Population Genetics of IFIH1: Ancient Population Structure, Local Selection, and Implications for Susceptibility to Type 1 Diabetes

Matteo Fumagalli,1,2 Rachele Cagliani,1 Stefania Riva,1 Uberto Pozzoli,1 Mara Biasin,3 Luca Piacentini,3 Giacomo P. Comi,4 Nereo Bresolin,1,4 Mario Clerici,5,6 and Manuela Sironi*,1
1Bioinformatic Lab, Scientific Institute Istituto di Ricovero e Cura a Carattere Scientifico E. Medea, Bosisio Parini, Lecco, Italy
2Bioengineering Department, Politecnico di Milano, Milan, Italy
3Chair of Immunology, Dipartimento di Scienze Cliniche, Laboratorio Interdisciplinare Tecnologie Avanzate Vialba, University of Milano, Milan, Italy
4Dino Ferrari Centre, Department of Neurological Sciences, University of Milan, IRCCS Ospedale Maggiore Policlinico, Mangiagalli and Regina Elena Foundation, Milan, Italy
5Chair of Immunology, Department of Biomedical Sciences and Technologies LITA Segrè, University of Milano, Milan, Italy
6Fondazione Don C. Gnocchi, Istituto di Ricovero e Cura a Carattere Scientifico, Milan, Italy
*Corresponding author: E-mail: manuela.sironi@bp.lnf.it.
Associate editor: Connie Mulligan

Abstract

The human interferon induced with helicase C domain 1 (IFIH1) gene encodes a sensor of double-strand RNA involved in innate immunity against viruses, indicating that this gene is a likely target of virus-driven selective pressure. Notably, IFIH1 also plays a role in autoimmunity, as common and rare polymorphisms in this gene have been associated with type 1 diabetes (T1D). We analyzed the evolutionary history of IFIH1 in human populations. Results herein suggest that two major IFIH1 haplotype clades originated from ancestral population structure (or balancing selection) in the African continent and that local selective pressures have acted on the gene. Specifically, directional selection in Europe and Asia resulted in the spread of a common IFIH1 haplotype carrying a derived His460 allele. This variant changes a highly conserved arginine residue in the helicase domain, possibly conferring altered specificity in viral recognition. An alternative common haplotype has swept to high frequency in South Americans as a result of recent positive selection. Previous studies suggested that a portion of risk alleles for autoimmune diseases could have been maintained in humans as they conferred a selective advantage against infections. This is not the case for IFIH1, as population genetic differentiation and haplotype analyses indicated that the T1D susceptibility alleles behaved as neutral or nearly neutral polymorphisms. Our findings suggest that variants in IFIH1 confer different susceptibility to diverse viral infections and provide insight into the relationship between adaptation to past infection and predisposition to autoimmunity in modern populations.

Key words: IFIH1, antiviral response, local selective pressure, positive selection, population structure, type 1 diabetes.

Introduction

The human interferon induced with helicase C domain 1 (IFIH1) gene (MIM 606951) encodes a cytoplasmic sensor of double-strand RNA (dsRNA) that mediates immune activation in response to viral infections (reviewed in Meylan et al. 2006). The protein product of IFIH1 interacts with dsRNA through its helicase domain and uses two N-terminal CARD domains for signaling through a set of adaptor molecules that converge on interferon responsive factors and nuclear factor kB for the production of β-interferon and other cytokines (reviewed in Meylan et al. 2006; Takeuchi and Akira 2008). Experiments in knockout mice have indicated that IFIH1 mainly acts as a sensor of picornavirus-derived dsRNA (Kato et al. 2006), and recent evidence has indicated that V proteins encoded by paramyxoviruses can bind the protein product of IFIH1 and inhibit dsRNA-mediated activation of the β-interferon promoter (Andrejeva et al. 2004).

Picornaviridae and Paramyxoviridae include several well-known human pathogens such as poliovirus, coxsackievirus, encephalomyocarditis virus, measles, and mumps. Consistently, we have previously demonstrated that a single nucleotide polymorphism (SNP) in IFIH1 displays signatures of virus-driven selective pressure in human populations (Fumagalli et al. 2010).

