Common Risk Alleles for Inflammatory Diseases Are Targets of Recent Positive Selection

Towfique Raj,[1,2,3,4] Manik Kuchroo,[1,4] Joseph M. Replogle,[2,4] Soumya Raychaudhuri,[2,3,4,5,6] Barbara E. Stranger,[2,3,4,7,8,9,*] and Philip L. De Jager[1,2,3,4,8,*]

Genome-wide association studies (GWASs) have identified hundreds of loci harboring genetic variation influencing inflammatory-disease susceptibility in humans. It has been hypothesized that present day inflammatory diseases may have arisen, in part, due to pleiotropic effects of host resistance to pathogens over the course of human history, with significant selective pressures acting to increase host resistance to pathogens. The extent to which genetic factors underlying inflammatory-disease susceptibility has been influenced by selective processes can now be quantified more comprehensively than previously possible. To understand the evolutionary forces that have shaped inflammatory-disease susceptibility and to elucidate functional pathways affected by selection, we performed a systems-based analysis to integrate (1) published GWASs for inflammatory diseases, (2) a genome-wide scan for signatures of positive selection in a population of European ancestry, (3) functional genomics data comprised of protein-protein interaction networks, and (4) a genome-wide expression quantitative trait locus (eQTL) mapping study in peripheral blood mononuclear cells (PBMCs). We demonstrate that loci for inflammatory-disease susceptibility are enriched for genomic signatures of recent positive natural selection, with selected loci forming a highly interconnected protein-protein interaction network. Further, we identify 21 loci for inflammatory-disease susceptibility that display signatures of recent positive selection, of which 13 also show evidence of *cis*-regulatory effects on genes within the associated locus. Thus, our integrated analyses highlight a set of susceptibility loci that might subserve a shared molecular function and has experienced selective pressure over the course of human history; today, these loci play a key role in influencing susceptibility to multiple different inflammatory diseases, in part through alterations of gene expression in immune cells.

Introduction

The availability of genome-wide genetic variation data from population-based cohorts, coupled with improved analytic methods, have led to a number of genome-wide scans for selection that have collectively identified hundreds of regions targeted by recent positive selection, many of which harbor genetic variants implicated in inflammatory diseases.[1–15] Interestingly, for many of these variants, selection has favored the haplotypes carrying disease risk alleles, suggesting a potential beneficial role for the risk allele in the course of human history, presumably for another function (e.g., resistance to pathogens).[1] For example, individuals homozygous for the Crohn disease (MIM 266600) risk allele at rs601338 in *FUT2* (MIM 182100) are protected from norovirus infection.[16,17] Similarly, it was shown that the positively selected celiac disease (MIM 212750) risk variant in *SH2B3* (MIM 605093) might have a protective role against bacterial infections.[18] Further, the major histocompatibility complex (MHC) contains several alleles that are strongly associated with susceptibility to different inflammatory diseases and have experienced strong selection that has been attributed to host-pathogen coevolution.[19–21] These observations are consistent with the "hygiene hypothesis," which posits that protective immune responses in a setting with high exposure to pathogens (which was the norm for most of human history) have become deleterious and contribute to inflammatory-disease susceptibility in the modern era of enhanced hygiene and extended lifetimes. With a genetic lens, the hypothesis states that risk alleles for inflammatory disease have been maintained in the human population as a result of adaptation to pathogen exposures that were endemic in the past but are no longer a significant burden in developed societies today.[22,23]

Although the studies listed above provide anecdotal evidence for a role for pathogen-driven selection at several inflammatory-disease loci, a systematic analysis exploring the role of natural selection in influencing susceptibility to inflammatory disease has not been performed. Until recently, the genetic architecture of inflammatory diseases was relatively uncharacterized, precluding a comprehensive analysis. Even with the recent catalogs of robust, replicated inflammatory-disease associations, several challenges to our understanding of disease-associated variation remain, including identifying the causal variant(s) within each locus and understanding how the dozens of known variants assemble into networks that ultimately lead to altered immune function, which is dysfunctional in our

[1]Program in Translational NeuroPsychiatric Genomics, Department of Neurology, Brigham & Women's Hospital, Boston, MA 02115, USA; [2]Division of Genetics, Department of Medicine, Brigham & Women's Hospital, Boston, MA 02115, USA; [3]Harvard Medical School, Boston, MA 02115, USA; [4]Program in Medical & Population Genetics, The Broad Institute, Cambridge, MA 02139, USA; [5]Division of Rheumatology, Immunology, and Allergy, Department of Medicine, Brigham and Women's, Hospital, Harvard Medical School, Boston, MA 02115, USA; [6]Partners Center for Personalized Genetic Medicine, Boston, MA 02115, USA; [7]Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, IL 60637, USA
[8]These authors contributed equally to this work
[9]Present address: Institute for Genomics and Systems Biology, University of Chicago, Chicago, IL 60637, USA
*Correspondence: bstranger@medicine.bsd.uchicago.edu (B.E.S.), pdejager@rics.bwh.harvard.edu (P.L.D.J.)
http://dx.doi.org/10.1016/j.ajhg.2013.03.001.

modern environment. Here, we perform a systematic analysis for evidence of natural selection to inform these two challenges, and we integrate our results with analyses of protein interaction networks and gene expression data (expression quantitative locus [eQTL] mapping) in immune cells. The results of these analyses support the hypothesis that a specific network of immune genes has undergone natural selection in populations of European origin—the population in which virtually all of the susceptibility alleles have been discovered—and plays an important role in susceptibility to inflammatory diseases in our current environment. These analyses may contribute to the identification of causal variants and the elucidation of functional mechanisms underlying disease susceptibility; they may also aid in the interpretation of likely agents and targets of positive selection through elucidation of functional effects on specific biological pathways or functional modules.[24]

Specifically, we performed a genome-wide scan for recent positive selection by using long-range haplotype tests and we identified 21 (out of 416) inflammatory-disease variants exhibiting significant signatures of selection in a population of European ancestry. We note that the genes found in these selected loci can be assembled, independently, into a molecular pathway via a protein-protein interaction (PPI) network approach. To identify potential mechanisms for disease susceptibility and potential gene targets of selection, we also performed an eQTL analysis in peripheral blood mononuclear cells (PBMCs) of 228 individuals with an inflammatory demyelinating disease of the central nervous system. Of the 21 susceptibility loci with signatures of selection, we identified 13 variants that exhibit *cis*-regulatory effects in PBMCs, a mixture of cell types with important roles in both adaptive and innate immunity. Further, previously published data on transcriptional response to pathogens provide functional evidence that the affected genes play a significant role in orchestrating pathogen-specific immune response.[25,26] Thus, our results suggest that a subset of genes for inflammatory-disease susceptibility can be assembled into a molecular pathway whose component molecules (1) have experienced a shared evolutionary history in response to selective pressures that may have been acting on host resistance to pathogens and (2) serve a shared molecular function in the immune system that is dysregulated in several inflammatory diseases.

Material and Methods

Genome-wide Association Studies Data

We obtained data from published GWASs via the National Human Genome Research Institute (NHGRI) GWAS catalog available online (December 20, 2011, version). To compile the most recent and comprehensive list, we also mined literature for inflammatory SNPs from GWASs. We used SNPs reported to be associated to complex traits with a p value of at least 5×10^{-8}. A list of the complex traits from the GWAS catalog included for each of the five major trait groups is in Table S3 available online.

To test for enrichment for selection, we selected ten inflammatory diseases: ankylosing spondylitis (AS [MIM 106300]), Crohn disease (CD), celiac disease (CeD), multiple sclerosis (MS [MIM 126200]), primary biliary cirrhosis (PBC [MIM 109720]), psoriasis (PS [MIM 177900]), rheumatoid arthritis (RA [MIM 180300]), systemic lupus erythematosus (SLE [MIM 152700]), type 1 diabetes (T1D [MIM 222100]), and ulcerative colitis (UC [MIM 191390]) (Table S1). These inflammatory diseases were selected because they had the largest repertoire of common, non-MHC variants (validated and replicated variants) at the beginning of this study. In addition, these ten inflammatory diseases were the focus of the ImmunoChip Consortium,[27] which designed a custom-made SNP chip relevant to multiple immune-mediated diseases.

The Integrated Haplotype Score

The integrated haplotype score (iHS) statistic was computed as described in Voight et al.[2]. We used the iHS software to compute genome-wide iHSs on SNPs with minor allele frequency (MAF) >0.05 using the Phased haplotype data from three HapMap II populations: African (Yoruba from Ibadan, Nigeria; YRI), CEU (Utah residents with ancestry from northern and western Europe from the CEPH collection; CEU), and East Asian (Japanese subjects from Tokyo and Han Chinese subjects from Beijing representing Asian populations; ASI). The iHSs can also be retrieved from the Haplotter website.

