B, Crombie C, Tischler J, Fortunato A, Fraser AG. Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. Nature genetics. 2006;38(8):896-903.

73. Huang W, Richards S, Carbone MA, Zhu D, Anholt RR, Ayroles JF, et al. Epistasis dominates the genetic architecture of Drosophila quantitative traits. Proceedings of the National Academy of Sciences of the United States of America. 2012;109(39):15553-9.

74. Mackay TF. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. Nature reviews Genetics. 2014;15(1):22-33.

75. Huang W, Mackay TF. The Genetic Architecture of Quantitative Traits Cannot Be Inferred from Variance Component Analysis. PLoS genetics. 2016;12(11):e1006421.

76. Liu Y, Xu H, Chen S, Chen X, Zhang Z, Zhu Z, et al. Genome-wide interaction-based association analysis identified multiple new susceptibility Loci for common diseases. PLoS genetics. 2011;7(3):e1001338.

77. Hu T, Chen Y, Kiralis JW, Collins RL, Wejse C, Sirugo G, et al. An information-gain approach to detecting three-way epistatic interactions in genetic association studies. J Am Med Inform Assoc. 2013;20(4):630-6.

78. Kirino Y, Bertsias G, Ishigatsubo Y, Mizuki N, Tugal-Tutkun I, Seyahi E, et al. Genome-wide association analysis identifies new susceptibility loci for Behcet's disease and epistasis between HLA-B*51 and ERAP1. Nature genetics. 2013;45(2):202-7.

79. Hemani G, Shakhbazov K, Westra HJ, Esko T, Henders AK, McRae AF, et al. Detection and replication of epistasis influencing transcription in humans. Nature. 2014;508(7495):249-53.

80. Huang Y, Wang C, Yao Y, Zuo X, Chen S, Xu C, et al. Molecular Basis of Gene-Gene Interaction: Cyclic Cross-Regulation of Gene Expression and Post-GWAS Gene-Gene Interaction Involved in Atrial Fibrillation. PLoS genetics. 2015;11(8):e1005393.

81. Verma SS, Cooke Bailey JN, Lucas A, Bradford Y, Linneman JG, Hauser MA, et al. Epistatic Gene-Based Interaction Analyses for Glaucoma in eMERGE and NEIGHBOR Consortium. PLoS genetics. 2016;12(9):e1006186.

82. Galarza-Munoz G, Briggs FB, Evsyukova I, Schott-Lerner G, Kennedy EM, Nyanhete T, et al. Human Epistatic Interaction Controls IL7R Splicing and Increases Multiple Sclerosis Risk. Cell. 2017;169(1):72-84 e13.

83. Sucheston L, Witonsky DB, Hastings D, Yildiz O, Clark VJ, Di Rienzo A, et al. Natural selection and functional genetic variation in the p53 pathway. Human molecular genetics. 2011;20(8):1502-8.

84. Sturm RA, Duffy DL. Human pigmentation genes under environmental selection. Genome Biol. 2012;13(9):248.

85. Bigham AW, Lee FS. Human high-altitude adaptation: forward genetics meets the HIF pathway. Genes Dev. 2014;28(20):2189-204.

86. Daub JT, Moretti S, Davydov, II, Excoffier L, Robinson-Rechavi M. Detection of Pathways Affected by Positive Selection in Primate Lineages Ancestral to Humans. Mol Biol Evol. 2017;34(6):1391-402.

87. Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, Moutsianas L, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature. 2011;476(7359):214-9.

88. Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. American journal of human genetics. 2007;81(6):1278-83.

89. Merico D, Isserlin R, Stueker O, Emili A, Bader GD. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PloS one. 2010;5(11):e13984.

90. Segal R. The Black Diaspora: Five Centuries of Black Experience Outside Africa. New York: Farrar, Straus and Giroux; 1995.

91. Thomas H. The Slave Trade: The Story of the Atlantic Slave Trade: 1440-1870: Simon & Schuster; 1999.

92. Eltis D. The volume and structure of the transatlantic slave trade: a reassessment. William Mary Q. 2001;58(1):17-46.

93. Zakharia F, Basu A, Absher D, Assimes TL, Go AS, Hlatky MA, et al. Characterizing the admixed African ancestry of African Americans. Genome Biol. 2009;10(12):R141.

94. Pengelly RJ, Tapper W, Gibson J, Knut M, Tearle R, Collins A, et al. Whole genome sequences are required to fully resolve the linkage disequilibrium structure of human populations. BMC Genomics. 2015;16:666.

95. Quach H, Quintana-Murci L. Living in an adaptive world: Genomic dissection of the genus Homo and its immune response. J Exp Med. 2017;214(4):877-94.

96. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, et al. AmiGO: online access to ontology and annotation data. Bioinformatics. 2009;25(2):288-9.

97. Jagannathan-Bogdan M, Zon LI. Hematopoiesis. Development. 2013;140(12):2463-7.

98. Hsia N, Zon LI. Transcriptional regulation of hematopoietic stem cell development in zebrafish. Exp Hematol. 2005;33(9):1007-14.

99. Palis J, Yoder MC. Yolk-sac hematopoiesis: the first blood cells of mouse and man. Exp Hematol. 2001;29(8):927-36.

100. Tavian M, Biasch K, Sinka L, Vallet J, Peault B. Embryonic origin of human hematopoiesis. Int J Dev Biol. 2010;54(6-7):1061-5.

101. Hoffmann JA, Kafatos FC, Janeway CA, Ezekowitz RA. Phylogenetic perspectives in innate immunity. Science. 1999;284(5418):1313-8.

102. Valanne S, Wang JH, Ramet M. The Drosophila Toll signaling pathway. J Immunol. 2011;186(2):649-56.

103. Takeda K, Kaisho T, Akira S. Toll-like receptors. Annu Rev Immunol. 2003;21:335-76.

104. Bachmann MF, Kopf M. On the role of the innate immunity in autoimmune disease. J Exp Med. 2001;193(12):F47-50.

105. Zhang H, Meltzer P, Davis S. RCircos: an R package for Circos 2D track plots. BMC Bioinformatics. 2013;14:244.

106. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, et al. Whole-genome patterns of common DNA variation in three human populations. Science. 2005;307(5712):1072-9.

107. Li MJ, Wang LY, Xia Z, Wong MP, Sham PC, Wang J. dbPSHP: a database of recent positive selection across human populations. Nucleic acids research. 2014;42(Database issue):D910-6.

108. Montinaro F, Busby GB, Pascali VL, Myers S, Hellenthal G, Capelli C. Unravelling the hidden ancestry of American admixed populations. Nat Commun. 2015;6:6596.

109. Vitti JJ, Grossman SR, Sabeti PC. Detecting natural selection in genomic data. Annu Rev Genet. 2013;47:97-120.

110. Casanova JL, Abel L, Quintana-Murci L. Immunology taught by human genetics. Cold Spring Harb Symp Quant Biol. 2013;78:157-72.

111. Vannberg FO, Chapman SJ, Hill AV. Human genetic susceptibility to intracellular pathogens. Immunol Rev. 2011;240(1):105-16.

112. Parkes M, Cortes A, van Heel DA, Brown MA. Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nature reviews Genetics. 2013;14(9):661-73.

113. Abel L, Alcais A, Schurr E. The dissection of complex susceptibility to infectious disease: bacterial, viral and parasitic infections. Curr Opin Immunol. 2014;30:72-8.

114. Fairfax BP, Knight JC. Genetics of gene expression in immunity to infection. Curr Opin Immunol. 2014;30:63-71.

115. Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, Cheung VG. Common genetic variants account for differences in gene expression among ethnic groups. Nature genetics. 2007;39(2):226-31.

116. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, et al. Patterns of cis regulatory variation in diverse human populations. PLoS genetics. 2012;8(4):e1002639.

117. Raj T, Rothamel K, Mostafavi S, Ye C, Lee MN, Replogle JM, et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science. 2014;344(6183):519-23.

118. Nedelec Y, Sanz J, Baharian G, Szpiech ZA, Pacis A, Dumaine A, et al. Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. Cell. 2016;167(3):657-69 e21.

119. Quach H, Rotival M, Pothlichet J, Loh YE, Dannemann M, Zidane N, et al. Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. Cell. 2016;167(3):643-56 e17.

120. Hoban MD, Cost GJ, Mendel MC, Romero Z, Kaufman ML, Joglekar AV, et al. Correction of the sickle cell disease mutation in human hematopoietic stem/progenitor cells. Blood. 2015;125(17):2597-604.

121. National Center for Biotechnology Information. Genes and Disease [Internet]. Bethesda, Maryland: National Center for Biotechnology Information; 1998.

122. Barreiro LB, Tailleux L, Pai AA, Gicquel B, Marioni JC, Gilad Y. Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. Proceedings of the National Academy of Sciences of the United States of America. 2012;109(4):1204-9.

123. Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science. 2014;343(6175):1246949.

124. Lee MN, Ye C, Villani AC, Raj T, Li W, Eisenhaure TM, et al. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. Science. 2014;343(6175): 1246980.

125. Ter Horst R, Jaeger M, Smeekens SP, Oosting M, Swertz MA, Li Y, et al. Host and Environmental Factors Influencing Individual Human Cytokine Responses. Cell. 2016;167(4): 1111-24 e13.

126. Hughes T, Adler A, Kelly JA, Kaufman KM, Williams AH, Langefeld CD, et al. Evidence for gene-gene epistatic interactions among susceptibility loci for systemic lupus erythematosus. Arthritis Rheum. 2012;64(2):485-92.

127. Bustamante MF, Fissolo N, Rio J, Espejo C, Costa C, Mansilla MJ, et al. Implication of the Toll-like receptor 4 pathway in the response to interferon-beta in multiple sclerosis. Ann Neurol. 2011;70(4):634-45.

128. Miranda-Hernandez S, Baxter AG. Role of toll-like receptors in multiple sclerosis. Am J Clin Exp Immunol. 2013;2(1):75-93.

129. Hossain MJ, Tanasescu R, Gran B. TLR2: an innate immune checkpoint in multiple sclerosis. Oncotarget. 2015;6(34):35131-2.

130. Rose AM, Bell LC. Epistasis and immunity: the role of genetic interactions in autoimmune diseases. Immunology. 2012;137(2):131-8.

131. Galanter JM, Gignoux CR, Oh SS, Torgerson D, Pino-Yanes M, Thakur N, et al. Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures. Elife. 2017;6.