Recent studies have shown that polymorphisms in IFIH1 also play a relevant role in the pathogenesis of autoimmune diseases. Both common and rare polymorphic variants in the gene have been reproducibly associated with the susceptibility to type 1 diabetes (T1D) (Smyth et al. 2006; Liu et al. 2009; Nejentsev et al. 2009; Shigemoto et al. 2009; Jermendy et al. 2010). In particular, the derived T allele of rs1990760 (exon 15, Ala946Thr) is common in Europeans...
and correlates with an increased risk to develop the disease (Smyth et al. 2006; Liu et al. 2009; Nejentsev et al. 2009; Jermendy et al. 2010), whereas the association of this same variant with susceptibility to rheumatoid arthritis, autoimmune thyroid disease, and multiple sclerosis is more controversial (Marinou et al. 2007; Sutherland et al. 2007; Martinez, Santiago, et al. 2008; Martinez, Varade, et al. 2008; Couturier et al. 2009; Enevold et al. 2009; Penna-Martinez et al. 2009). In the case of T1D, rare variants that decrease or disable IFIH1 expression have a protective role (Nejentsev et al. 2009), whereas higher gene expression is observed in peripheral blood mononuclear cells of individuals carrying the common susceptible genotype (Liu et al. 2009).

On the one hand, these data suggest that increased efficiency of IFIH1 transcription or protein function may be associated with the development of autoimmunity. On the other hand, given the central role of IFIH1 in antiviral response, it is conceivable that sustained activity of this helicase might confer strong protection against infections and therefore be favored by natural selection.

Recent studies addressing the relationship between adaptation and immune diseases in humans (Fumagalli, Pozzoli, et al. 2009; Barreiro and Quintana-Murci 2010) suggested that a portion of risk alleles has been selected because they could provide protection against infectious diseases.

Here, we analyzed the evolutionary history of IFIH1 in humans. Results indicate that local selective pressures have acted on this gene, favoring the spread of different alleles/haplotypes in distinct geographic areas. In particular, population genetics analyses suggest that ancient population structure (or balancing selection) resulted in the maintenance of two major haplotype clades in African populations, whereas a selective sweep has driven a derived nonsynonymous variant to high frequency in Asians and Europeans; as for South Americans, a recent, possibly ongoing, selective sweep originated population-specific haplotypes. Finally, population genetic differentiation and haplotype analysis suggested that the T1D risk alleles in Europeans have not been the selection target but rather behaved as neutral variants.

Materials and Methods

DNA Samples and Sequencing/Genotyping

Human genomic DNA for Europeans (CEU), Yoruba (YRI), Asians (AS), and South Americans (SAM) was obtained from the Coriell Institute for Medical Research. The analyzed region was polymerase chain reaction (PCR) amplified in overlapping fragments and directly sequenced; primer sequences are available on request. PCR products were treated with ExoSAP-IT (USB Corporation Cleveland Ohio, Ohio, OH), directly sequenced on both strands with a Big Dye Terminator Sequencing Kit (v3.1 Applied Biosystems) and run on an Applied Biosystems ABI 3130 XL Genetic Analyzer (Applied Biosystems). Sequences were assembled using AutoAssembler version 1.4.0 (Applied Biosystems), inspected manually by two distinct operators, and singletons were reamplified and resequenced.

Data Retrieval and Haplotype Construction

Genotype data for 238 resequenced human genes were derived from the National Institute of Environmental Health Sciences (NIEHS) SNPs Program Web site (http://epgs.washington.edu). In particular, we selected genes that had been resequenced in populations of defined ethnicity including Europeans, Yoruba, and Asians (NIEHS panel 2).

For each gene, a 5-kb window was randomly selected; windows with resequencing gaps longer than 500 bp or containing ≤5 SNPs were discarded. The number of windows for YRI, CEU, and AS were 203, 193, and 186, respectively.

Haplotypes were inferred using PHASE version 2.1 (Stephens et al. 2001; Stephens and Scheet 2005), a program for reconstructing haplotypes from unrelated genotype data through a Bayesian statistical method. Haplotypes for individuals resequenced in this study are available as supplementary material (supplementary table S1, Supplementary Material online).

Data concerning the HGDP–CEPH panel derive from a previous work (Li et al. 2008). Atypical or duplicated samples and pairs of close relatives were removed (Rosenberg 2006). Following previous indications (Fumagalli, Cagliani, et al. 2009; Fumagalli, Pozzoli, et al. 2009), Bantu individuals (South Africa) were considered as one population.

Annotation regarding conserved sequences among different species was derived from UCSC Genome Browser (http://genome.ucsc.edu/, phastConsElements28wayPlac-Mammal table).