The CEU HapMap II Phased haplotypes for each chromosome were downloaded from the HapMap FTP site. The ancestral states for HapMap phase II data were based on alignment to the chimpanzee sequence and were downloaded from the Haplotter website. The population-specific recombination map positions were downloaded from the HapMap FTP site. We standardized the iHS values by normalizing the given unstandardized iHS by the mean and standard deviation of other iHSs observed across the genome, with similar derived allele frequencies.

Calculating the F_{ST} Statistic

We calculated unbiased estimates of F_{ST} as described by Weir and Cockerham.[28] We calculated the global F_{ST} for each single SNP among the three HapMap populations. For autosomal HapMap II SNPs, we considered a candidate selection if the SNP $F_{ST} \geq 0.58$, which corresponds to a genome-wide empirical significance level of $\alpha = 0.01$.

Estimating the Age of Selective Sweep

To obtain a rough estimate of the age of the selective sweep, we first computed the extended haplotype homozygosity (EHH)[3] by a web-based tool (Web-EHH) for the chromosomal segments with evidence for selection.[29] Using the EHH scores, we computed the recombination distance r by using the distance (in cM) between the points where EHH $= x$ to the left and to the right of the core SNP. We can obtain r from any given x (i.e., EHH $=$ 0.25). The generation time G can be calculated as $G = (\ln x / -r) \times 100$.[18] Assuming a mean generation time of 25 years, the age of the selective sweep for the derived allele is $25 \times G$.

Protein-Protein Interaction Network

A protein-protein interaction (PPI) network was reconstructed with the online tool DAPPLE with 1,000 permutations and 2 interacting binding degree as a cutoff.[30] The 37 inflammatory

disease SNPs consistent with patterns of recent positive selection ($|iHS| > 1.65$; $p < 0.1$) were included as an input to DAPPLE. DAPPLE then converts into genes based on overlapping wingspan, a region containing SNPs with $r^2 > 0.5$; this region is then extended to the nearest recombination hotspot. DAPPLE reconstructs networks of PPI representing proteins as nodes connected by an edge if there is in vitro evidence of interaction based on the InWeb database, which contains 169,810 high-confidence pair-wise interactions involving 12,793 proteins.[31] DAPPLE reconstructs two types of networks: (1) direct network, two associated proteins can be connected by exactly one edge, and (2) indirect network, where associated proteins can be connected via common interactor proteins with which the associated proteins each share an edge. To assess the statistical significance of the PPI networks, DAPPLE constructs random networks (mimicking the structure of the original network) from a within-degree node-label permutation strategy (see Rossin et al.[32] for further details). An empirical distribution is constructed for each network connectivity parameter and used to evaluate the statistical significance of networks.

Expression QTL Analysis

Gene expression levels were quantified with mRNA derived from peripheral blood mononuclear cells (PBMCs) of 228 subjects of European ancestry with relapsing remitting (RR) multiple sclerosis (MS) via an Affymetrix Human Genome U133 Plus 2.0 Array. These data were collected between July 2002 and October 2007 as part of the Comprehensive Longitudinal Investigation of MS at the Brigham and Women's Hospital. The expression levels were adjusted for confounding factors, such as subject's use of immunomodulatory drugs, age, gender, and batch effects via principle components analysis.[33]

DNA from each individual was genotyped on the Affymetrix 550K GeneChip 6.0 platform as a part of a previously published study.[33] The genotyped data sets were imputed with the CEU samples in HapMap Phase II (~2.5 million SNPs) as the reference.

Associations between SNP genotypes and adjusted expression values were conducted by Spearman rank correlation (SRC). For the *cis* analysis, we considered only SNPs within a 1 MB window from the transcript start site (TSS) of genes. Significance of the nominal p values was determined by comparing the distribution of the most significant p values generated by permuting expression phenotypes 10,000 times independently for each gene.[34] We call a *cis*-eQTL significant if the nominal association p value is greater than the 0.05 tail of the minimal p value distribution resulting from the permuted associations.

Similar methods were used to evaluate the *cis*-regulatory effects in $CD4^+$ T lymphocytes and $CD14^+CD16^-$ monocytes data sets consisting of 80 healthy individuals of European ancestry. These analyses were conducted under the auspices of a protocol approved by the institutional review board of Partners Healthcare.

Results

Loci for Inflammatory-Disease Susceptibility Are Enriched for Signatures of Recent Positive Selection

To test whether inflammatory-disease susceptibility loci are enriched for signals of positive selection, we first compiled a list of single-nucleotide polymorphisms (SNPs) associated with complex traits from the December 10, 2011, version of the National Human Genome Research Institute (NHGRI) GWAS catalog, which includes 1,107 published association studies of 5,481 SNPs (p value $< 10^{-6}$) to more than 550 traits, of which 588 SNPs are associated with ten inflammatory diseases, i.e., ankylosing spondylitis (AS), Crohn disease (CD), ulcerative colitis (UC), celiac disease (CeD), multiple sclerosis (MS), type 1 diabetes (T1D), rheumatoid arthritis (RA), primary biliary cirrhosis (PBC), systemic lupus erythematosus (SLE), and psoriasis (PS) (Table S1). To identify a nonredundant set of associated variants, we LD pruned (keeping the most strongly associated SNP and removing those with an $r^2 > 0.4$) the 588 SNPs, which resulted in 416 independent associations, and this set was used for all the subsequent analysis. All of the discovery studies used subjects of European ancestry.

To quantify signatures of positive selection in the human genome, we searched for regions of the genome where the pattern of genetic variation is consistent with that expected under a model of recent positive selection. For our analysis we use the integrated haplotype score (iHS) as our primary measure of selection. The iHS statistic was primarily developed to detect positive selection and is sensitive to detecting soft sweeps or incomplete sweeps. Many of the GWAS SNPs are probably evolving under such sweeps.[10,35] The iHS statistic uses the lengths of the haplotypes surrounding each core SNP to identify SNPs for which alleles have rapidly risen in frequency.[2,3] We applied iHS to genome-wide HapMap Phase II SNPs (MAF > 0.05) in the CEU population (Utah residents with ancestry from northern and western Europe from the CEPH collection). As a secondary analysis, we also examined differences in allele frequencies (global F_{ST}) between the three HapMap populations (subjects of East Asian [Han Chinese from Beijing and Japanese from Tokyo], European [CEU], and African [Yoruba from Ibadan, Nigeria] ancestry) at disease-associated SNPs. Because F_{ST} provides a metric of the magnitude of global allele frequency differentiation, it is less sensitive to detecting population-specific positive selection signals than iHS,[36] and it is these population-specific signals that are most pertinent in our evaluation of variants associated with disease in European populations.

In our primary analysis, we therefore searched for inflammatory-disease SNPs that localize to regions of the genome where patterns of genetic variation are consistent with evidence for recent positive selection. Of 416 SNPs associated with inflammatory diseases that are considered in our analyses, 21 SNPs exhibited a significant signal of selection at $p < 0.05$ ($|iHS| > 2$) after correcting for genome-wide testing, with 13 loci meeting or exceeding the more rigorous corrected threshold of $p < 0.01$ ($|iHS| > 2.6$) (Table 1). To assess how extreme these results are, we also performed simulations to test for an enrichment of positive selection among SNPs associated with inflammatory disease relative to other SNP sets: we generated 10,000 SNP sets sampled at random from bins matched for minor allele frequency (MAF), haplotype block length, and genic/nongenic status to the 416 susceptibility variants. With this distribution, the 13

Table 1. Inflammatory Disease-Associated Variants that Localize to Regions with Evidence of Positive Selection