132. Weigelt B, Reis-Filho JS. Epistatic interactions and drug response. J Pathol. 2014;232(2): 255-63.

133. Wong A. Epistasis and the Evolution of Antimicrobial Resistance. Front Microbiol. 2017;8:246.

134. Sved JA. The covariance of heterozygosity as a measure of linkage disequilibrium between blocks of linked and unlinked sites in Hapmap. Genet Res (Camb). 2011;93(4):285-90.

135. Black WCt, Baer CF, Antolin MF, DuTeau NM. Population genomics: genome-wide sampling of insect populations. Annu Rev Entomol. 2001;46:441-69.

136. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449(7164):851-61.

137. Ma J, Amos CI. Principal components analysis of population admixture. PloS one. 2012;7(7):e40115.

138. Chen GK, Marjoram P, Wall JD. Fast and flexible simulation of DNA sequence data. Genome Res. 2009;19(1):136-42.

139. Pickrell JK, Reich D. Toward a new history and geography of human genes informed by ancient DNA. Trends in genetics : TIG. 2014;30(9):377-89.

140. Mai CW, Chung FF, Leong CO. Targeting Legumain As a Novel Therapeutic Strategy in Cancers. Curr Drug Targets. 2017;18(11):1259-68.

141. Barreiro LB, Ben-Ali M, Quach H, Laval G, Patin E, Pickrell JK, et al. Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. PLoS genetics. 2009;5(7):e1000562.

142. Netea MG, Wijmenga C, O'Neill LA. Genetic variation in Toll-like receptors and disease susceptibility. Nat Immunol. 2012;13(6):535-42.

143. Milacic M, Haw R, Rothfels K, Wu G, Croft D, Hermjakob H, et al. Annotating cancer variants and anti-cancer therapeutics in reactome. Cancers (Basel). 2012;4(4):1180-211.

144. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The Reactome pathway Knowledgebase. Nucleic acids research. 2016;44(D1):D481-7.

145. Wlasiuk G, Nachman MW. Adaptation and constraint at Toll-like receptors in primates. Mol Biol Evol. 2010;27(9):2172-86.

146. Medvedev AE. Toll-like receptor polymorphisms, inflammatory and infectious diseases, allergies, and cancer. J Interferon Cytokine Res. 2013;33(9):467-84.

147. Ferwerda B, McCall MB, Verheijen K, Kullberg BJ, van der Ven AJ, Van der Meer JW, et al. Functional consequences of toll-like receptor 4 polymorphisms. Mol Med. 2008;14(5-6):346-52.

148. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, et al. Linkage disequilibrium in the human genome. Nature. 2001;411(6834):199-204.

149. International HapMap Consortium, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52-8.

150. Christoforou A, Espeseth T, Davies G, Fernandes CP, Giddaluru S, Mattheisen M, et al. GWAS-based pathway analysis differentiates between fluid and crystallized intelligence. Genes Brain Behav. 2014;13(7):663-74.

151. Shim U, Kim HN, Lee H, Oh JY, Sung YA, Kim HL. Pathway Analysis Based on a Genome-Wide Association Study of Polycystic Ovary Syndrome. PloS one. 2015;10(8):e0136609.

152. Michailidou K, Lindstrom S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017;551(7678):92-4.

153. Pai S, Hui S, Isserlin R, Shah MA, Kaka H, Bader GD. netDx: Interpretable patient classification using integrated patient similarity networks. bioRxiv. 2017.

154. Wang W, Xu ZZ, Costanzo M, Boone C, Lange CA, Myers CL. Pathway-based discovery of genetic interactions in breast cancer. PLoS genetics. 2017;13(9):e1006973.

155. Kelley R, Ideker T. Systematic interpretation of genetic interactions using protein networks. Nature biotechnology. 2005;23(5):561-6.

156. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2017.

157. Purcell SM, Chang CC. PLINK 1.9.

158. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.

159. Satterthwaite TD, Connolly JJ, Ruparel K, Calkins ME, Jackson C, Elliott MA, et al. The Philadelphia Neurodevelopmental Cohort: A publicly available resource for the study of normal and abnormal brain development in youth. NeuroImage. 2016;124(Pt B):1115-9.

160. Pathway Analysis Subgroup of Psychiatric Genomics Consortium. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. Nature neuroscience. 2015;18(2):199-209.

161. Holmans P. Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. Adv Genet. 2010;72:141-79.

162. Elbers CC, van Eijk KR, Franke L, Mulder F, van der Schouw YT, Wijmenga C, et al. Using genome-wide pathway analysis to unravel the etiology of complex diseases. Genet Epidemiol. 2009;33(5):419-31.

163. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. 2007;81(3):559-75.

164. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498-504.

165. Kucera M, Isserlin R, Arkhangorodsky A, Bader GD. AutoAnnotate: A Cytoscape app for summarizing networks with semantic annotations. F1000Res. 2016;5:1717.

166. Clayton D. snpStats: SnpMatrix and XSnpMatrix classes and methods. 1.27.0 ed: R package; 2015.

167. Clayton D, Leung HT. An R package for analysis of whole-genome association studies. Hum Hered. 2007;64(1):45-51.

168. Cadzow M, Boocock J, Nguyen HT, Wilcox P, Merriman TR, Black MA. A bioinformatics workflow for detecting signatures of selection in genomic data. Front Genet. 2014;5:293.

169. Sergushichev A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. bioRxiv. 2016.

7. Appendix

Supplementary Table 1. Significant GSEA pathway enrichment results using the HM3 dataset.

Pathway	Size	ES	NES	NominalP	FDR
SEX DIFFERENTIATION	84	0.362	4.072	0.000	0.026
DEVELOPMENT OF PRIMARY SEXUAL CHARACTERISTICS	58	0.405	4.024	0.000	0.027
PROTEIN DEACETYLATION	48	0.406	3.985	0.000	0.028
CELL FATE COMMITMENT	109	0.372	4.138	0.000	0.029
MYELOID LEUKOCYTE ACTIVATION	60	0.414	4.078	0.000	0.029
REGULATION OF HEMOPOIESIS	183	0.304	3.956	0.000	0.029
ID	26	0.537	3.931	0.001	0.029
HISTONE DEACETYLATION	44	0.440	4.286	0.000	0.030
MACROMOLECULE DEACYLATION	54	0.389	4.152	0.000	0.033
PROTEIN DEACYLATION	53	0.395	4.188	0.000	0.035
REPRODUCTIVE STRUCTURE DEVELOPMENT	103	0.376	3.832	0.000	0.035
REPRODUCTIVE SYSTEM DEVELOPMENT	103	0.376	3.832	0.000	0.035
ENERGY DERIVATION BY OXIDATION OF ORGANIC COMPOUNDS	148	0.267	3.812	0.000	0.035
REGULATION OF CAMP METABOLIC PROCESS	85	0.431	3.799	0.001	0.035
GONAD DEVELOPMENT	57	0.399	3.775	0.000	0.036
PROTEIN COMPLEX DISASSEMBLY	157	0.226	3.742	0.001	0.037
REGULATION OF LEUKOCYTE DIFFERENTIATION	123	0.327	3.711	0.000	0.039
CELLULAR SENESCENCE	27	0.526	3.682	0.000	0.041
CELLULAR RESPIRATION	105	0.258	3.65	0.000	0.041
PURINE NUCLEOSIDE TRIPHOSPHATE METABOLIC PROCESS	139	0.231	3.639	0.000	0.041
NOTCH SIGNALING PATHWAY	46	0.439	3.625	0.000	0.041
LYMPHOCYTE ACTIVATION	183	0.321	3.652	0.000	0.043

NUCLEOSIDE TRIPHOSPHATE METABOLIC PROCESS	155	0.248	4.499	0.000	0.044
TP53 REGULATES METABOLIC GENES	73	0.363	4.289	0.000	0.045
D-<I>MYO< I>-INOSITOL (1,4,5)-TRISPHOSPHATE BIOSYNTHESIS	24	0.567	3.555	0.001	0.047
APPENDAGE DEVELOPMENT	57	0.442	3.528	0.000	0.047
LIMB DEVELOPMENT	57	0.442	3.528	0.000	0.047
REGULATION OF CYCLIC NUCLEOTIDE METABOLIC PROCESS	102	0.387	3.517	0.002	0.047
SENSORY ORGAN MORPHOGENESIS	88	0.419	3.54	0.000	0.048
MACROMOLECULAR COMPLEX DISASSEMBLY	165	0.219	3.559	0.001	0.049
HORMONE TRANSPORT	35	0.438	3.466	0.001	0.054
TRANSMEMBRANE RECEPTOR PROTEIN SERINE/THREONINE KINASE SIGNALING PATHWAY	135	0.347	3.429	0.001	0.058
PLASMA MEMBRANE ORGANIZATION	156	0.387	3.397	0.002	0.059
PURINE RIBONUCLEOSIDE TRIPHOSPHATE METABOLIC PROCESS	135	0.232	3.404	0.000	0.060
POLYOL METABOLIC PROCESS	62	0.412	3.408	0.000	0.061
DIGESTIVE TRACT DEVELOPMENT	55	0.424	3.367	0.000	0.063
ISOPRENOID METABOLIC PROCESS	89	0.338	3.331	0.002	0.068
TOLL-LIKE RECEPTOR SIGNALING PATHWAY	74	0.374	3.293	0.004	0.070
VACUOLE ORGANIZATION	137	0.294	3.272	0.001	0.070
CELLULAR RESPONSE TO BMP STIMULUS	59	0.410	3.274	0.001	0.071
RESPONSE TO BMP	59	0.410	3.274	0.001	0.071
REGULATION OF MYELOID LEUKOCYTE DIFFERENTIATION	61	0.368	3.294	0.001	0.072
NEGATIVE REGULATION OF PROTEIN KINASE B SIGNALING	26	0.502	3.296	0.001	0.073
BMP SIGNALING PATHWAY	55	0.412	3.229	0.001	0.075
DIGESTIVE SYSTEM DEVELOPMENT	58	0.404	3.238	0.001	0.076