We calculated conserved noncoding sequence (CNS) densities (proportion of conserved bases per intron length) for a set of 20,978 nonalternative introns. Density distributions (percentiles) were calculated independently for ten intron length classes (breaks: 1; 128; 294; 549; 886; 1,300; 1,880; 2,700; 4,080; 8,020; and 53,600). CNS densities for IFIH1 introns 3 and 4 have been compared with the corresponding class distribution based on their length and correspond to the 97th and 99th percentiles, respectively.

Statistical Analysis

A detailed description of all tests we applied and their meaning are available in the supplementary material (supplementary table S2, Supplementary Material online).

Tajima’s D (Tajima 1989), Fu and Li’s D* and F* (Fu and Li 1993) statistics, as well as diversity parameters (Watterson 1975) and (Nei and Li 1979), and Fay and Wu’s H (Fay and Wu 2000) were calculated using “libsequence” (Thornton 2003), a C++ class library providing an object-oriented framework for the analysis of molecular population genetic data.

Calibrated coalescent simulations were performed using the cosi package (Schaffner et al. 2005) and its best-fit parameters for YRI, CEU, and AS populations with 10,000 iterations. For SAM, a previously reported demographic model (Ray et al. 2010) was used and included in the cosi best-fit model. Coalescent simulations were conditioned on mutation rate, and recombination rate was derived from UCSC
tables (http://genome.ucsc.edu/, snpRecombRateHamap table).

Composite likelihood ratio (CLR) test and coalescent simulations under a selective sweep regime were performed using clsw and ssw programs kindly provided by Yuseob Kim.

The \( F_{ST} \) statistic (Wright 1950) estimates genetic differentiation among populations, and it was calculated among continental groups using the R package HIERFSTAT (Goudet 2005).

\( S^* \), a measure of linkage disequilibrium (LD) based on the number of congruent or almost congruent mutations, was calculated as proposed by Plagnol and Wall (2006) using LDstruct software (http://www-gene.cimr.cam.ac.uk/vplagnol/). Significance was assessed by calibrated coalescent simulations conditioned on the number of segregating sites and incorporating different demographic models (Schaffner et al. 2005; Voight et al. 2005; Gutenkunst et al. 2009).

Data concerning haplotype-based tests iHS and XP-EHH were derived from the HGDP Selection Browser (http://hgdp.uchicago.edu/cgi-bin/gbrowse/HGDP/) (Pickrell et al. 2009).

The maximum likelihood ratio Hudson-Kreitman-Aguadé (MLHKA) test was performed using the MLHKA software (Wright and Charlesworth 2004) as previously proposed (Fumagalli et al. 2009). Briefly, 16 reference loci were randomly selected among NIEHS loci shorter than 20 kb that have been resequenced in the three populations; the only criterion was that Tajima’s D did not suggest the action of natural selection (i.e., Tajima’s D is higher than the 5th and lower than the 95th percentiles in the distribution of NIEHS genes). The reference set was accounted for by the following genes: VNN3 (MIM 606592), PL2G2D (MIM 606630), MB (MIM 160000), MAD2L2 (MIM 604094), HRAS (MIM 190020), CYP17A1 (MIM 609300), ATOX1 (MIM 602270), BNI3P (MIM 603293), CDC20 (MIM 603618), NGB (MIM 605304), TUBA1 (MIM 191110), MT3 (MIM 139255), NUDT1 (MIM 600312), PRDX5 (MIM 605683), RETN (MIM 605565), and JUND (MIM 165162).

The MWU high test was performed as previously described (Andres et al. 2009; Nielsen et al. 2009). Significance was assessed by performing 10,000 coalescent simulations conditioned on the number of segregating sites and incorporating demographic scenarios (Schaffner et al. 2005).

Median-joining networks to infer haplotype genealogy were constructed using NETWORK 4.5 (Bandelt et al. 1999). Estimate of the time to the most recent common ancestor (TMRCA) was obtained using a phylogeny-based approach implemented in NETWORK based on the average distance from the root (Morral et al. 1994; Saillard et al. 2000) and using a mutation rate based on the number of fixed differences between human and chimpanzee and assuming a separation time from humans of 6 My (Glazko and Nei 2003). A second TMRCA estimate derived from application of a maximum likelihood coalescent method implemented in GENETREE (Griffiths and Tavare 1994, 1995). Again, the mutation rate \( \mu \) was obtained on the basis of the divergence between human and chimpanzee and under the assumption of a generation time of 25 years. Using this \( \mu \) and \( \theta \) maximum likelihood (\( \theta_{ML} \)), we estimated the effective population size parameter (\( N_e \)). With these assumptions, the coalescence time, scaled in \( 2Ne \) units, was converted into years. For the coalescence process, \( 10^6 \) simulations were performed.