Chr	SNP	Disease[a]	Derived/Ancestral Allele	Risk Allele[b]	Derived Allele Freq.	F_{ST}	iHS	Genes within 50 kb
7q11	rs1167796	SLE	G/A	G	0.53	0.24	5.64	*HIP1* (MIM 601767)
6p21	rs3131379	SLE	A/G	A*	0.10	0.06	−3.18	*VARS* (MIM 192150), *LSM2* (MIM 607282)
12q24	rs17696736	T1D	G/A	G*	0.35	0.35	−3.10	*TMEM116*
19q13	rs281379	CD	G/A	A*	0.42	0.45	2.98	*FUT2* (MIM 182100), *FUT1* (MIM 211100)
12q24	rs3184504	RA, T1D	T/C	T*	0.41	0.40	−2.97	*SH2B3* (MIM 605093), *ATXN2* (MIM 601517)
6q21	rs11962089	MS	G/A	G	0.16	0.47	2.81	*POPDC3* (MIM 605824)
4q27	rs6822844	CeD, RA, UC	T/G	G	0.21	0.20	−2.82	*IL2* (MIM 147680), *IL21* (MIM 605384)
3q13	rs4308217	MS	A/C	C	0.43	0.45	−2.73	*CD86* (MIM 601020)
3q25	rs12638253	MS	T/C	NR	0.48	0.20	−2.71	*LEKR1* (MIM 613536)
3q25	rs17810546	CeD	G/A	G*	0.10	0.09	−2.66	*SCHIP1* (MIM 611622), *IL12A* (MIM 161560)
3q13	rs1132200	MS	T/C	NR	0.12	0.06	−2.68	*ARHGAP31* (MIM 610911), *STAT1* (MIM 600555)
19q13	rs307896	MS	T/C	C	0.23	0.41	−2.62	*SAE1* (MIM 613294)
5q31	rs2188962	CD	T/C	T*	0.45	0.43	−2.60	*SLC22A5* (MIM 603377), *IRF1* (MIM 147575)
16p13	rs12708716	MS, T1D	G/A	A*	0.28	0.03	2.32	*CLEC16A* (MIM 611303)
6p21	rs11755393	SLE	A/G	NR	0.63	0.12	−2.28	*UHRF1BP1* (MIM 612253)
2q37	rs10210302	CD	C/T	T*	0.45	0.14	2.18	*INPP5D* (MIM 601582), *ATG16L1* (MIM 610767)
6q27	rs415890	CD	G/C	NR	0.48	0.12	2.15	*RNASET2* (MIM 612944)
6p21	rs3129934	MS	T/C	T*	0.25	0.09	−2.12	*BTNL2* (MIM 606000)
5p13	rs10440635	AS	G/A	A*	0.36	0.43	2.03	*PTGER4* (MIM 601586)
19q13	rs3745516	PBC	A/G	A	0.20	0.49	2.01	*SPIB* (MIM 606802)
21q22	rs2838519	UC	A/G	G*	0.49	0.25	2.00	*ICOSLG* (MIM 605717)
20q13	rs2248359	MS	C/T	C*	0.63	0.24	−1.93	*CYP24A1* (MIM 126065)
16p13	rs6498169	MS	A/G	G*	0.61	0.06	1.90	*CLEC16A* (MIM 611303)
11q23	rs678170	UC	A/G	NR	0.69	0.20	1.90	*FAM55A, FAM55D*
16q12	rs5743289	CD	C/T	T*	0.73	0.26	1.90	*SNX20* (MIM 613281), *NOD2* (MIM 605956)
2q12	rs2058660	CD	G/A	G*	0.23	0.08	−1.88	*IL18RAP* (MIM 604509)
10p15	rs12722489	CD, MS	T/C	C	0.16	0.07	−1.88	*IL2RA* (MIM 147730)
10q24	rs10786436	T1D	T/C	T	0.34	0.08	1.82	*HPSE2* (MIM 613469)
2p14	rs17035378	CeD	T/C	NR	0.70	0.16	1.82	*PLEK* (MIM 173570)
15q23	rs17374222	RA	A/C	A*	0.49	0.08	−1.81	*KIF3* (MIM 604683)
6p21	rs2647044	T1D	G/A	A*	0.91	0.02	1.78	*HLA-DQB1* (MIM 604305), *HLA-DQA2* (MIM 613503)
6q23	rs11154801	MS	A/C	A	0.32	0.15	1.77	*AHI1* (MIM 608894)
6p22	rs2285797	MS	G/A	NR	0.32	0.06	−1.77	*TRIM10* (MIM 605701)
2q32	rs10931468	PBC	C/A	A*	0.82	0.10	1.74	*STAT4* (MIM 600558)

(*Continued on next page*)

Table 1. *Continued*

Chr	SNP	Disease[a]	Derived/Ancestral Allele	Risk Allele[b]	Derived Allele Freq.	F_{ST}	iHS	Genes within 50 kb
4q21	rs4333130	AS	C/T	NR	0.45	0.66	1.74	*ANTXR2* (MIM 608041)
17q21	rs744166	CD, MS	A/G	A*	0.56	0.10	−1.72	*STAT3* (MIM 102582)
5p13	rs6897932	MS, T1D	C/T	C*	0.24	0.11	−1.66	*IL7R* (MIM 146661)

Abbreviations are as follows: iHS, integrated haplotype score; NR, not reported; Freq., frequency.
[a]Abbreviations are as follows: AS, ankylosing spondylitis (MIM 106300); CD, Crohn disease (MIM 266600); CeD, celiac disease (MIM 212750); MS, multiple sclerosis (MIM 126200); PBC, primary biliary cirrhosis (MIM 109720); PS, psoriasis (MIM 177900); RA, rheumatoid arthritis (MIM 180300); SLE, systemic lupus erythematosus (MIM 152700); T1D, type 1 diabetes (MIM 222100); UC, ulcerative colitis (MIM 191390).
[b]An asterisk (*) indicates that the risk allele lies on the positively selected haplotype.

haplotypes with putative positively selected SNPs meeting an $|iHS| > 2.6$ threshold are unlikely to have been observed by chance alone ($p < 0.0001$; Figure 1, left). This result suggests that loci for inflammatory-disease susceptibility are more likely to be located on positively selected haplotypes than matched SNP sets.

We have also performed an analysis to assess whether the loci harboring risk variants associated with common inflammatory diseases are significantly enriched for signals of selection relative to immune response genes that are not known to be associated with susceptibility to inflammatory disease. In this analysis, we find that, at an $|iHS| > 2.6$, there are 3% (13/413 SNPs) of susceptibility variants under positive selection compared to 0.6% (23/3,654 SNPs) of LD-pruned variants found in the vicinity of genes previously defined as being involved in immune function[37] ($p < 0.0001$) (see Supplemental Data for details of the analysis; Figure S1). This result suggests that risk variants for inflammatory diseases are significantly enriched for signatures of selection relative to variants in immune genes that influence immune function but are not associated with disease (Figures S1 and S2).

Interestingly, we find that 21 of the 30 reported risk alleles (out of 37; Table 1) with evidence of selection lie on the positively selected haplotypes, suggesting that the disease-associated risk alleles are, on average, more likely to be targets of selection than the protective allele (binomial test $p = 0.043$). However, we note that these analyses are complicated by the fact that some variants are associated with more than one inflammatory disease, and indeed there are examples where a given allele is protective against one disease and a risk allele for another.[38,39]

In a secondary analysis, we deployed the F_{ST} metric, a different method for assessing evidence of natural selection, and compared the number of SNPs associated with inflammatory diseases in the 1% tail of global F_{ST} to a distribution comprised of MAF- and genic/nongenic status-matched SNPs sampled at random 10,000 times from the empirical distribution; here, we find that inflammatory disease SNPs are more highly differentiated across human populations than expected by chance (observed nine SNPs meeting this threshold; $p = 0.006$; Figure 1, right). Of the 13 putative positively selected SNPs in the $|iHS|$ analysis, 10 also have high global F_{ST} score (above genome-wide average; $F_{ST} > 0.11$), suggesting that these variants might represent the true variants that are under selection.

To determine whether SNPs associated with inflammatory diseases show greater enrichment for signatures of selection relative to those associated with other disease classes, we compared the overall distribution of iHS across different classes of complex traits. By using the 5,481 SNPs from the GWAS catalog, we selected SNPs associated with one of five trait classes (inflammatory disease, cancer, neurological/psychiatric disease, metabolic disease, and height), and we LD pruned variants within each group by removing SNPs with an $r^2 > 0.4$ with a SNP with a higher reported p value for a given trait (Table S2).

Although we do not observe a major shift in the entire distribution of iHSs between inflammatory disease SNPs and those of the other trait classes, we note that variants associated with inflammatory diseases may be somewhat enriched in the 1% tail of the empirical distribution for iHS (data not shown). So, we tested for enrichment among inflammatory-disease SNPs at the top 5% and 1% of iHS (Figure 2). Among the top 1% of iHS signals ($|iHS| > 2.6$), we observe that inflammatory diseases have a higher proportion of SNPs (3.1%) targeted by positive selection than the other trait groups (Figure 2). This enrichment is significant when compared to the distribution observed in all SNPs genotyped in CEU subjects of the HapMap Phase II resource (Pearson's χ^2 test, $p = 1.1 \times 10^{-5}$). Thus, overall, it appears that the class of inflammatory-disease variants has experienced more selection in subjects of European ancestry than is expected by chance alone.