RIBONUCLEOSIDE TRIPHOSPHATE METABOLIC PROCESS	136	0.226	3.232	0.000	0.076
RHYTHMIC PROCESS	101	0.360	3.217	0.002	0.076
DEVELOPMENT OF PRIMARY MALE SEXUAL CHARACTERISTICS	33	0.443	3.186	0.002	0.077
MALE GONAD DEVELOPMENT	33	0.443	3.186	0.002	0.077
REGULATION OF NUCLEOTIDE METABOLIC PROCESS	141	0.362	3.171	0.001	0.077
REGULATION OF CAMP BIOSYNTHETIC PROCESS	76	0.433	3.194	0.001	0.078
MTOR SIGNALLING	37	0.410	3.172	0.001	0.078
REGULATION OF PURINE NUCLEOTIDE METABOLIC PROCESS	131	0.371	3.159	0.002	0.078
ATP METABOLIC PROCESS	131	0.229	3.197	0.000	0.079
REGULATION OF TOR SIGNALING	58	0.400	3.14	0.001	0.082
NUCLEOSIDE MONOPHOSPHATE METABOLIC PROCESS	169	0.236	3.128	0.002	0.083
LIPOSACCHARIDE METABOLIC PROCESS	91	0.346	3.114	0.004	0.085
CIRCADIAN REGULATION OF GENE EXPRESSION	49	0.417	3.105	0.004	0.087
NEGATIVE REGULATION OF SEQUENCE-SPECIFIC DNA BINDING TRANSCRIPTION FACTOR ACTIVITY	113	0.334	3.087	0.002	0.087
PI3K CASCADE	66	0.358	3.081	0.001	0.087
INTRINSIC APOPTOTIC SIGNALING PATHWAY IN RESPONSE TO DNA DAMAGE BY P53 CLASS MEDIATOR	21	0.508	3.092	0.000	0.088
LYMPHOCYTE DIFFERENTIATION	99	0.350	3.089	0.000	0.088
PEPTIDE CHAIN ELONGATION	72	0.118	3.06	0.019	0.091
ODONTOGENESIS	41	0.409	3.047	0.002	0.093
EUKARYOTIC TRANSLATION ELONGATION	76	0.120	3.041	0.013	0.094
REGULATION OF ADENYLATE CYCLASE ACTIVITY	48	0.498	3.029	0.000	0.096

REGULATION OF BONE MINERALIZATION	46	0.458	2.995	0.004	0.098
MYD88-INDEPENDENT TOLL-LIKE RECEPTOR SIGNALING PATHWAY	28	0.438	3.014	0.000	0.099
NEGATIVE REGULATION OF LEUKOCYTE DIFFERENTIATION	47	0.371	3.006	0.003	0.099
MALE SEX DIFFERENTIATION	38	0.411	3.003	0.000	0.099
NEGATIVE REGULATION OF HEMOPOIESIS	80	0.317	2.998	0.001	0.099
TERPENOID METABOLIC PROCESS	73	0.356	2.932	0.002	0.099
MORPHOGENESIS OF AN EPITHELIUM	168	0.327	2.984	0.002	0.100
REGULATION OF BIOMINERAL TISSUE DEVELOPMENT	51	0.436	2.955	0.007	0.100
EUKARYOTIC TRANSLATION TERMINATION	75	0.117	2.939	0.018	0.100
NOTCH SIGNALING PATHWAY	56	0.439	2.934	0.006	0.100

Note—A total of 76 pathways are displayed at FDR ≤ 0.1. **Abbreviations**: ES, enrichment score; NES, normalized enrichment score; NominalP, nominal *p* value; FDR, false discovery rate.

Supplementary Table 2. Significant GSEA pathway enrichment results using the PNC dataset.

Pathway	Size	ES	NES	NominalP	FDR
LYMPHOCYTE AGGREGATION	117	0.330	6.005	0.000	0.000
SIRT1 NEGATIVELY REGULATES RRNA EXPRESSION	51	0.214	6.001	0.000	0.000
T CELL ACTIVATION	116	0.329	5.931	0.000	0.000
T CELL AGGREGATION	116	0.329	5.931	0.000	0.000
RNA POLYMERASE I PROMOTER OPENING	47	0.156	5.706	0.000	0.000
DNA METHYLATION	49	0.181	5.697	0.000	0.000
TCF DEPENDENT SIGNALING IN RESPONSE TO WNT	188	0.249	5.654	0.000	0.000
RESPONSE TO TOXIC SUBSTANCE	84	0.331	5.535	0.000	0.000
LEUKOCYTE AGGREGATION	123	0.314	5.390	0.000	0.000
REGULATION OF NEURON DEATH	103	0.367	5.375	0.000	0.000
LYMPHOCYTE ACTIVATION	173	0.293	5.236	0.000	0.000
LEUKOCYTE CELL-CELL ADHESION	145	0.286	4.893	0.000	0.001
CARBOHYDRATE DERIVATIVE CATABOLIC PROCESS	119	0.340	4.884	0.000	0.001
CELLULAR RESPIRATION	98	0.284	4.770	0.000	0.001
ASPARTATE FAMILY AMINO ACID METABOLIC PROCESS	45	0.395	4.725	0.000	0.002
POSITIVE REGULATION OF CELLULAR PROTEIN CATABOLIC PROCESS	144	0.277	4.717	0.000	0.002
POSITIVE REGULATION OF PROTEOLYSIS INVOLVED IN CELLULAR PROTEIN CATABOLIC PROCESS	137	0.278	4.669	0.000	0.002
PROTEIN COMPLEX DISASSEMBLY	138	0.238	4.654	0.000	0.002
G ALPHA (I) SIGNALLING EVENTS	177	0.246	4.569	0.000	0.002
PROTEIN TETRAMERIZATION	84	0.307	4.565	0.000	0.002
REGULATION OF NEURON APOPTOTIC PROCESS	73	0.353	4.549	0.000	0.002
REGULATION OF MYELOID CELL DIFFERENTIATION	101	0.280	4.518	0.000	0.003
NICOTINAMIDE NUCLEOTIDE METABOLIC PROCESS	58	0.355	4.428	0.000	0.003
PYRIDINE NUCLEOTIDE METABOLIC PROCESS	58	0.355	4.428	0.000	0.003
ENERGY DERIVATION BY OXIDATION OF ORGANIC COMPOUNDS	135	0.271	4.415	0.000	0.003

PACKAGING OF TELOMERE ENDS	39	0.171	4.402	0.003	0.003
LYSOSOMAL TRANSPORT	54	0.436	4.398	0.000	0.003
CELL CYCLE CHECKPOINTS	171	0.220	4.387	0.000	0.003
ACTIVATION OF INNATE IMMUNE RESPONSE	171	0.263	4.345	0.000	0.003
CYTOKINE PRODUCTION	66	0.349	4.334	0.000	0.003
RESPONSE TO VIRUS	128	0.233	4.331	0.000	0.003
NEGATIVE REGULATION OF SECRETION	85	0.330	4.329	0.000	0.003
VACUOLE ORGANIZATION	119	0.323	4.328	0.000	0.003
REGULATION OF PROTEOLYSIS INVOLVED IN CELLULAR PROTEIN CATABOLIC PROCESS	197	0.257	4.300	0.000	0.003
GENERATION OF PRECURSOR METABOLITES AND ENERGY	196	0.216	4.292	0.000	0.003
REGULATION OF CALCIUM ION TRANSPORT	115	0.377	4.260	0.000	0.004
REGULATION OF ION TRANSMEMBRANE TRANSPORT	196	0.366	4.209	0.000	0.005
HEXOSE METABOLIC PROCESS	83	0.308	4.177	0.000	0.005
LYMPHOCYTE DIFFERENTIATION	93	0.318	4.146	0.000	0.005
NEGATIVE REGULATION OF SECRETION BY CELL	77	0.323	4.135	0.000	0.005
INNATE IMMUNE RESPONSE-ACTIVATING SIGNAL TRANSDUCTION	164	0.267	4.130	0.000	0.005
NONSENSE MEDIATED DECAY (NMD) INDEPENDENT OF THE EXON JUNCTION COMPLEX (EJC)	64	0.126	4.120	0.004	0.005
POSITIVE REGULATION OF PROTEIN CATABOLIC PROCESS	192	0.255	4.115	0.000	0.005
ACTIVATION OF RRNA EXPRESSION BY ERCC6 (CSB) AND EHMT2 (G9A)	57	0.127	4.104	0.008	0.005
REGULATION OF PROTEIN MODIFICATION BY SMALL PROTEIN CONJUGATION OR REMOVAL	200	0.237	4.082	0.001	0.006
POSITIVE REGULATION OF PROTEASOMAL UBIQUITIN-DEPENDENT PROTEIN CATABOLIC PROCESS	62	0.378	4.077	0.000	0.006
PEPTIDE LIGAND-BINDING RECEPTORS	150	0.221	4.070	0.000	0.006
PROTEIN SECRETION	64	0.322	4.059	0.000	0.006
AMINE LIGAND-BINDING RECEPTORS	33	0.449	4.022	0.000	0.007
ENDOSOME ORGANIZATION	46	0.404	4.010	0.000	0.007
B CELL ACTIVATION	74	0.326	4.004	0.001	0.007

BEHAVIOR	132	0.367	3.973	0.000	0.007
NEGATIVE REGULATION OF TRANSMEMBRANE TRANSPORT	56	0.377	3.965	0.001	0.007
MACROMOLECULAR COMPLEX DISASSEMBLY	146	0.225	3.937	0.000	0.007
OSTEOBLAST DIFFERENTIATION	65	0.329	3.946	0.000	0.008
ENDOSOME TO LYSOSOME TRANSPORT	30	0.478	3.931	0.000	0.008
ACTIVATED PKN1 STIMULATES TRANSCRIPTION OF AR (ANDROGEN RECEPTOR) REGULATED GENES KLK2 AND KLK3	50	0.150	3.890	0.003	0.009
LEARNING OR MEMORY	56	0.513	3.838	0.000	0.010
B CELL ACTIVATION	178	0.319	3.834	0.001	0.010
PROTEIN HOMOTETRAMERIZATION	35	0.449	3.831	0.000	0.010
BIOSYNTHESIS OF THE N-GLYCAN PRECURSOR (DOLICHOL LIPID-LINKED OLIGOSACCHARIDE, LLO) AND TRANSFER TO A NASCENT PROTEIN	65	0.372	3.814	0.000	0.010
AEROBIC RESPIRATION	36	0.420	3.808	0.000	0.010
LIPOSACCHARIDE METABOLIC PROCESS	84	0.349	3.807	0.000	0.010
RNA POLYMERASE I CHAIN ELONGATION	71	0.189	3.800	0.001	0.010
FOXO FAMILY SIGNALING	47	0.405	3.785	0.000	0.010
REGULATION OF METAL ION TRANSPORT	194	0.350	3.780	0.001	0.010
NUCLEOSIDE TRIPHOSPHATE METABOLIC PROCESS	140	0.219	3.773	0.000	0.011
OSSIFICATION	102	0.300	3.767	0.000	0.011
OXIDOREDUCTION COENZYME METABOLIC PROCESS	67	0.329	3.753	0.000	0.011
PYRIDINE-CONTAINING COMPOUND METABOLIC PROCESS	71	0.312	3.709	0.000	0.011
DNA REPLICATION-DEPENDENT NUCLEOSOME ASSEMBLY	28	0.184	3.705	0.008	0.011
DNA REPLICATION-DEPENDENT NUCLEOSOME ORGANIZATION	28	0.184	3.705	0.008	0.011
REGULATION OF NUCLEAR BETA CATENIN SIGNALING AND TARGET GENE TRANSCRIPTION	66	0.382	3.730	0.001	0.012
G ALPHA (S) SIGNALLING EVENTS	119	0.325	3.729	0.000	0.012
METABOLISM OF FAT-SOLUBLE VITAMINS	42	0.367	3.722	0.000	0.012
REGULATION OF PROTEIN UBIQUITINATION	185	0.233	3.722	0.001	0.012
CLASS I PI3K SIGNALING EVENTS MEDIATED BY AKT	32	0.454	3.716	0.000	0.012