All calculations were performed in the R environment (www.r-project.org).

**Results**

Nucleotide Diversity and Haplotype Structure

We had previously identified an SNP in \( IFIH1 \) that strongly correlates with virus diversity in human populations, suggesting that this variant or a linked one has been subjected to virus-driven selective pressure (Fumagalli et al. 2010). The SNP (rs10439256) is located within intron 4 and falls within a region that is highly conserved among mammals. In general, the region surrounding exon 4 harbors several CNSs (fig. 1); a comparison with the intronic density of CNSs in human genes indicated that both introns 3 and 4 display an exceptionally high number of conserved sequences (see Materials and Methods), suggesting the presence of gene regulatory elements. As shown in figure 2, the allele frequency distribution for rs10439256 displays high continental differentiation: The ancestral T allele is fixed or almost fixed in Europe, the Middle East, and Asia, whereas the C-derived allele shows intermediate or high frequency in Africa and Central/South America, respectively. In line with this observation, analysis of population genetic differentiation (\( F_{ST} \)) across 52 human populations indicated that \( F_{ST} \) among SNPs genotyped in the HGDP–CEPH panel and located within \( IFIH1 \) (fig. 1). In particular, a striking difference in allele frequency is observed between populations living in East Asia and those in Central/South America (fig. 2), suggesting local adaptation. In order to explore this possibility, we calculated the distribution of \( F_{ST} \) values for all SNPs in the HGDP–CEPH panel (more than 660,000 variants, Li et al. 2008) between the Yakut (located in Siberia and considered as the closest ancestors of modern Americans Li et al. 2008) and the Maya. rs10439256 displayed an \( F_{ST} \) of 0.32, which corresponds to a percentile rank of 0.957 in the distribution of HGDP–CEPH SNPs, suggesting the action of local selective pressures on \( IFIH1 \) in the Americas.

To gain further insight into the evolutionary history of the gene, we resequenced a \( \sim 10 \)-kb region encompassing rs10439256 and covering several CNSs (fig. 1) in three HapMap populations, namely Yoruba (YRI), Europeans (CEU), and East Asians (AS), as well as in SAM. The number of SNPs identified in each population is reported in table 1, together with \( \theta_W \) and \( \pi \), two nucleotide diversity measures (Watterson 1975; Nei and Li 1979). As an empirical comparison, \( \theta_W \) and \( \pi \) were also calculated for 5-kb windows deriving from 238 genes resequenced by the NieHS program in YRI, CEU, and AS (no extensive resequencing data are available for
SAM) (see supplementary table S3, Supplementary Material online, for a comparison with 10-kb windows). As shown in table 1, the CEU sample displayed a significant reduction of nucleotide diversity in this gene region.

We next calculated $F_{ST}$ over the whole resequenced region. Generally, high $F_{ST}$ values were obtained (table 1) and again, these were compared with the distribution of $F_{ST}$ calculated for the 5-kb reference windows: YRI/CEU values ranked above the 95th percentile (see supplementary table S3, Supplementary Material online, for a comparison with 10-kb windows).

In order to analyze the haplotype structure of the resequenced IFIH1 gene region, we constructed a median-joining network (fig. 3). A genealogy with two major deeply separated clades (A and B) was evident. In line with the $F_{ST}$ results, haplotype frequency is extremely diverse across populations with all CEU chromosomes clustering in clade B together with the majority of Asian haplotypes.

**Fig. 1.** Schematic diagram of the exon–intron structure of IFIH1. Gray boxes represent exons, and the region we resequenced is indicated by the shaded box. The location and ID of SNPs genotyped in the HGDP–CEPH panel are reported, and the line is proportional to $F_{ST}$ calculated across continental groups. The location of CNSs is shown below the gene diagram (black boxes).