Positively Selected Genes Encode Proteins that Directly Interact in PPI Network

Having observed robust evidence of natural selection in a subset of loci for inflammatory-disease susceptibility, we next investigated whether the selected loci might serve a shared cellular function that could have been the target of selective pressure over the course of human history. We used an approach that leverages protein-protein interaction maps to see whether the selected loci could be assembled into a single network of genes that have a correlated evolutionary history.[32,40,41] We tested this hypothesis by using a robust method for the assessment

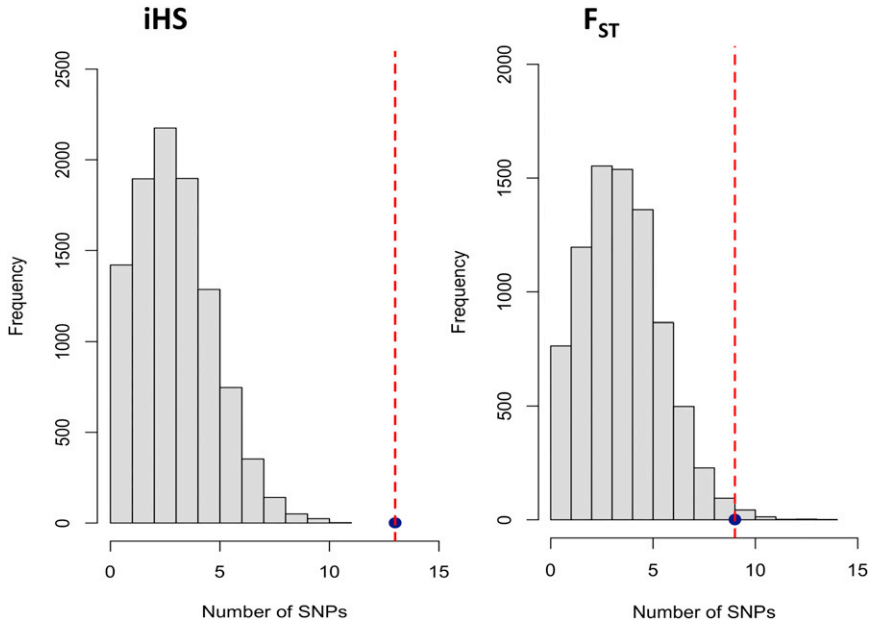

Figure 1. Inflammatory-Disease SNPs Are More Likely to Have Experienced Recent Positive Selection

The distribution of the number of SNPs with |iHS| > 2.6 (left) and highly differentiated SNPs (top 1% of the empirical genome-wide F_{ST} distribution) (right) observed for each of 10,000 draws of 416 SNPs matched for MAF, haplotype length, and genic/nongeneic status to the 416 inflammatory disease SNPs (see Table S1 for a list of diseases). Shown in solid blue circle are the observed number of SNPs with |iHS| > 2.6 (13 SNPs; $p < 0.0001$) and high F_{ST} (8 SNPs; $p = 0.006$). A dotted red line further highlights this result.

of functional connectivity, the disease association protein-protein link evaluator (DAPPLE) algorithm, that requires evidence of functional connectivity as well as coexpression in at least one tissue.[32] DAPPLE defines a genomic region around each SNP in terms of LD and considers any protein encoded within that region. We pursued the DAPPLE analyses with a slightly more relaxed threshold (|iHS| > 1.6) to enhance the dissection of the pathways. Using more relaxed threshold in the network analyses is one approach to identify which of the loci with suggestive evidence of selection share certain functional information with the significant loci.

We found that 15 of 37 putative positively selected loci ($p < 0.1$; |iHS| > 1.6) are directly connected and 28 loci are indirectly connected (via an interactor protein that is not known to be under selection). These results suggest the existence of functional connections between proteins encoded in the selected regions. When compared to 10,000 random networks, we found that the direct network connectivity was statistically significant beyond that expected by chance (permuted $p = 9.91 \times 10^{-5}$, Figure 3). The model for indirect connectivity via common interactors that are not among the most selected loci was also significant (permuted $p = 0.0027$; Figure S3).

Our observation that there are significant numbers of direct connections between proteins in inflammatory-disease loci under selection is not surprising, as it has been previously shown that genes in loci associated with inflammatory diseases (i.e., RA and CD) are significantly connected via protein-protein interactions.[32] Whereas these studies have identified elements of the network that we present (Figure 3), we explored the possibility that the proteins encoded in loci for inflammatory-disease susceptibility that have evidence for selection are more densely connected than those located near random sets of loci for inflammatory-disease susceptibility. Connectivity within the latter, randomly selected pools of disease-associated SNPs may simply reflect the more general connectivity of immune pathways.[32] To statistically evaluate the connectivity of the positively selected PPI network, we therefore generated 1,000 random sets of 37 SNPs (to match the number of putatively selected loci) from the 416 SNPs associated with inflammatory diseases and constructed PPI networks from each set. With these results, we observe that the mean direct connectivity for the PPI generated from selected variants (Figure 3) falls in the top 1% of the empirical distribution for random inflammatory-disease susceptibility networks and is unlikely to have occurred by chance alone ($p < 0.014$; Figure S4). This result suggests that proteins encoded in the positively selected inflammatory-disease loci are more highly interconnected than those encoded by loci for inflammatory-disease susceptibility in general. Given their similar evolutionary history and the independent evidence of functional connectivity between them, this group of selected susceptibility genes is more likely to serve a shared molecular function that was advantageous in response to a selective pressure encountered by European populations over the course of their history.

To investigate whether the genes of the core selected pathway have been targets of selection in response to a shared selective pressure, we estimated the age of the haplotypes carrying the selected risk alleles to see whether their ages converge. To estimate the age of a putative selective event, we first calculated the extended haplotype homozygosity (EHH) for each haplotype, assuming a star phylogeny of the haplotypes and an average generation time of 25 years (see Material and Methods). In the European populations where the susceptibility alleles have been discovered, the estimated age of a putative selective sweep for the components in the core, directly connected network (Figure 3) is in the range of 1,200–2,600 years ago (estimates of the age of the selective sweep for the derived allele are listed in Table S4). These estimates are coarse approximations, and this method is particularly well powered for recent selective events. Appreciating these

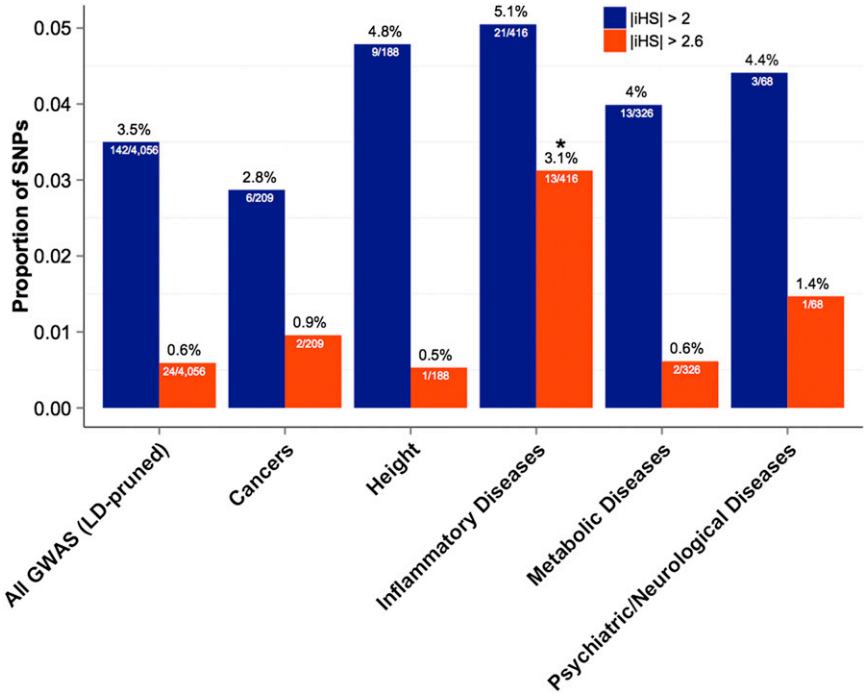

Figure 2. Proportion of Trait-Associated SNPs that Localize to Regions with Evidence of Recent Positive Selection

Shown here are GWAS SNPs from five major trait groups: all GWAS SNPs (LD pruned; n = 4,056), cancers (n = 209), height (n = 188), inflammatory diseases (n = 416), metabolic diseases (n = 326), and neurological/psychiatric diseases (n = 68). The trait-associated SNPs are LD pruned to $r^2 < 0.4$; shown here are SNPs with |iHS| > 2 in blue bars and |iHS| > 2.6 in red bars. The proportion of SNPs at different iHS thresholds are shown in white inside each bar. Inflammatory diseases have a higher proportion of SNPs (3.1% at iHS > |2.6|) targeted by positive selection than other trait groups (Pearson's χ^2 test, $p = 1.1 \times 10^{-5}$). We highlight this category with an asterisk.

limitations, it nonetheless appears that the selective forces influencing allele frequencies in the core selected network that we have identified occurred over the course of recent, recorded human history.