REGULATION OF PHOSPHOPROTEIN PHOSPHATASE ACTIVITY	42	0.445	3.714	0.000	0.012
PURINE NUCLEOSIDE METABOLIC PROCESS	184	0.215	3.691	0.001	0.012
REGULATION OF CELLULAR KETONE METABOLIC PROCESS	112	0.258	3.687	0.000	0.012
DOWNSTREAM SIGNALING EVENTS OF B CELL RECEPTOR (BCR)	154	0.306	3.680	0.000	0.012
REGULATION OF GENE EXPRESSION, EPIGENETIC	171	0.215	3.661	0.000	0.012
CELLULAR PROTEIN COMPLEX DISASSEMBLY	97	0.242	3.655	0.000	0.012
REGULATION OF NEUROTRANSMITTER LEVELS	113	0.367	3.650	0.000	0.012
POSITIVE REGULATION OF BONE MINERALIZATION	27	0.553	3.645	0.000	0.012
GLYCOLIPID METABOLIC PROCESS	82	0.350	3.633	0.000	0.013
G2 M CHECKPOINTS	144	0.218	3.631	0.000	0.013
REGULATION OF PROTEASOMAL UBIQUITIN-DEPENDENT PROTEIN CATABOLIC PROCESS	111	0.312	3.625	0.000	0.013
EUKARYOTIC TRANSLATION TERMINATION	62	0.115	3.622	0.007	0.013
PURINE RIBONUCLEOSIDE METABOLIC PROCESS	181	0.210	3.600	0.003	0.014
COGNITION	81	0.455	3.597	0.002	0.014
NEGATIVE REGULATION OF ION TRANSMEMBRANE TRANSPORT	52	0.382	3.594	0.002	0.014
ANION TRANSMEMBRANE TRANSPORT	162	0.337	3.576	0.001	0.014
GLUCOSE METABOLIC PROCESS	63	0.328	3.563	0.000	0.015
CELL RECOGNITION	73	0.369	3.544	0.001	0.016
NEURON-NEURON SYNAPTIC TRANSMISSION	23	0.562	3.540	0.001	0.016
NEGATIVE REGULATION OF NEURON DEATH	59	0.341	3.537	0.002	0.016
LEARNING	29	0.579	3.531	0.001	0.016
INNATE IMMUNE RESPONSE ACTIVATING CELL SURFACE RECEPTOR SIGNALING PATHWAY	90	0.282	3.470	0.001	0.019
CELLULAR MODIFIED AMINO ACID METABOLIC PROCESS	140	0.240	3.464	0.000	0.019
INFLAMMATORY RESPONSE	194	0.194	3.455	0.000	0.019
CELLULAR SENESCENCE	158	0.198	3.446	0.000	0.019
GLYCOSPHINGOLIPID METABOLIC PROCESS	49	0.446	3.448	0.002	0.020
REGULATION OF HEMOPOIESIS	168	0.223	3.411	0.001	0.021

STIMULATORY C-TYPE LECTIN RECEPTOR SIGNALING PATHWAY	88	0.284	3.386	0.001	0.022
AMINOGLYCAN METABOLIC PROCESS	116	0.316	3.382	0.001	0.022
CELLULAR RESPONSE TO DECREASED OXYGEN LEVELS	79	0.288	3.378	0.001	0.022
NEGATIVE REGULATION OF ION TRANSPORT	69	0.360	3.375	0.001	0.022
CARBOXYLIC ACID TRANSMEMBRANE TRANSPORT	65	0.388	3.372	0.001	0.022
APOPTOSIS	140	0.259	3.363	0.000	0.022
3' -UTR-MEDIATED TRANSLATIONAL REGULATION	77	0.154	3.358	0.000	0.022
L13A-MEDIATED TRANSLATIONAL SILENCING OF CERULOPLASMIN EXPRESSION	77	0.154	3.358	0.000	0.022
REGULATION OF CELLULAR RESPONSE TO GROWTH FACTOR STIMULUS	138	0.281	3.389	0.002	0.023
LEUKOCYTE MIGRATION	181	0.229	3.387	0.000	0.023
REGULATION OF SYNAPSE STRUCTURE OR ACTIVITY	76	0.403	3.365	0.000	0.023
ROLE OF CALCINEURIN-DEPENDENT NFAT SIGNALING IN LYMPHOCYTES	50	0.400	3.358	0.001	0.023
PRC2 METHYLATES HISTONES AND DNA	56	0.228	3.350	0.000	0.023
PROGRAMMED CELL DEATH	143	0.253	3.335	0.001	0.023
LOCOMOTORY BEHAVIOR	28	0.445	3.335	0.000	0.023
CELL-CELL ADHESION VIA PLASMA-MEMBRANE ADHESION MOLECULES	89	0.402	3.334	0.000	0.023
RESPONSE TO LIGHT STIMULUS	152	0.278	3.320	0.001	0.024
CHROMATIN SILENCING	62	0.127	3.318	0.006	0.024
NICOTINIC ACETYLCHOLINE RECEPTOR SIGNALING PATHWAY	52	0.346	3.316	0.000	0.024
ORGANIC ACID TRANSMEMBRANE TRANSPORT	69	0.385	3.303	0.001	0.024
NEGATIVE REGULATION OF PROTEIN COMPLEX ASSEMBLY	54	0.387	3.302	0.001	0.024
POSITIVE REGULATION OF PROTEASOMAL PROTEIN CATABOLIC PROCESS	73	0.310	3.302	0.000	0.024
METHYLATION	162	0.268	3.303	0.000	0.025
A6B1 AND A6B4 INTEGRIN SIGNALING	43	0.426	3.292	0.000	0.025
NEURON MIGRATION	33	0.540	3.284	0.003	0.025
RESPONSE TO STEROID HORMONE	144	0.269	3.273	0.003	0.025
CIRCADIAN RHYTHM	73	0.322	3.273	0.001	0.025

TRANSPORT OF INORGANIC CATIONS ANIONS AND AMINO ACIDS OLIGOPEPTIDES	87	0.365	3.272	0.001	0.025
EMBRYONIC ORGAN DEVELOPMENT	126	0.252	3.269	0.001	0.025
RHYTHMIC PROCESS	93	0.314	3.257	0.000	0.026
RESPONSE TO TUMOR NECROSIS FACTOR	157	0.189	3.256	0.002	0.026
PROTEIN ALKYLATION	75	0.309	3.248	0.005	0.026
PROTEIN METHYLATION	75	0.309	3.248	0.005	0.026
NEGATIVE REGULATION OF PROTEIN MODIFICATION BY SMALL PROTEIN CONJUGATION OR REMOVAL	101	0.257	3.245	0.001	0.026
NEUROTRANSMITTER TRANSPORT	88	0.362	3.234	0.003	0.027
HUMORAL IMMUNE RESPONSE	106	0.220	3.230	0.006	0.027
TP53 REGULATES METABOLIC GENES	70	0.265	3.227	0.002	0.027
PHOSPHOLIPASES	35	0.435	3.214	0.002	0.027
NEGATIVE REGULATION OF NEURON APOPTOTIC PROCESS	47	0.331	3.199	0.003	0.028
NEGATIVE REGULATION OF CYTOKINE PRODUCTION	126	0.236	3.197	0.006	0.028
HOMOPHILIC CELL ADHESION VIA PLASMA MEMBRANE ADHESION MOLECULES	37	0.555	3.197	0.001	0.028
B CELL DIFFERENTIATION	51	0.339	3.196	0.002	0.028
ANTIGEN RECEPTOR-MEDIATED SIGNALING PATHWAY	134	0.335	3.191	0.001	0.028
DEACTIVATION OF THE BETA-CATENIN TRANSACTIVATING COMPLEX	34	0.387	3.171	0.002	0.029
MUSCLE CONTRACTION	151	0.314	3.170	0.001	0.029
POSITIVE REGULATION OF NEURON DEATH	24	0.486	3.167	0.000	0.029
REGULATION OF CALCIUM ION IMPORT	62	0.373	3.167	0.000	0.029
RIBONUCLEOSIDE METABOLIC PROCESS	197	0.201	3.175	0.002	0.030
REGULATION OF ESTABLISHMENT OF PROTEIN LOCALIZATION TO MITOCHONDRION	113	0.257	3.172	0.003	0.030
POSITIVE REGULATION OF BIOMINERAL TISSUE DEVELOPMENT	29	0.516	3.158	0.001	0.030
REGULATION OF SYNAPTIC PLASTICITY	47	0.427	3.152	0.001	0.030
SYNTHESIS OF SUBSTRATES IN N-GLYCAN BIOSYTHESIS	53	0.366	3.150	0.000	0.030
LEUKOCYTE DIFFERENTIATION	145	0.260	3.140	0.002	0.031
PI3K AKT SIGNALING IN CANCER	64	0.381	3.134	0.002	0.031