**Fig. 2.** Worldwide allele frequency distribution for rs10439256. Each pie represents one HGDP–CEPH population. The ancestral T allele is shown in blue and the derived C allele in yellow. Yakut and Maya are circled in red.
Conversely, most SAM chromosomes belong to clade A and account for population-specific haplogroups. In order to estimate the TMRCA of the IFIH1 haplotype genealogy, we applied a phylogeny-based method (Bandelt et al. 1999). Using a mutation rate based on 63 fixed differences between chimpanzees and humans and a separation time of 6 My (Glazko and Nei 2003), we estimated a TMRCA of 3.01 My [standard deviation (SD): 0.531 My]. In order to obtain a more robust estimate and given the relatively low recombination rate in the region, we calculated a second TMRCA using GENETREE, which is based on a maximum likelihood coalescent analysis (Griffiths and Tavare 1995). The method assumes an infinite-site model without recombination: eight segregating sites and one haplotype had to be removed as they violated these assumptions. The resulting gene tree, rooted using the chimpanzee

Table 1. Nucleotide Diversity and $F_{ST}$ for the IFIH1 Region We Analyzed.

<table>
<thead>
<tr>
<th>Population</th>
<th>N</th>
<th>$s^b$</th>
<th>$\theta_{w^c}$ (rankd)</th>
<th>$\pi^{c}$ (rankd)</th>
<th>$F_{ST}$ (rankd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>YRI</td>
<td>40</td>
<td>49</td>
<td>10.62 (0.69)</td>
<td>13.86 (0.93)</td>
<td>CEU: 0.41 (0.97), AS: 0.20 (0.72), and SAM: 0.52</td>
</tr>
<tr>
<td>CEU</td>
<td>40</td>
<td>6</td>
<td>1.30 (&lt;0.001)</td>
<td>0.74 (0.005)</td>
<td>AS: 0.16 (0.78) and SAM: 0.88</td>
</tr>
<tr>
<td>AS</td>
<td>40</td>
<td>31</td>
<td>6.72 (0.68)</td>
<td>6.93 (0.69)</td>
<td>SAM: 0.72</td>
</tr>
<tr>
<td>SAM</td>
<td>32</td>
<td>31</td>
<td>5.34</td>
<td>7.09</td>
<td></td>
</tr>
</tbody>
</table>

$^a$ Sample size (chromosomes).
$^b$ Number of segregating sites.
$^c$ Nucleotide diversity measures ($\times 10^{-4}$)
$^d$ Percentile rank relative to the distribution of 5-kb windows from NIEHS genes.

In order to estimate the TMRCA of the IFIH1 haplotype genealogy, we applied a phylogeny-based method (Bandelt et al. 1999). Using a mutation rate based on 63 fixed differences between chimpanzees and humans and a separation time of 6 My (Glazko and Nei 2003), we estimated a TMRCA of 3.01 My [standard deviation (SD): 0.531 My]. In order to obtain a more robust estimate and given the relatively low recombination rate in the region, we calculated a second TMRCA using GENETREE, which is based on a maximum likelihood coalescent analysis (Griffiths and Tavare 1995). The method assumes an infinite-site model without recombination: eight segregating sites and one haplotype had to be removed as they violated these assumptions. The resulting gene tree, rooted using the chimpanzee.

Fig. 3. Genealogy of IFIH1 haplotypes reconstructed through a median-joining network. Each node represents a different haplotype, with the size of the circle proportional to frequency. Nucleotide differences between haplotypes are indicated on the branches of the network. Circles are color coded according to population (green: YRI, red: CEU, blue: AS, and gray: SAM). The MRCA is also shown (black circle). The relative position of mutations along a branch is arbitrary. Mutation 75 identifies rs10930046 (His460Arg). The genealogy of clade B haplotypes with the inclusion of rs3747517 (mutation 76) and rs1990760 (mutation 77) is shown in the smaller panel. Color codes are as in the main network. The allelic status at aminoacid positions 843 (rs3747517) and 946 (rs1990760) is also shown.

Downloaded from https://academic.oup.com/mbe/article-abstract/27/11/2555/1119076 by guest on 03 April 2019
sequence, is partitioned into two deep branches (Fig. 4). A maximum likelihood estimate of $\theta (\theta_{ML})$ of 9.4 was obtained, resulting in an estimated effective population size ($N_e$) of 17,904. Using this method, the TMRCA of the IFIH1 haplotype lineages amounted to 2.31 My (SD: 0.537 My). These TMRCA estimates are deeper than those obtained for the great majority of neutrally evolving autosomal loci (Tishkoff and Verrelli 2003; Garrigan and Hammer 2006).