Positively Selected Inflammatory Risk Variants Have *cis*-Regulatory Effects in Activated Immune Cells

To examine the mechanism(s) by which the selected variants may influence immune function, we assessed each of the 21 inflammatory disease loci with |iHS| > 2 for evidence of an effect on RNA expression levels in *cis*. That is, we performed an expression quantitative trait locus (eQTL) analysis in each locus for genes in the vicinity of the disease-associated SNP. We first investigated genetic variation and mRNA expression levels in a data set derived from peripheral blood mononuclear cells (PBMCs) of 228 individuals of European ancestry with an inflammatory demyelinating disease of the central nervous system (a clinically isolated demyelinating syndrome or multiple sclerosis), representing a set of subjects with an activated immune system. To investigate our findings in other immune-relevant cell types, we also performed eQTL analyses in independent data sets generated from flow-sorted $CD4^+$ T lymphocytes and $CD14^+CD16^-$ monocytes (both of which are constituents of the PBMC mixture) of 80 healthy subjects from the Immunological Variation (ImmVar) project and also evaluated previously published eQTL data sets including HapMap LCLs,[42] $CD4^+$ T lymphocytes from asthmatics,[43] and monocytes from healthy individuals.[30] These genome-wide eQTL data from multiple immune cell types enhance our ability to identify functional effects of genetic variation on gene expression levels and allow us to assemble a framework for understanding which cells and cell functions are affected by each variant.

Of the 21 inflammatory-disease variants exhibiting signatures of selection, 13 variants had *cis*-regulatory effects on gene expression levels in the available data (Table 2). Among the inflammatory diseases, MS-, CD-, and T1D-associated risk variants were the most abundant among those exhibiting selection and eQTL effects, most probably due to large number of variants discovered for these diseases (Table 2). One locus with overlapping signals (selection, eQTL, and disease association) contains the SLE risk allele $rs11755393^G$, which is under selection (iHS = −2.28) and is strongly associated with PBMC RNA expression of the ICBP90 binding protein 1 (*UHRF1BP1* [MIM 612253]; $p = 7.53 \times 10^{-42}$), a putative binding partner of UHRF1, a RING-finger type E3 ubiquitin ligase[44] (Figure 4). For this locus, we further replicated the *cis*-regulatory effect in data generated from $CD4^+$ T lymphocytes ($p = 5.65 \times 10^{-10}$; Figure S5) and $CD14^+$ $CD16^-$ monocytes ($p = 3.57 \times 10^{-10}$; Figure S6) of 80 healthy individuals of European ancestry: in all three data sets, the SLE susceptibility allele $rs11755393^G$ is associated with decreased *UHRF1BP1* RNA expression. In PBMCs, the regulatory trait concordance scores (RTCs),[46] an index integrating eQTL and GWAS data to detect disease-causing regulatory effects, were equal to one (RTC = 1), suggesting that the genetic variant underlying disease susceptibility and *cis*-regulatory signals are the same. Other genes with *cis*-eQTLs and signals of selection are listed in Table 2.

Discussion

The results presented here provide support for recent positive selection having shaped a portion of genetic variation influencing inflammatory-disease susceptibility to

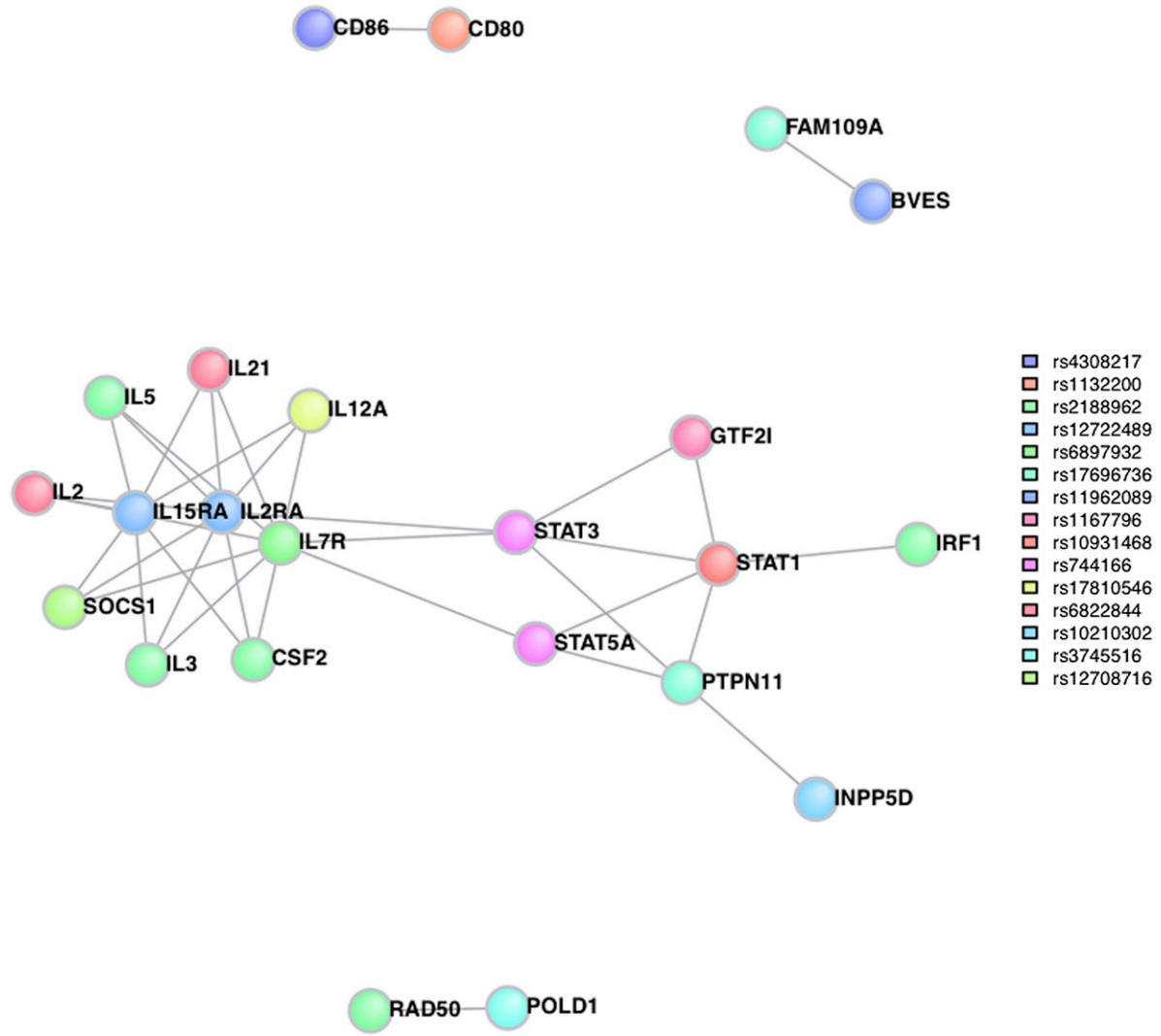

Figure 3. Directly Connected Protein-Protein Interaction Network
The protein-protein interaction (PPI) network was generated from genes found in inflammatory disease haplotypes with evidence for positive selection. Human RefSeq Genes are represented as nodes (colored by locus); edges indicate known direct functional connection according to the PPI algorithm.[32] The network is statistically significant for direct connectivity more than would be expected by chance (mean direct connectivity, 3.13; expected, 1.18; permuted $p = 9.91 \times 10^{-5}$).

a greater extent than genetic variation associated with other common diseases. We show that several genes found in these disease-associated, positively selected haplotypes can be assembled, independently, into a molecular complex via a protein-protein interaction (PPI) network approach. This network of selected genes is significantly more connected than random networks of genes for inflammatory-disease susceptibility, suggesting that these selected genes may have served a shared molecular mechanism that experienced selective pressure over human history. More specifically, the core of the directly connected, selected protein network (Figure 3) captures molecular pathways involved in lymphocyte activation. Although many of these genes are expressed in different subsets of immune cells and our understanding of the role of specific genes across immune cell types is incomplete, a subset of T cells stands out as being influenced by a majority of the genes in the directly interconnected network: pathogenic Th17 and Th1 cells that play an important role in several inflammatory diseases such as Crohn disease, multiple sclerosis, rheumatoid arthritis, and type 1 diabetes mellitus.[47–50]