CHROMATIN SILENCING AT RDNA	28	0.198	3.118	0.008	0.032
GLYCOSPHINGOLIPID METABOLISM	30	0.428	3.107	0.001	0.033
POSITIVE REGULATION OF PROTEIN SECRETION	111	0.278	3.104	0.001	0.033
VIRAL LIFE CYCLE	189	0.210	3.097	0.003	0.033
NEGATIVE REGULATION OF CELL MOTILITY	142	0.284	3.093	0.000	0.033
NEGATIVE REGULATION OF CELL MIGRATION	135	0.295	3.083	0.000	0.034
CELLULAR METABOLIC COMPOUND SALVAGE	28	0.443	3.079	0.000	0.034
PURINE-CONTAINING COMPOUND BIOSYNTHETIC PROCESS	82	0.312	3.077	0.003	0.034
NUCLEOTIDE CATABOLIC PROCESS	42	0.397	3.065	0.003	0.035
COMPLEMENT CASCADE	29	0.417	3.063	0.002	0.035
PROTEIN AUTOUBIQUITINATION	40	0.385	3.062	0.006	0.035
C-TYPE LECTIN RECEPTORS (CLRS)	105	0.282	3.058	0.004	0.035
REGULATION OF BONE MINERALIZATION	45	0.405	3.058	0.001	0.035
NEGATIVE REGULATION OF WNT SIGNALING PATHWAY	139	0.278	3.056	0.001	0.035
NUCLEOSIDE MONOPHOSPHATE METABOLIC PROCESS	156	0.208	3.055	0.002	0.035
B-WICH COMPLEX POSITIVELY REGULATES RRNA EXPRESSION	70	0.228	3.053	0.000	0.035
POSITIVE REGULATION OF PROTEIN MODIFICATION BY SMALL PROTEIN CONJUGATION OR REMOVAL	142	0.220	3.048	0.001	0.035
ANATOMICAL STRUCTURE HOMEOSTASIS	159	0.255	3.046	0.004	0.035
REGULATION OF CYTOSOLIC CALCIUM ION CONCENTRATION	112	0.356	3.044	0.000	0.035
TOLL-LIKE RECEPTOR SIGNALING PATHWAY	68	0.308	3.043	0.001	0.035
TRK RECEPTOR SIGNALING MEDIATED BY PI3K AND PLC-GAMMA	34	0.431	3.040	0.000	0.035
MONOSACCHARIDE METABOLIC PROCESS	112	0.270	3.038	0.003	0.035
HIV INFECTION	183	0.254	3.026	0.002	0.036
MULTI-ORGANISM BEHAVIOR	24	0.529	3.020	0.003	0.037
GLYCOSAMINOGLYCAN METABOLIC PROCESS	108	0.312	3.008	0.004	0.037
ORGANIC ACID TRANSPORT	190	0.304	3.004	0.002	0.038
GTP HYDROLYSIS AND JOINING OF THE 60S RIBOSOMAL SUBUNIT	78	0.147	3.002	0.001	0.038

ETHANOLAMINE-CONTAINING COMPOUND METABOLIC PROCESS	67	0.316	2.999	0.001	0.038
REGULATION OF CYTOKINE SECRETION	82	0.300	2.996	0.003	0.038
SULFUR COMPOUND BIOSYNTHETIC PROCESS	146	0.281	2.994	0.004	0.038
SELENOCYSTEINE SYNTHESIS	63	0.094	2.984	0.032	0.038
REGULATION OF TRANSPORTER ACTIVITY	143	0.382	2.980	0.001	0.039
MODULATION OF SYNAPTIC TRANSMISSION	121	0.389	2.968	0.002	0.040
REGULATION OF MITOCHONDRION ORGANIZATION	183	0.218	2.963	0.004	0.040
NEGATIVE REGULATION OF CANONICAL WNT SIGNALING PATHWAY	119	0.283	2.960	0.004	0.040
PI3 KINASE PATHWAY	34	0.447	2.954	0.001	0.040
CARBOHYDRATE BIOSYNTHETIC PROCESS	75	0.288	2.941	0.001	0.042
CELLULAR RESPONSE TO OXYGEN LEVELS	83	0.286	2.936	0.003	0.042
MULTIVESICULAR BODY ASSEMBLY	24	0.476	2.927	0.000	0.042
SIG PIP3 SIGNALING IN CARDIAC MYOCTES	54	0.394	2.926	0.003	0.042
APOPTOTIC SIGNALING PATHWAY	193	0.222	2.925	0.003	0.042
POSITIVE REGULATION OF MITOCHONDRION ORGANIZATION	144	0.227	2.921	0.005	0.042
ENDOSOMAL TRANSPORT	185	0.290	2.928	0.002	0.043
REGULATION OF CYCLIC NUCLEOTIDE BIOSYNTHETIC PROCESS	79	0.367	2.918	0.003	0.043
CONNECTIVE TISSUE DEVELOPMENT	65	0.322	2.915	0.003	0.043
INSULIN-MEDIATED GLUCOSE TRANSPORT	27	0.427	2.912	0.000	0.043
SIGNAL TRANSDUCTION BY P53 CLASS MEDIATOR	97	0.305	2.910	0.001	0.043
SKELETAL SYSTEM MORPHOGENESIS	55	0.308	2.908	0.007	0.043
REGULATION OF TRANSLATION	199	0.248	2.900	0.004	0.043
EPIGENETIC REGULATION OF GENE EXPRESSION	122	0.223	2.901	0.000	0.044
POSITIVE REGULATION OF ESTABLISHMENT OF PROTEIN LOCALIZATION TO MITOCHONDRION	104	0.249	2.896	0.006	0.044
CAP-DEPENDENT TRANSLATION INITIATION	84	0.142	2.885	0.001	0.044
EUKARYOTIC TRANSLATION INITIATION	84	0.142	2.885	0.001	0.044
REGULATION OF GLYCOPROTEIN BIOSYNTHETIC PROCESS	26	0.426	2.886	0.001	0.045

REGULATION OF CYTOKINE-MEDIATED SIGNALING PATHWAY	94	0.257	2.880	0.005	0.045
PURINE NUCLEOSIDE MONOPHOSPHATE METABOLIC PROCESS	148	0.202	2.875	0.008	0.045
PURINE RIBONUCLEOSIDE MONOPHOSPHATE METABOLIC PROCESS	148	0.202	2.875	0.008	0.045
REGULATION OF GLYCOPROTEIN METABOLIC PROCESS	30	0.394	2.874	0.002	0.045
RIBONUCLEOSIDE MONOPHOSPHATE METABOLIC PROCESS	151	0.209	2.856	0.006	0.047
FORMATION OF A POOL OF FREE 40S SUBUNITS	70	0.136	2.852	0.007	0.047
REGULATION OF CYCLIC NUCLEOTIDE METABOLIC PROCESS	90	0.336	2.852	0.004	0.047
MACROMOLECULE METHYLATION	141	0.280	2.850	0.005	0.047
MONOCARBOXYLIC ACID TRANSPORT	80	0.354	2.845	0.004	0.047
MORPHOGENESIS OF AN EPITHELIUM	144	0.255	2.842	0.004	0.047
CARBOXYLIC ACID TRANSPORT	187	0.300	2.839	0.005	0.047
SINGLE-ORGANISM BEHAVIOR	93	0.406	2.837	0.003	0.048
CASPASE CASCADE IN APOPTOSIS	48	0.325	2.833	0.008	0.048
CARBOXYLIC ACID BIOSYNTHETIC PROCESS	158	0.264	2.829	0.008	0.048
ORGANIC ACID BIOSYNTHETIC PROCESS	158	0.264	2.829	0.008	0.048
AMMONIUM ION METABOLIC PROCESS	114	0.293	2.826	0.003	0.048
GLYCOSYL COMPOUND CATABOLIC PROCESS	31	0.383	2.822	0.002	0.048
TUBE DEVELOPMENT	174	0.250	2.822	0.009	0.048
SENESCENCE-ASSOCIATED SECRETORY PHENOTYPE (SASP)	84	0.184	2.821	0.000	0.048
HDACS DEACETYLATE HISTONES	75	0.133	2.820	0.004	0.048
PROTEIN ACTIVATION CASCADE	42	0.339	2.808	0.005	0.049
PEPTIDYL-LYSINE METHYLATION	47	0.327	2.802	0.010	0.049
REGULATION OF CAMP BIOSYNTHETIC PROCESS	70	0.383	2.802	0.006	0.050
GENE SILENCING	146	0.221	2.799	0.003	0.050
NUCLEOSIDE PHOSPHATE CATABOLIC PROCESS	47	0.356	2.794	0.006	0.050

Note—A total of 237 pathways are displayed at FDR ≤ 0.05. **Abbreviations**: ES, enrichment score; NES, normalized enrichment score; NominalP, nominal *p* value; FDR, false discovery rate.

Supplementary Table 3. Leading-edge subset per confidently enriched and nonenriched pathway.

ID #	Enriched				Nonenriched			
	HM3		PNC		HM3		PNC	
	Size	LE	Size	LE	Size	LE	Size	LE
1	105	51	98	54	39	23	37	14
2	148	76	135	61	92	43	86	34
3	75	52	62	10	64	10	63	15
4	91	47	84	44	28	35	28	27
5	183	87	173	68	24	14	22	6
6	99	51	93	35	33	16	32	22
7	165	100	146	78	25	20	23	15
8	168	102	144	70	94	58	86	39
9	169	91	156	61	48	24	41	22
10	155	79	140	65	47	36	36	28
11	157	96	138	76	48	23	40	22
12	46	26	45	16	29	7	25	18
13	76	44	70	31	20	13	20	10
14	102	59	90	38	42	9	37	26
15	183	88	168	87	45	34	42	31
16	101	49	93	36	24	16	24	12
17	74	43	68	31	31	14	25	19
18	73	30	70	22	28	22	26	15
19	137	68	119	67	-	-	-	-

Abbreviations: HM3, HapMap Phase 3; PNC, Philadelphia Neurodevelopmental Cohort; LE, leading-edge.

Supplementary Table 4. *Regulation of hemopoiesis g*enes with evidence for recent positive selection via dbPSHP.