**Neutrality Tests**

We calculated Tajima’s $D$ ($D_T$) (Tajima 1989), Fu and Li’s $F^*$ and $D^*$ (Fu and Li 1993), as well as Fay and Wu’s $H$ (Fay and Wu 2000) for the resequenced IFIH1 region and evaluated whether these statistics significantly deviate from expectations under neutrality using both coalescent simulations and the empirical distribution of 5-kb reference windows (see supplementary table S3, Supplementary Material online, for a comparison with 10-kb windows). For coalescent simulations, we applied models that incorporate demographic scenarios (see Materials and Methods). Results are summarized in table 2 and indicate that $D_T$ is significantly high in YRI but no other test rejects neutrality in this population. Conversely, low values of $D_T$, $D^*$, and $F^*$ were obtained for CEU with borderline significant $P$ values. As for AS and SAM, $D^*$ was unusually high in both populations and $H$ was significantly negative; this latter result indicates that AS and SAM display an excess of high-frequency derived alleles.

Under neutral evolution, the amount of within-species diversity is predicted to correlate with levels of between-species divergence because both depend on the neutral mutation rate (Kimura 1983). To test this expectation, we applied a MLHKA test (Wright and Charlesworth 2004) by comparing polymorphisms and divergence levels at the IFIH1 genomic region with 16 NIEHS genes resequenced in YRI, CEU, and AS (see Materials and Methods). The results are shown in table 2 and indicate that a significant reduction in nucleotide diversity versus divergence is detectable in the CEU sample, whereas the opposite situation is observed in YRI and AS.

**Selection Pattern in Human Populations and Possible Selection Targets**

As determined above, IFIH1 haplotypes display an unusually deep coalescent time. There are at least two possible explanations for this finding: long-standing balancing selection and ancient population structure in the African continent. The two processes originate different gene tree

---

**Fig. 4.** Estimated haplotype tree for the IFIH1 gene region we resequenced. Mutations are represented as black dots and named for their physical position along the regions. The absolute frequency of each haplotype is also reported.
topologies in that balancing selection elongates the whole neutral genealogy, whereas population structure results in a longer proportion of genealogical time occupied by single lineages (i.e., longer than expected basal branches) (Takahata 1990; Wall 2000). Therefore, admixture of structured populations is expected to result in a specific pattern of LD. Plagnol and Wall (2006) recently proposed a new measure of LD, denoted $S^*$, that is specifically devised to test for an excess of congruent or almost congruent mutations. Calculation of $S^*$ for the YRI sample resulted in a value of 69,014; coalescent simulations with different demographic models yielded the following $P$ values: 0.038 (Schaffner et al. 2005), 0.034 (Voight et al. 2005), and 0.017 (Gutenkunst et al. 2009) (see Materials and Methods). These data are therefore consistent with the idea that the two major clades of the $IFIH1$ genealogy result from ancient population admixture in Africa. Yet, it is worth mentioning that a more complex situation of long-standing balancing selection with episodic selective sweeps or with multiple positively selected alleles (see below) might also result in unusual LD patterns.

Whatever the reason for the maintenance of the two deep clades, the large number of mutations on the basal branches (i.e., mutations that differentiate clades A and B) affects neutrality tests, irrespective of the evolutionary patterns that have acted on this gene region in different populations following the establishment of balancing selection or after population admixture. Indeed, these mutations contribute to nucleotide diversity estimates and are likely to account for the significant MLHKA test we obtained for YRI and AS and for the significantly high Fu and Li’s $D^*$ values in SAM and AS. For the CEU sample, the lack of haplotypes in clade A simplifies the analysis and the data we obtained (reduced nucleotide diversity and significant MLHKA test) suggest that $IFIH1$ has undergone a selective sweep in this population.