A careful balance exists among Th17, Th1, Th2, regulatory T (Treg) cells, and other immune cells to maintain the homeostasis of the immune system. In inflammatory diseases, this balance is disrupted, with the frequency and function of these cell types being altered. In the core network of 17 directly connected genes (Figure 3), derived from inflammatory disease susceptibility genes with evidence of selection, we find genes that (1) induce Th17 cell differentiation (*IL21* [MIM 605384], *STAT3* [MIM 102582]), (2) prevent Th17 cell differentiation (*IL12A* [MIM 161560]), (3) regulate Th17 cell function (*IRF1* [MIM 147575], *STAT1* [MIM 600555]), (4) are involved

Table 2. Inflammatory Disease-Associated Variants that Localize to Regions with Evidence for *cis*-Regulatory Function and Positive Selection

	Locus												
	7q11	**6p21**	**12q24**	**19q13**	**12q24**	**3q13**	**19q13**	**3q13**	**5q31**	**16p13**	**6p21**	**2q37**	**6q27**
SNP	rs1167796	rs3131379	rs17696736	rs281379	rs3184504	rs4308217	rs307896	rs1132200	rs2188962	rs12708716	rs11755393	rs10210302	rs415890
Disease	SLE	SLE	T1D	CD	RA, T1D	MS	MS	MS	CD	MS, T1D	SLE	CD	CD
Minor (risk) allele	A(G)	A(A)	G(G)	A(A)	T(T)	A(C)	A(G)	A(NR)	T(T)	G(G)	G(NR)	T(T)	C(C)
MAF	0.29	0.05	0.2	0.27	0.21	0.19	0.13	0.12	0.17	0.34	0.45	0.39	0.4
iHS	5.64	−3.18	−3.1	2.98	−2.97	−2.73	−2.62	−2.54	−2.34	2.32	−2.28	2.18	2.15
cis-regulated gene	*PMS2L3*	*VARS2*	*C12orf47*	*LILRB5*	*TMEM116*	*CD86*	*SAE1*	*POGLUT1*	*IRF1*	*DEXI*	*UHRF1BP1*	*ATG16L1*	*RNASET2*
eQTL Cell Types[a]													
This Study													
PBMC (n = 228)	+	–	+	–	+	–	–	–	+	+	+	–	+
RTC (PBMC)	0.91	0.78	0.62	NA	0.76	NA	NA	NA	0.85	0.73	1	NA	0.94
$CD14^{+}CD16^{-}$ Monocytes (n = 80)	–	–	–	+	–	–	–	–	+	–	+	–	+
$CD4^{+}$ T Cells (Healthy) (n = 80)	–	–	–	–	–	–	–	–	–	–	+	–	+
Published													
Blood Monocytes (n = 1,490)[30]	+	+	+	+	–	–	–	–	–	+	+	–	+
$CD4^{+}$ T Cells (Asthmatics) (n = 200)[43]	–	–	–	–	–	–	+	+	+	–	–	+	+
LCLs (CEU HapMap) (n = 109)[42]	–	–	–	–	–	–	–	–	–	–	–	+	–
LCLs (Asthmatics) (n = 400)[60]	–	–	–	–	–	+	–	+	–	–	–	+	–

Abbreviations are as follows: iHS, integrated haplotype score; MAF, minor allele frequency; RTC, relative trait concordance, a score integrating eQTL and GWAS data to detect disease-causing *cis*-regulatory effects. RTC scores are calculated only in PBMC.

[a]Significant *cis*-eQTLs are indicated by "+." The PBMC is our primary data set and for replication we used gene expression data from $CD14^{+}CD16^{-}$ monocytes and $CD4^{+}$ T cells of 80 healthy subjects (data not shown). The other *cis*-eQTL results presented here were reported as significant in the published studies that we reference.

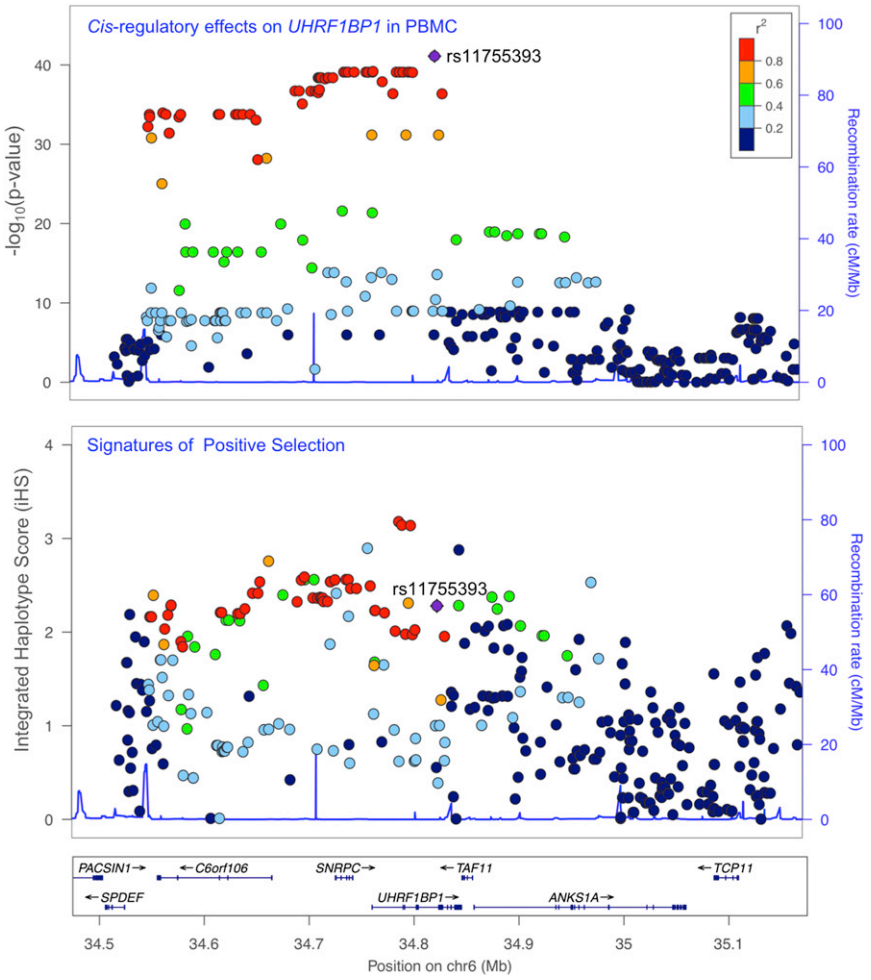

Figure 4. Colocalization of *cis*-Regulatory Effects and Positive Selection Signal in the *UHRF1BP1* Locus

The top panel reports the *cis*-regulatory effects of each SNP on *UHRF1BP1* RNA expression in PBMC; $-\log_{10}$ (p value) is reported on the *y* axis. The lower panel reports the scores for evidence of positive selection (iHS) for each SNP over the same chromosomal segment. The RefSeq genes in the region are shown at the bottom of the figure. The LD (in r^2) for each SNP with the index SLE-associated *UHRF1BP1* SNP (rs11755393; shown in purple) is illustrated with the use of colors, as indicated in the top right of the figure. This figure was generated by LocusZoom.[45]

in Th17 cell function (*CSF2* [MIM 138960], *IL3* [MIM 147740], *STAT3* [MIM 102582]), and (5) induce and maintain Treg cells that inhibit Th17 and other pathogenic cells (*IL2* [MIM 147730], *IL2RA* [MIM 147730], *SOCS1* [MIM 603597], *STAT5* [MIM 601511]).[51] Several of these genes also influence the function of other immune cell subsets, but the overlay of our core selected network with molecular pathways that are important for Th17 and Treg cell function is consistent with an extensive, recent immunologic literature that has defined a prominent role of these cell types in inflammatory diseases.[47,52] At this time, it is not clear what selective pressure may have led to the signals of natural selection that we have uncovered in populations of European ancestry: favorable responses to one or more pathogen during the history of European populations is a reasonable scenario, particularly since the Th17 cell pathway that emerges from our analysis has evolved to confer protective immunity against extracellular bacteria as well as intracellular pathogens.[53–56]

Although it is impossible to recreate natural selection in vitro, functional experiments examining cellular response to pathogens have helped to characterize pathogen-specific host response networks. We find that four genes with evidence for selection and *cis*-eQTL effects (*IRF1*, *PLEK* [MIM 173570], *DEXI*, *CD86* [MIM 601020]) are part of regulatory networks that control the transcriptional response to a viral or bacterial construct in mouse primary dendritic cells (DCs).[25] In addition, several of the selected genes (*TMEM116, UHRF1BP1, IRF1, DEXI, RNASET2* [MIM 612944], *CD86, PLEK*) are significantly differentially expressed in a recent study of 65 individuals whose primary dendritic cells (DCs) were challenged with *Mycobacterium tuberculosis* (MTB) infection.[26] Finally, three of the selected genes (*FUT2*, *SH2B3*, *IRF1*) contain variants that are highly correlated with current estimates of local pathogen diversity,[8] lending additional support for a role of pathogens in shaping levels and patterns of genetic variation at loci influencing inflammatory-disease susceptibility.