Chrom	Gene	Population	Description	Function
chr17	*ACE*	52 worldwide populations	-	climate adaption; metabolism
chr2	*ACVR2A*	CEU-YRI	23.6/CLR	-
chr10	*ADAM8*	CEU+YRI	-	-
chr10	*ADAM8*	CEU-ASN	-	-
chr6	*AGER*	Mandenka	-	-
chr6	*AGER*	CHB	-2.189/Tajima's D; -3.508/Fay and Wu's H	-
chr6	*AGER*	ASN	-	-
chr1	*ARNT*	MKK	2.6089/absolute value of iHS	-
chr1	*ARNT*	MKK	2.47585/absolute value of iHS	-
chr1	*ARNT*	MKK	2.4139/absolute value of iHS	-
chr1	*ARNT*	MKK	2.49442/absolute value of iHS	-
chr1	*ARNT*	MKK	2.62799/absolute value of iHS	-
chr1	*ARNT*	MKK	2.47308/absolute value of iHS	-
chr1	*ARNT*	MKK	2.7361/average iHS	-
chr1	*ARNT*	MKK	2.22633/absolute value of iHS	-
chr1	*ARNT*	MKK	2.65039/absolute value of iHS	-
chr1	*ARNT*	MKK	2.72292/absolute value of iHS	-

chr1	*ARNT*	MKK	2.42302/absolute value of iHS	-
chr5	*CAMK4*	CEU	-	-
chr5	*CARTPT*	Yoruba-Hazara-Yakut-Mongola-Xibo- Oroqen-Hezhen-Maya	-	-
chr3	*CCR1*	CEU+YRI	-	-
chr3	*CCR1*	YRI-CHB-CEU	0.5505/TD CEU; -1.2962/TD CHB; -1.7328/TD YRI	-
chr3	*CD80*	CEU-ASN	-	-
chr16	*CD86*	CEU	0.45/Fst; -2.73/iHS	multiple sclerosis
chr1	*CDC73*	Bangladeshi	5.49/Max CMS score	cholera susceptibility; chloride secretion; innate immune system
chr7	*CDK6*	GIH	3.56E-02/adjusted haploPS score	-
chr9	*CDKN2A*	ASN-CEU	-	type 2 diabetes
chr11	*CTR9*	CHB+JPT	-	-
chr17	*FLCN*	JPT	2.36E-02/adjusted haploPS score	-
chr5	*FNIP1*	Mbutipygmy	-	-
chr5	*FNIP1*	TSI	4.42E-02/adjusted haploPS score	-
chr5	*FNIP1*	CEU	2.60E-02/adjusted haploPS score	-
chr5	*FNIP1*	African Pygmy	-	human height; bone homeostasis
chr6	*FOXO3*	CHM,CEU	3.3625/Maximum iHS absolute value	-
chr3	*FOXP1*	Kalash	-	-
chr3	*FOXP1*	CHB+JPT	-	-
chr3	*FOXP1*	Oceania	-	-

chr3	*FOXP1*	YRI	-	-
chr3	*FOXP1*	CHM-CEU	3.5358/Maximum XP-EHH score	-
chr3	*FOXP1*	CHM-CEU	3.2946/Maximum XP-EHH score	-
chr3	*FOXP1*	ASN-CEU-YRI	9/SNP with FST>0.65	-
chr2	*GLI2*	MKK	2.30136/absolute value of iHS	-
chr2	*GLI2*	ASN-CEU-YRI	19/SNP with FST>0.65	-
chr2	*GLI2*	ASN-YRI	341.1/max XP-CLR; 0.0089/XP-EHH; 0.2221/iHS(ASN); 0.0079/iHS(YRI)	-
chr2	*GLI2*	CEU	4.94E-02/adjusted haploPS score	-
chr2	*GLI2*	MAS	2.51E-02/adjusted haploPS score	-
chr2	*GLI2*	MKK	2.96849/absolute value of iHS	-
chr2	*GLI2*	MKK	3.35722/absolute value of iHS	-
chr2	*GLI2*	MKK	2.20054/absolute value of iHS	-
chr2	*GLI2*	CHB	1.42E-02/adjusted haploPS score	-
chr2	*GLI2*	TSI	3.20E-02/adjusted haploPS score	-
chr2	*GLI2*	MKK	4.04988/absolute value of iHS	-
chr2	*GLI2*	JPT	2.32E-02/adjusted haploPS score	-
chr2	*GLI2*	MKK	3.49522/absolute value of iHS	-

chr2	*GLI2*	MKK	2.96849/absolute value of iHS	-
chr2	*GLI2*	MEX	1.21E-03/adjusted haploPS score	-
chr2	*GLI2*	MKK	2.64/absolute value of iHS	-
chr2	*GLI2*	MKK	2.71923/absolute value of iHS	-
chr2	*GLI2*	MKK	3.35722/absolute value of iHS	-
chr2	*GLI2*	CHD	4.87E-02/adjusted haploPS score	-
chr2	*GLI2*	GIH	3.41E-02/adjusted haploPS score	-
chr7	*GLI3*	MKK	0.2521/Fst	-
chr7	*GLI3*	MKK	0.2604/Fst	-
chr7	*GLI3*	MKK	0.1711/Fst	-
chr7	*GLI3*	MKK	0.1688/Fst	-
chr7	*GLI3*	MKK	0.1752/Fst	-
chr7	*GLI3*	MKK	0.1948/Fst	-
chr7	*GLI3*	YRI	-	-
chr7	*GLI3*	MKK	0.2738/Fst	-
chr7	*GLI3*	MKK	0.1726/Fst	-
chr7	*GLI3*	MKK	0.1736/Fst	-
chr7	*GLI3*	MKK	0.2509/Fst	-
chr7	*GLI3*	MKK	0.1809/Fst	-
chr7	*GLI3*	MKK	0.2266/Fst	-
chr7	*GLI3*	MKK	0.2421/Fst	-
chr7	*GLI3*	MKK	0.1948/Fst	-
chr7	*GLI3*	MKK	0.1747/Fst	-
chr7	*GLI3*	LWK	2.33E-02/adjusted haploPS score	-

chr7	*GLI3*	MKK	0.2105/Fst	-
chr7	*GLI3*	ASN-CEU-YRI	1/SNP with FST>0.65	-
chr3	*HCLS1*	ASN-YRI-CEU	-	Positive regulation of cell proliferation; Immune-related
chr3	*HCLS1*	CEU-ASN	-	-
chr3	*HCLS1*	MEX	1.60E-02/adjusted haploPS score	-
chr14	*HIF1A*	Daghestani	-	high-altitude adaptation
chr14	*HIF1A*	CEU	-	-
chr14	*HIF1A*	YRI	-	-
chr6	*HIST1H4A*	GIH	1.33E-02/adjusted haploPS score	-
chr6	*HIST1H4B*	GIH	1.33E-02/adjusted haploPS score	-
chr6	*HIST1H4C*	GIH	1.33E-02/adjusted haploPS score	-
chr6	*HIST1H4D*	JPT	3.94E-02/adjusted haploPS score	-
chr6	*HIST1H4E*	JPT	3.94E-02/adjusted haploPS score	-
chr6	*HIST1H4F*	JPT	3.94E-02/adjusted haploPS score	-
chr6	*HIST1H4F*	CEU-CHB-YRI	44.027/CLR	-
chr6	*HIST1H4J*	TSI	3.54E-02/adjusted haploPS score	-
chr6	*HIST1H4J*	MEX	4.58E-02/adjusted haploPS score	-
chr6	*HIST1H4J*	CEU	2.91E-02/adjusted haploPS score	-
chr6	*HIST1H4K*	CEU	2.91E-02/adjusted haploPS score	-

chr6	*HIST1H4K*	TSI	3.54E-02/adjusted haploPS score	-
chr6	*HIST1H4K*	MEX	4.58E-02/adjusted haploPS score	-
chr6	*HIST1H4L*	MEX	4.58E-02/adjusted haploPS score	-
chr6	*HIST1H4L*	TSI	3.54E-02/adjusted haploPS score	-
chr6	*HIST1H4L*	CEU	2.91E-02/adjusted haploPS score	-
chr6	*HLA-G*	Oceania	-	-
chr6	*HLA-G*	CHB	-	-
chr6	*HLA-G*	Makrani-Yizu-Miaozu	-	-
chr12	*IFNG*	CEU	-	-
chr6	*IL17A*	America	-	-
chr5	*IL3*	Mbutipygmy	-	-
chr5	*IL3*	Csasia-Mideast-Europe-America	-	-
chr5	*IL4*	MKK	0.2460/Fst	-
chr5	*IL4*	CEU	-	-
chr5	*IL4*	MKK	0.2483/Fst	-
chr7	*INHBA*	LWK	2.33E-02/adjusted haploPS score	-
chr7	*INHBA*	present-day human-Neandertal	-	-
chr5	*IRF1*	CEU	-	Crohn's disease
chr5	*IRF1*	CEU	0.43/Fst; -2.6/iHS	Crohn disease
chr6	*IRF4*	YRI-CEU-ASN	-	Hair colour
chr6	*IRF4*	YRI-CEU-ASN	-	Hair colour
chr15	*LEO1*	CHM-CEU	2.913/Maximum XP-EHH score	-

chr22	*LIF*	CHM-YRI	3.2607/Maximum XP-EHH score	-
chr19	*LILRB2*	CHB+JPT	-	HLA class I-recognizing receptors; innate and adaptive immunity
chr3	*LTF*	CEU+YRI	-	-
chr8	*LYN*	CHM-CEU	7.2437/Maximum XP-EHH score	-
chr8	*LYN*	CEU-CHB-YRI	25.735/CLR	-
chr8	*LYN*	CHM-YRI	6.008/Maximum XP-EHH score	-
chr6	*MAPK14*	52 worldwide populations	-	climate adaption; metabolism
chr19	*PAF1*	Bangladeshi	3.52/Max CMS score	cholera susceptibility; chloride secretion; innate immune system
chr4	*PF4*	Mideast-Easia	-	-
chr4	*PF4*	Druze-Uygur-Mongola-Naxi	-	-
chr1	*PRDM16*	CEU	-	-
chr1	*PRDM16*	ASN-CEU-YRI	2/SNP with FST>0.65	-
chr17	*PRKCA*	CEU	-	-
chr17	*PRKCA*	CEU	2.87E-04/CMS test P-value	-
chr17	*PRKCA*	CEU	-	-
chr8	*PTK2B*	52 worldwide populations	-	climate adaption; metabolism
chr8	*PTK2B*	52 worldwide populations	-	climate adaption; metabolism
chr8	*PTK2B*	52 worldwide populations	-	climate adaption; metabolism
chr8	*PTK2B*	52 worldwide populations	-	climate adaption; metabolism

chr8	*PTK2B*	52 worldwide populations	-	climate adaption; metabolism
chr8	*PTK2B*	ASN-YRI-CEU	-	Positive regulation of cell proliferation; Immune-related
chr6	*RIPK1*	CEU	-	-
chr21	*SOD1*	52 worldwide populations	-	climate adaption; metabolism
chr21	*SOD1*	52 worldwide populations	-	climate adaption; metabolism
chr11	*SPI1*	CEU-YRI-CHB	-	-
chr6	*TNF*	CEU	1.16E-02/adjusted haploPS score	-
chr6	*TNF*	JPT	2.45E-02/adjusted haploPS score	-
chr13	*TNFSF11*	Bedouin	-	-
chr6	*VNN1*	YRI	-	-
chr6	*VNN1*	YRI	3.18E-04/CMS test P-value	-
chr6	*VNN1*	LWK	2.27E-03/adjusted haploPS score	-
chr2	*ZAP70*	GIH	3.57E-02/adjusted haploPS score	-
chr2	*ZAP70*	India	0.00017/Empirical p-value, India	-
chr2	*ZAP70*	Bangladeshi	7.2/Max CMS score	cholera susceptibility; chloride secretion; innate immune system

Note—57 unique genes (from 183 total) in the pathway demonstrate evidence for recent positive selection. Genes with multiple entires represent different positively selected loci identified within that gene. The ‘Description’ column indicates the particular statistical term for which the selective signal was determined. **Abbreviations:** YRI, Nigerian; CEU, European-American; GIH, Gujarati-Indian-American; MKK, Maasai Kenya; TSI, Italian; CHB, Han Chinese; JPT, Japanese; LWK, Luhya Kenya; CHD, Chinese-American; MEX, Mexican-American; FST, fixation index; iHS, integrated haplotype score; XP-EHH, cross-population extended haplotype homozygosity; XP-CLR, cross-population composite likelihood ratio; CMS, composite of multiple signals (unknown/undefined acronyms: CHM, MAS).