It is perhaps not surprising that many of the selected regions harbor genes and molecular pathways that are directly involved in immune response to pathogens. Infectious diseases have been one of the major causes of death throughout human history. Since the agricultural revolution, human populations expanded rapidly and developed resistance to a number of microbial and parasitic pathogens, and as a consequence, natural selection probably played a major role in shaping human genetic variation. In the context of inflammatory diseases, it is plausible that signatures of recent positive selection observed in the susceptibility alleles are due to an overactive immune system, driven at least in part by immune response to pathogen exposure in the past. This is consistent with the "hygiene hypothesis," which postulates that lack of exposure to infectious agents in the modern environment increases susceptibility to inflammatory disease as a result of imbalance of the immune system.[22]

Given the timing of the putative selective event exerting an influence on the core selected PPI network

(1,200–2,600 years ago) (Figure 3; Table S4) and the availability of recorded history in European populations during this time, we can speculate as to the nature of the selective pressure. The confidence interval on these estimates is large (400–700 years), and thus we must be cautious in interpreting the estimated dates for the selective events. Although many pathogens were doubtlessly influencing human populations during the time period of selection that we estimate, few had as dramatic effects as the recurrent epidemics of bubonic plague in European populations. Although intriguing, the convergence of our results with this historical record should be seen as speculative because our understanding of the role of pathogens during this period is incomplete. Nonetheless, this scenario is biologically plausible because the core selected network that influences Th17 and other cells of the adaptive immune system has been implicated in resistance to *Yersinia pestis* (the organism causing plague) in a mouse model.[57]

Overall, our integrated analyses contribute to understanding the organization of the large collection of susceptibility genes for inflammatory diseases, many of which are shared by several diseases:[32,38,58,59] it provides a framework with which to generate hypotheses and design new experiments to better understand how genetic variation contributing to variation in immune function via alterations in gene expression and other mechanisms ultimately leads to immune dysfunction in some individuals as they encounter environmental risk factors in our modern environment. For example, our core network of interconnected, selected susceptibility genes highlights several genes not known to be involved in pathogenic Th17 cell differentiation, function, or regulation. Our data suggest that these genes should be investigated in this context, and our eQTL analyses suggest that alterations in gene expression may be one mechanism by which several of the selected variants might influence immune function. These results may also be helpful in refining causal variants within each locus for future experimental validation and characterization. Finally, our analyses of natural selection provide a framework for the remarkable extent of shared genetic architecture among inflammatory diseases: it may be the by-product of beneficial variation in key molecular pathways influencing the activation of the immune system in response to pathogens that now influences aberrant response to self-antigens. The core selected network (Figure 3) contains genes that influence susceptibility to multiple different diseases and may be critical to setting the likelihood of responses to self-antigens, whereas disease-specific susceptibility loci may play a greater role in the syndromic manifestations of an immune reaction against self.

Supplemental Data

Supplemental Data include six figures and four tables and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

We thank Christophe Benoist for his leadership on RC2 GM093080. We thank Michelle Lee for sample collection and Katherine Rothamel for data generation. The authors are grateful to Vijay Kuchroo, Eli Stahl, and Chris Cotsapas for comments on a previous version of the manuscript. This work is supported by the National Institutes of Health (NIH) (RC2 GM093080, R01 NS067305, and F32 AG043267). P.L.D. is a Harry Weaver Neuroscience Scholar of the National Multiple Sclerosis Society (JF2138A1).

Received: October 9, 2012
Revised: November 13, 2012
Accepted: March 1, 2013
Published: March 21, 2013

Web Resources

The URLs for data presented herein are as follows:

DAPPLE, http://www.broadinstitute.org/mpg/dapple/dapple.php
GWAS Catalog, http://www.genome.gov/gwastudies/
Haplotter, http://haplotter.uchicago.edu/
iHS software, http://hgdp.uchicago.edu/Software/
InnateDB, http://www.innatedb.ca/
International HapMap Project, http://hapmap.ncbi.nlm.nih.gov/
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org/
RefSeq, http://www.ncbi.nlm.nih.gov/RefSeq

Accession Numbers

The data are available on the Gene Expression Omnibus website (GSE16214).

References

1. Barreiro, L.B., and Quintana-Murci, L. (2010). From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat. Rev. Genet. *11*, 17–30.
2. Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. PLoS Biol. *4*, e72.
3. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al.; International HapMap Consortium. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature *449*, 913–918.
4. Nielsen, R., Williamson, S., Kim, Y., Hubisz, M.J., Clark, A.G., and Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. Genome Res. *15*, 1566–1575.
5. Barreiro, L.B., Laval, G., Quach, H., Patin, E., and Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. Nat. Genet. *40*, 340–345.
6. Pickrell, J.K., Coop, G., Novembre, J., Kudaravalli, S., Li, J.Z., Absher, D., Srinivasan, B.S., Barsh, G.S., Myers, R.M., Feldman, M.W., and Pritchard, J.K. (2009). Signals of recent positive selection in a worldwide sample of human populations. Genome Res. *19*, 826–837.
7. Fumagalli, M., Pozzoli, U., Cagliani, R., Comi, G.P., Bresolin, N., Clerici, M., and Sironi, M. (2010). Genome-wide

identification of susceptibility alleles for viral infections through a population genetics approach. PLoS Genet. *6*, e1000849.
8. Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admetlla, A., Pattini, L., and Nielsen, R. (2011). Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genet. *7*, e1002355.
9. Corona, E., Dudley, J.T., and Butte, A.J. (2010). Extreme evolutionary disparities seen in positive selection across seven complex diseases. PLoS ONE *5*, e12236.
10. Casto, A.M., and Feldman, M.W. (2011). Genome-wide association study SNPs in the human genome diversity project populations: does selection affect unlinked SNPs with shared trait associations? PLoS Genet. *7*, e1001266.
11. Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D., and Lander, E.S. (2006). Positive natural selection in the human lineage. Science *312*, 1614–1620.
12. Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., and Clark, A.G. (2007). Recent and ongoing selection in the human genome. Nat. Rev. Genet. *8*, 857–868.
13. Walsh, E.C., Sabeti, P., Hutcheson, H.B., Fry, B., Schaffner, S.F., de Bakker, P.I., Varilly, P., Palma, A.A., Roy, J., Cooper, R., et al. (2006). Searching for signals of evolutionary selection in 168 genes related to immune function. Hum. Genet. *119*, 92–102.
14. Akey, J.M., Zhang, G., Zhang, K., Jin, L., and Shriver, M.D. (2002). Interrogating a high-density SNP map for signatures of natural selection. Genome Res. *12*, 1805–1814.
15. Akey, J.M. (2009). Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res. *19*, 711–722.
16. Franke, A., McGovern, D.P., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J., Roberts, R., et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat. Genet. *42*, 1118–1125.
17. Carlsson, B., Kindberg, E., Buesa, J., Rydell, G.E., Lidón, M.F., Montava, R., Abu Mallouh, R., Grahn, A., Rodríguez-Díaz, J., Bellido, J., et al. (2009). The G428A nonsense mutation in FUT2 provides strong but not absolute protection against symptomatic GII.4 Norovirus infection. PLoS ONE *4*, e5593.
18. Zhernakova, A., Elbers, C.C., Ferwerda, B., Romanos, J., Trynka, G., Dubois, P.C., de Kovel, C.G., Franke, L., Oosting, M., Barisani, D., et al.; Finnish Celiac Disease Study Group. (2010). Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. Am. J. Hum. Genet. *86*, 970–977.
19. Hughes, A.L., and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature *335*, 167–170.
20. Prugnolle, F., Manica, A., Charpentier, M., Guégan, J.F., Guernier, V., and Balloux, F. (2005). Pathogen-driven selection and worldwide HLA class I diversity. Curr. Biol. *15*, 1022–1027.
21. Qutob, N., Balloux, F., Raj, T., Liu, H., Marion de Procé, S., Trowsdale, J., and Manica, A. (2012). Signatures of historical demography and pathogen richness on MHC class I genes. Immunogenetics *64*, 165–175.
22. Strachan, D.P. (1989). Hay fever, hygiene, and household size. BMJ *299*, 1259–1260.
23. Strachan, D.P. (2000). Family size, infection and atopy: the first decade of the "hygiene hypothesis". Thorax *55*(*Suppl 1*), S2–S10.
24. Stranger, B.E., Stahl, E.A., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. Genetics *187*, 367–383.
25. Amit, I., Garber, M., Chevrier, N., Leite, A.P., Donner, Y., Eisenhaure, T., Guttman, M., Grenier, J.K., Li, W., Zuk, O., et al. (2009). Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. Science *326*, 257–263.
26. Barreiro, L.B., Tailleux, L., Pai, A.A., Gicquel, B., Marioni, J.C., and Gilad, Y. (2012). Deciphering the genetic architecture of variation in the immune response to *Mycobacterium tuberculosis* infection. Proc. Natl. Acad. Sci. USA *109*, 1204–1209.
27. Cortes, A., and Brown, M.A. (2011). Promise and pitfalls of the Immunochip. Arthritis Res. Ther. *13*, 101.
28. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. Evolution *38*, 1358–1370.
29. Mueller, J.C., and Andreoli, C. (2004). Plotting haplotype-specific linkage disequilibrium patterns by extended haplotype homozygosity. Bioinformatics *20*, 786–787.
30. Zeller, T., Wild, P., Szymczak, S., Rotival, M., Schillert, A., Castagne, R., Maouche, S., Germain, M., Lackner, K., Rossmann, H., et al. (2010). Genetics and beyond—the transcriptome of human monocytes and disease susceptibility. PLoS ONE *5*, e10693.
31. Lage, K., Hansen, N.T., Karlberg, E.O., Eklund, A.C., Roque, F.S., Donahoe, P.K., Szallasi, Z., Jensen, T.S., and Brunak, S. (2008). A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. Proc. Natl. Acad. Sci. USA *105*, 20870–20875.
32. Rossin, E.J., Lage, K., Raychaudhuri, S., Xavier, R.J., Tatar, D., Benita, Y., Cotsapas, C., and Daly, M.J.; International Inflammatory Bowel Disease Genetics Constortium. (2011). Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genet. *7*, e1001273.
33. De Jager, P.L., Jia, X., Wang, J., de Bakker, P.I., Ottoboni, L., Aggarwal, N.T., Piccio, L., Raychaudhuri, S., Tran, D., Aubin, C., et al.; International MS Genetics Consortium. (2009). Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. Nat. Genet. *41*, 776–782.
34. Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D., et al. (2007). Population genomics of human gene expression. Nat. Genet. *39*, 1217–1224.
35. Pritchard, J.K., Pickrell, J.K., and Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr. Biol. *20*, R208–R215.
36. Myles, S., Davison, D., Barrett, J., Stoneking, M., and Timpson, N. (2008). Worldwide population differentiation at disease-associated SNPs. BMC Med. Genomics *1*, 22.
37. Lynn, D.J., Winsor, G.L., Chan, C., Richard, N., Laird, M.R., Barsky, A., Gardy, J.L., Roche, F.M., Chan, T.H., Shah, N., et al. (2008). InnateDB: facilitating systems-level analyses of the mammalian innate immune response. Mol. Syst. Biol. *4*, 218.
38. Cotsapas, C., Voight, B.F., Rossin, E., Lage, K., Neale, B.M., Wallace, C., Abecasis, G.R., Barrett, J.C., Behrens, T., Cho, J., et al.; FOCiS Network of Consortia. (2011). Pervasive sharing of genetic effects in autoimmune disease. PLoS Genet. *7*, e1002254.