Supplementary Table 5. *Toll-like receptor pathway g*enes with evidence for recent positive selection via dbPSHP.

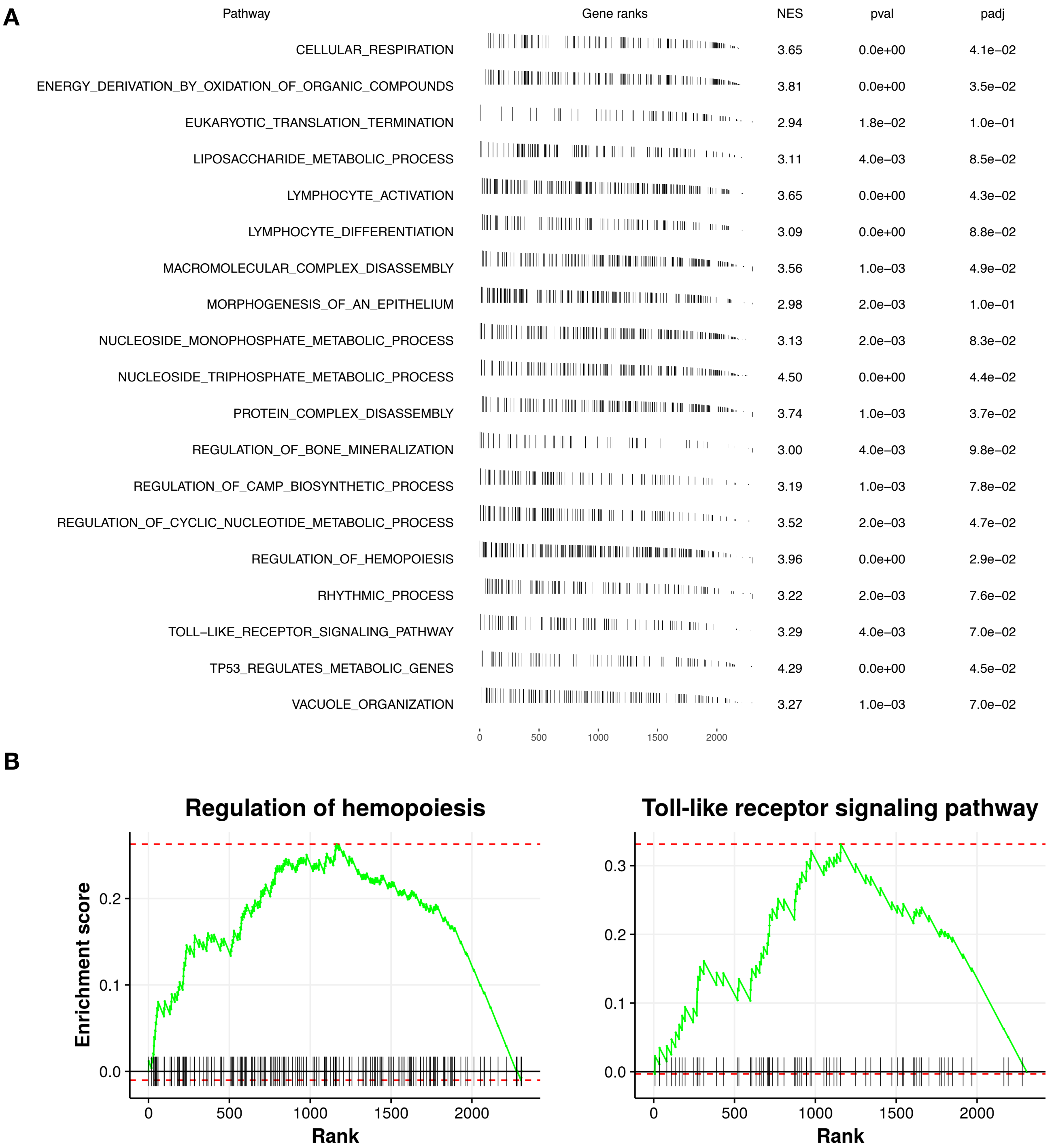
Chrom	Gene	Population	Description	Function
chr5	*CD180*	CEU	-	CD180 molecule
chr7	*CD36*	YRI	-	-
chr7	*CD36*	YRI	7.05E-09/p value	-
chr7	*CD36*	YRI	3.78E-08/p value	-
chr7	*CD36*	52 worldwide populations	-	climate adaption; metabolism
chr7	*CD36*	YRI	2.72E-06/Empirical P-value	-
chr7	*CD36*	YRI	1.30E-02/adjusted haploPS score	-
chr10	*CHUK*	YRI	3.16E-02/adjusted haploPS score	-
chr10	*CHUK*	YRI	1.47E-03/CMS test P-value	-
chr8	*CTSB*	Adygei-Mongola-Tu-Brahui-Pathan	-	-
chr8	*CTSB*	Bergamo-Brahui-Uygur-Xibo-Oroqen-Daur-Tujia-Miaozu-Dai-Lahu- Nasioi-Papuan-Maya-Colombian	-	-
chr8	*CTSB*	Csasia-America	-	-
chr8	*CTSB*	CHM-CEU	3.5081/Maximum XP-EHH score	-
chr8	*CTSB*	CHM-YRI	4.2413/Maximum XP-EHH score	-
chr1	*CTSK*	MKK	2.7361/average iHS	-

chr1	*CTSK*	MKK	2.68535/absolute value of iHS	-
chr1	*CTSS*	MKK	2.70558/absolute value of iHS	-
chr1	*CTSS*	MKK	3.58697/absolute value of iHS	-
chr1	*CTSS*	MKK	3.40564/absolute value of iHS	-
chr1	*CTSS*	MKK	3.3589/absolute value of iHS	-
chr1	*CTSS*	MKK	2.49308/absolute value of iHS	-
chr1	*CTSS*	MKK	2.71415/absolute value of iHS	-
chr1	*CTSS*	MKK	3.47909/absolute value of iHS	-
chr1	*CTSS*	MKK	3.45836/absolute value of iHS	-
chr1	*CTSS*	MKK	2.45722/absolute value of iHS	-
chr1	*CTSS*	MKK	3.31742/absolute value of iHS	-
chr1	*CTSS*	MKK	2.7361/average iHS	-
chr1	*CTSS*	MKK	2.47647/absolute value of iHS	-
chr12	*HSP90B1*	MKK	1.47E-03/adjusted haploPS score	-
chr2	*HSPD1*	CEU-CHB-YRI	0.972/CLR P-value	heat shock genes
chr8	*IKBKB*	CEU	-	-
chr8	*IKBKB*	MEX	1.62E-02/adjusted haploPS score	-
chr3	*IRAK2*	CEU	-	-
chr12	*IRAK4*	GIH	4.42E-03/adjusted haploPS score	-

chr12	*IRAK4*	Cambodian	-	-
chr16	*ITGAM*	MKK	0.1714/Fst	-
chr16	*ITGAM*	CEU-ASN	-	-
chr16	*ITGAM*	ASN-CEU-YRI	2/SNP with FST>0.65	-
chr16	*ITGAM*	Yakut-Colombian	-	-
chr3	*MAPKAPK3*	CEU+YRI	-	-
chr3	*MAPKAPK3*	Bangladeshi	4/Max CMS score	cholera susceptibility; chloride secretion; innate immune system
chr12	*NR1H4*	CEU-CHB-YRI	37.529/CLR	-
chr12	*NR1H4*	YRI-CHB-CEU	-1.8667/TD CEU; 1.9972/TD CHB; 1.317/TD YRI	-
chr3	*RFTN1*	YRI	-	-
chr6	*RIPK1*	CEU	-	-
chr2	*TANK*	Karitiana	-	-
chr19	*TICAM1*	CEU-ASN-YRI	-	TIR-containing adaptors of innate immune system
chr4	*TLR1*	YRI-CHB-CEU	-0.4247/TD CEU; 2.5188/TD CHB; -1.1971/TD YRI	-
chr4	*TLR1*	Pacific Islander	5.999/REHH	activation of innate immune system; sepsis, leprosy and tuberculosis
chr4	*TLR1*	CEU	0.53/SIFT score	activation of innate immune system; sepsis, leprosy and tuberculosis
chr4	*TLR1*	CEU	0.473/Fst	activation of innate immune system; sepsis, leprosy and tuberculosis
chr1	*TLR5*	CHM-YRI	4.3193/Maximum XP-EHH score	-