39. Patsopoulos, N.A., Esposito, F., Reischl, J., Lehr, S., Bauer, D., Heubach, J., Sandbrink, R., Pohl, C., Edan, G., Kappos, L., et al.; Bayer Pharma MS Genetics Working Group; Steering Committees of Studies Evaluating IFNβ-1b and a CCR1-Antagonist; ANZgene Consortium; GeneMSA; International Multiple Sclerosis Genetics Consortium. (2011). Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. Ann. Neurol. *70*, 897–912.
40. Lage, K., Karlberg, E.O., Størling, Z.M., Olason, P.I., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tümer, Z., Pociot, F., Tommerup, N., et al. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat. Biotechnol. *25*, 309–316.
41. Raj, T., Shulman, J.M., Keenan, B.T., Chibnik, L.B., Evans, D.A., Bennett, D.A., Stranger, B.E., and De Jager, P.L. (2012). Alzheimer disease susceptibility loci: evidence for a protein network under natural selection. Am. J. Hum. Genet. *90*, 720–726.
42. Stranger, B.E., Montgomery, S.B., Dimas, A.S., Parts, L., Stegle, O., Ingle, C.E., Sekowska, M., Smith, G.D., Evans, D., Gutierrez-Arcelus, M., et al. (2012). Patterns of cis regulatory variation in diverse human populations. PLoS Genet. *8*, e1002639.
43. Murphy, A., Chu, J.H., Xu, M., Carey, V.J., Lazarus, R., Liu, A., Szefler, S.J., Strunk, R., Demuth, K., Castro, M., et al. (2010). Mapping of numerous disease-associated expression polymorphisms in primary peripheral blood CD4+ lymphocytes. Hum. Mol. Genet. *19*, 4745–4757.
44. Gateva, V., Sandling, J.K., Hom, G., Taylor, K.E., Chung, S.A., Sun, X., Ortmann, W., Kosoy, R., Ferreira, R.C., Nordmark, G., et al. (2009). A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. Nat. Genet. *41*, 1228–1233.
45. Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R., and Willer, C.J. (2010). LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics *26*, 2336–2337.
46. Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I., and Dermitzakis, E.T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet. *6*, e1000895.
47. Korn, T., Bettelli, E., Oukka, M., and Kuchroo, V.K. (2009). IL-17 and Th17 cells. Annu. Rev. Immunol. *27*, 485–517.
48. Cooke, A. (2006). Th17 cells in inflammatory conditions. Rev. Diabet. Stud. *3*, 72–75.
49. Abraham, C., and Cho, J.H. (2009). IL-23 and autoimmunity: new insights into the pathogenesis of inflammatory bowel disease. Annu. Rev. Med. *60*, 97–110.
50. Stockinger, B., and Veldhoen, M. (2007). Differentiation and function of Th17 T cells. Curr. Opin. Immunol. *19*, 281–286.
51. Peters, A., Lee, Y., and Kuchroo, V.K. (2011). The many faces of Th17 cells. Curr. Opin. Immunol. *23*, 702–706.
52. Maddur, M.S., Miossec, P., Kaveri, S.V., and Bayry, J. (2012). Th17 cells: biology, pathogenesis of autoimmune and inflammatory diseases, and therapeutic strategies. Am. J. Pathol. *181*, 8–18.
53. Happel, K.I., Dubin, P.J., Zheng, M., Ghilardi, N., Lockhart, C., Quinton, L.J., Odden, A.R., Shellito, J.E., Bagby, G.J., Nelson, S., and Kolls, J.K. (2005). Divergent roles of IL-23 and IL-12 in host defense against *Klebsiella pneumoniae*. J. Exp. Med. *202*, 761–769.
54. Aujla, S.J., Chan, Y.R., Zheng, M., Fei, M., Askew, D.J., Pociask, D.A., Reinhart, T.A., McAllister, F., Edeal, J., Gaus, K., et al. (2008). IL-22 mediates mucosal host defense against Gram-negative bacterial pneumonia. Nat. Med. *14*, 275–281.
55. Zheng, Y., Valdez, P.A., Danilenko, D.M., Hu, Y., Sa, S.M., Gong, Q., Abbas, A.R., Modrusan, Z., Ghilardi, N., de Sauvage, F.J., and Ouyang, W. (2008). Interleukin-22 mediates early host defense against attaching and effacing bacterial pathogens. Nat. Med. *14*, 282–289.
56. Khader, S.A., and Gopal, R. (2010). IL-17 in protective immunity to intracellular pathogens. Virulence *1*, 423–427.
57. Derbise, A., Cerda Marin, A., Ave, P., Blisnick, T., Huerre, M., Carniel, E., and Demeure, C.E. (2012). An encapsulated *Yersinia pseudotuberculosis* is a highly efficient vaccine against pneumonic plague. PLoS Negl. Trop. Dis. *6*, e1528.
58. Zhernakova, A., Stahl, E.A., Trynka, G., Raychaudhuri, S., Festen, E.A., Franke, L., Westra, H.J., Fehrmann, R.S., Kurreeman, F.A., Thomson, B., et al. (2011). Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. PLoS Genet. *7*, e1002004.
59. Zhernakova, A., van Diemen, C.C., and Wijmenga, C. (2009). Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nat. Rev. Genet. *10*, 43–55.
60. Dixon, A.L., Liang, L., Moffatt, M.F., Chen, W., Heath, S., Wong, K.C., Taylor, J., Burnett, E., Gut, I., Farrall, M., et al. (2007). A genome-wide association study of global gene expression. Nat. Genet. *39*, 1202–1207.