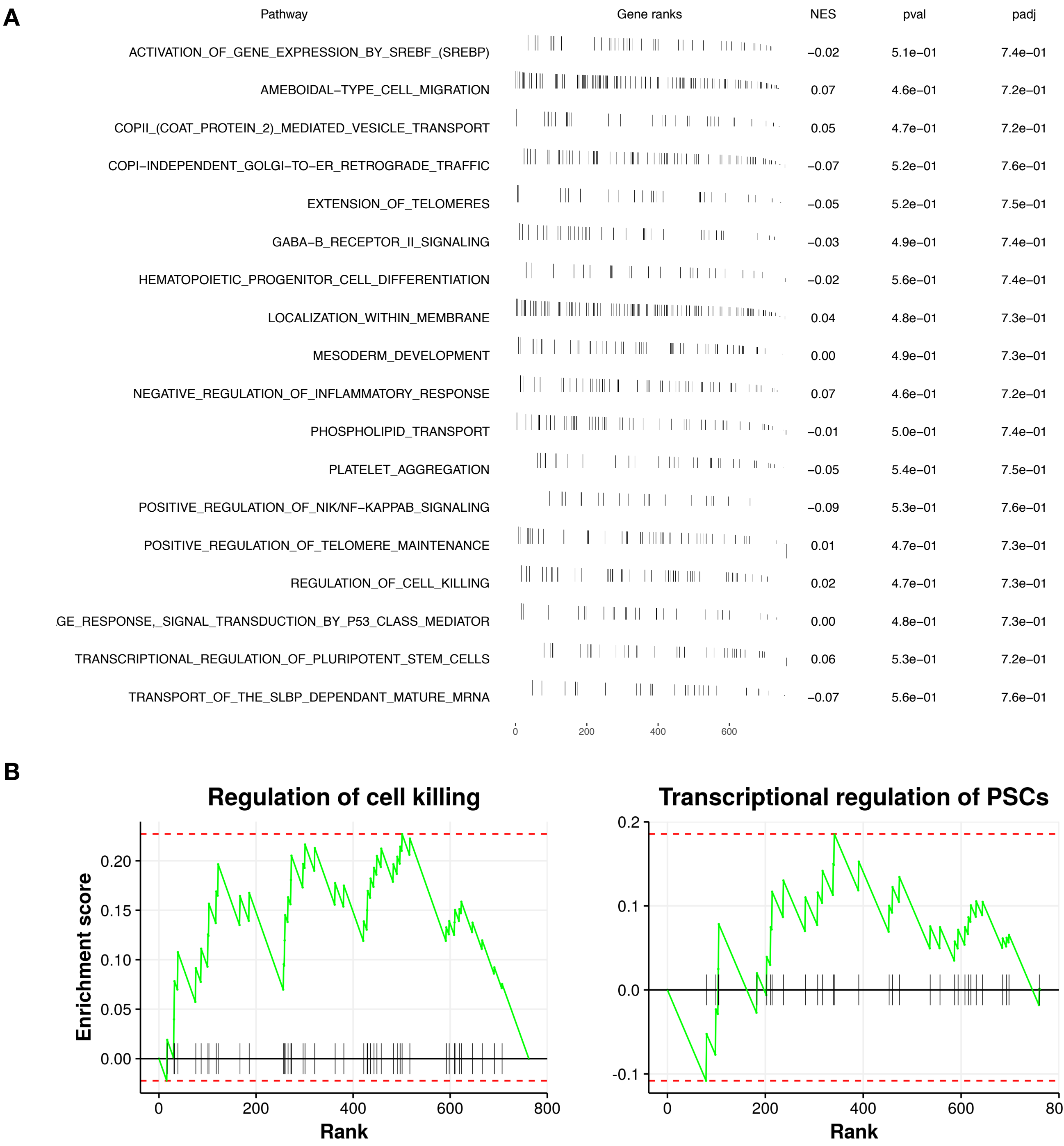
chr1	*TLR5*	MKK	2.86157/absolute value of iHS	-
chr1	*TLR5*	MKK	2.6131/absolute value of iHS	-
chr1	*TLR5*	YRI	-	-
chr1	*TLR5*	MKK	2.57621/absolute value of iHS	-
chr1	*TLR5*	CHM-YRI	3.5345/Maximum XP-EHH score	-
chr1	*TLR5*	MKK	2.92336/absolute value of iHS	-
chr1	*TLR5*	MKK	2.02128/absolute value of iHS	-
chr1	*TLR5*	MKK	2.50958/absolute value of iHS	-
chr1	*TLR5*	CHM-CEU	2.9636/Maximum XP-EHH score	-
chr1	*TLR5*	CHM-CEU	4.0257/Maximum XP-EHH score	-
chr5	*TNIP1*	Karitiana	-	-
chr5	*TNIP1*	America	-	-

Note—21 unique genes (from 74 total) in the pathway demonstrate evidence for recent positive selection. Genes with multiple entires represent different positively selected loci identified within that gene. The ‘Description’ column indicates the particular statistical term for which the selective signal was determined. **Abbreviations:** YRI, Nigerian; CEU, European-American; GIH, Gujarati-Indian-American; MKK, Maasai Kenya; TSI, Italian; CHB, Han Chinese; JPT, Japanese; LWK, Luhya Kenya; CHD, Chinese-American; MEX, Mexican-American; REHH, extended homozygosity haplotype (EHH) based tests; FST, fixation index; iHS, integrated haplotype score; XP-EHH, cross-population extended haplotype homozygosity; XP-CLR, cross-population composite likelihood ratio; CMS, composite of multiple signals (unknown/undefined acronyms: CHM, MAS).



Pathway	NES	pval	padj
CELLULAR_RESPIRATION	3.65	0.0e+00	4.1e−02
ENERGY_DERIVATION_BY_OXIDATION_OF_ORGANIC_COMPOUNDS	3.81	0.0e+00	3.5e−02
EUKARYOTIC_TRANSLATION_TERMINATION	2.94	1.8e−02	1.0e−01
LIPOSACCHARIDE_METABOLIC_PROCESS	3.11	4.0e−03	8.5e−02
LYMPHOCYTE_ACTIVATION	3.65	0.0e+00	4.3e−02
LYMPHOCYTE_DIFFERENTIATION	3.09	0.0e+00	8.8e−02
MACROMOLECULAR_COMPLEX_DISASSEMBLY	3.56	1.0e−03	4.9e−02
MORPHOGENESIS_OF_AN_EPITHELIUM	2.98	2.0e−03	1.0e−01
NUCLEOSIDE_MONOPHOSPHATE_METABOLIC_PROCESS	3.13	2.0e−03	8.3e−02
NUCLEOSIDE_TRIPHOSPHATE_METABOLIC_PROCESS	4.50	0.0e+00	4.4e−02
PROTEIN_COMPLEX_DISASSEMBLY	3.74	1.0e−03	3.7e−02
REGULATION_OF_BONE_MINERALIZATION	3.00	4.0e−03	9.8e−02
REGULATION_OF_CAMP_BIOSYNTHETIC_PROCESS	3.19	1.0e−03	7.8e−02
REGULATION_OF_CYCLIC_NUCLEOTIDE_METABOLIC_PROCESS	3.52	2.0e−03	4.7e−02
REGULATION_OF_HEMOPOIESIS	3.96	0.0e+00	2.9e−02
RHYTHMIC_PROCESS	3.22	2.0e−03	7.6e−02
TOLL−LIKE_RECEPTOR_SIGNALING_PATHWAY	3.29	4.0e−03	7.0e−02
TP53_REGULATES_METABOLIC_GENES	4.29	0.0e+00	4.5e−02
VACUOLE_ORGANIZATION	3.27	1.0e−03	7.0e−02

Supplementary Figure 1. GSEA enrichment plot and associated statistics per ancestry-enriched pathway. Tabular (**A**) and graphical (**B**) representations of a pathway enrichment score. The ES reflects the degree to which the genes in a pathway are overrepresented at the top or bottom of the entire ranked list of genes (see Materials and Methods for additional details). As demonstrated by the 'Gene ranks' column of **A**, a nonenriched pathway will have its genes spread more or less uniformly through the ranked list. A positively enriched pathway, on the other hand, will have a larger portion of its genes at the left end of the ranked list (and vice versa for a negatively enriched pathway). Graphically, the ES is represented by the maximum deviation from 0 as GSEA walks down the ranked list of genes, and is depicted by the top red dashed line in **B**. The green line represents the running sum enrichment score as GSEA walks through the ranked list of genes in a pathway. Both figures are generated via the fgsea R package (v1.4.1) [169]. **Abbreviations**: NES, normalized enrichment score; pval, Nominal *p* value; padj, adjusted *p* value (i.e., FDR).



Pathway	Gene ranks	NES	pval	padj
ACTIVATION_OF_GENE_EXPRESSION_BY_SREBF_(SREBP)		−0.02	5.1e−01	7.4e−01
AMEBOIDAL−TYPE_CELL_MIGRATION		0.07	4.6e−01	7.2e−01
COPII_(COAT_PROTEIN_2)_MEDIATED_VESICLE_TRANSPORT		0.05	4.7e−01	7.2e−01
COPI−INDEPENDENT_GOLGI−TO−ER_RETROGRADE_TRAFFIC		−0.07	5.2e−01	7.6e−01
EXTENSION_OF_TELOMERES		−0.05	5.2e−01	7.5e−01
GABA−B_RECEPTOR_II_SIGNALING		−0.03	4.9e−01	7.4e−01
HEMATOPOIETIC_PROGENITOR_CELL_DIFFERENTIATION		−0.02	5.6e−01	7.4e−01
LOCALIZATION_WITHIN_MEMBRANE		0.04	4.8e−01	7.3e−01
MESODERM_DEVELOPMENT		0.00	4.9e−01	7.3e−01
NEGATIVE_REGULATION_OF_INFLAMMATORY_RESPONSE		0.07	4.6e−01	7.2e−01
PHOSPHOLIPID_TRANSPORT		−0.01	5.0e−01	7.4e−01
PLATELET_AGGREGATION		−0.05	5.4e−01	7.5e−01
POSITIVE_REGULATION_OF_NIK/NF−KAPPAB_SIGNALING		−0.09	5.3e−01	7.6e−01
POSITIVE_REGULATION_OF_TELOMERE_MAINTENANCE		0.01	4.7e−01	7.3e−01
REGULATION_OF_CELL_KILLING		0.02	4.7e−01	7.3e−01
GE_RESPONSE,_SIGNAL_TRANSDUCTION_BY_P53_CLASS_MEDIATOR		0.00	4.8e−01	7.3e−01
TRANSCRIPTIONAL_REGULATION_OF_PLURIPOTENT_STEM_CELLS		0.06	5.3e−01	7.2e−01
TRANSPORT_OF_THE_SLBP_DEPENDANT_MATURE_MRNA		−0.07	5.6e−01	7.6e−01
	0 200 400 600			

Supplementary Figure 2. GSEA enrichment plot and associated statistics per nonenriched pathway. [See description for **Supplementary Figure 1**]. **Abbreviations**: PSCs, pluripotent stem cells.