Further analyses on the evolutionary pattern of $IFIH1$ in CEU, AS, and SAM were performed by applying a CLR test (Kim and Stephan 2002) that evaluates the local reduction of variation and skew of the frequency spectrum. Specifically, all CEU chromosomes were included in the analysis, whereas for AS and SAM only haplotypes in clades B and A, respectively, were used. We evaluated statistical significance of likelihood ratio values in two ways: distinguishing ancestral from derived alleles and then not distinguishing allele states (Test 1 and Test 2). As shown in table 3, all tests rejected neutral evolution for all populations. Given that CLR is not robust to demographic history, we applied a goodness-of-fit (GOF) test (Jensen et al. 2005); this method specifically tests how well a selective sweep model fits the data, as opposed to a generalized alternative model, by simulating genealogies under directional selection. Thus, nonsignificant $P$ values represent a good fit of the sweep model to the data, whereas low $P$ values fall within the range of effects that can also be generated by demographic events (e.g., population bottlenecks). For CEU, AS, and SAM, the GOF $P$ values suggest that rejection of neutrality by the CLR test is more likely due to a selective sweep than to demographic history (table 3). These data therefore support the idea that different $IFIH1$ haplotypes have increased in frequency in CEU/AS and SAM as a result of directional selection.

It has recently been shown that biased gene conversion (BGC) affects neutral substitution patterns (reviewed in Duret and Galtier 2009); the effect of BGC is particularly strong in subtelomeric regions and in regions with high male-specific recombination rates (Webster et al. 2005; Dreszer et al. 2007; Duret and Arndt 2008). $IFIH1$ is not subtelomeric and male-specific recombination rate along the gene amounts to 0.4 cm/Mb, a value substantially lower than the genome average for autosomes (0.98 cm/Mb) (Kong et al. 2002). In the region we analyzed, 17 polymorphisms (those located on the branch leading to clade B haplotypes) display a high frequency of the derived allele (frequency >0.6 averaged over all populations); of these only seven are A/T $\rightarrow$ G/C mutations, one is a T $\rightarrow$ A and the remaining are G/C $\rightarrow$ T/A substitutions. Overall, these data suggest that BGC does not play a major role in shaping nucleotide variability at $IFIH1$.

The rise in frequency of a selected allele may generate an extended haplotype that results from the long-range association with nearby polymorphisms—depending on the

### Table 2. Summary Statistics and MLHKA Test for the $IFIH1$ Gene Region.

<table>
<thead>
<tr>
<th>Population</th>
<th>Tajima’s $D$</th>
<th>Fu and Li’s $D^*$</th>
<th>Fu and Li’s $F^*$</th>
<th>Fay and Wu’s $H$</th>
<th>MLHKA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Value</td>
<td>Rank $^a$</td>
<td>$p^b$</td>
<td>Value</td>
<td>Rank $^a$</td>
</tr>
<tr>
<td>YRI</td>
<td>1.16</td>
<td>0.975</td>
<td>0.014</td>
<td>$-0.73$</td>
<td>0.350</td>
</tr>
<tr>
<td>CEU</td>
<td>$-1.14$</td>
<td>0.145</td>
<td>0.104</td>
<td>$-2.10$</td>
<td>0.067</td>
</tr>
<tr>
<td>AS</td>
<td>0.11</td>
<td>0.548</td>
<td>0.498</td>
<td>1.32</td>
<td>0.962</td>
</tr>
<tr>
<td>SAM</td>
<td>$-0.91$</td>
<td>NA</td>
<td>0.148</td>
<td>1.53</td>
<td>NA</td>
</tr>
</tbody>
</table>

*NOTE.—NA, not available.

$^a$ Percentile rank relative to the distribution of 5-kb windows from NEHS genes.

$^b$ $P$ value obtained by coalescent simulations using demographic models.

$^c$ Selection parameter ($k > 1$ indicates an excess of polymorphism compared with divergence and $k < 1$ indicates the opposite situation).

### Table 3. CLR Test Results.

<table>
<thead>
<tr>
<th>Population</th>
<th>Test 1 $^a$</th>
<th>Test 2 $^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LR</td>
<td>GOF</td>
</tr>
<tr>
<td>CEU</td>
<td>6.95 (0.004)</td>
<td>10.19 (0.25)</td>
</tr>
<tr>
<td>AS</td>
<td>6.01 (0.007)</td>
<td>15.61 (0.81)</td>
</tr>
<tr>
<td>SAM</td>
<td>9.90 (0.006)</td>
<td>9.18 (0.87)</td>
</tr>
</tbody>
</table>

*NOTE.—LR, likelihood ratio. $P$ values are given in parentheses.

$^a$ Distinguishing ancestral/derived allele.

$^b$ Not distinguishing ancestral/derived allele.
timing and strength of the selective event. To verify whether this is the case for IFIH1, we derived empirical P values for two haplotype-based tests, namely iHS (Voight et al. 2006) and XP-EHH (Sabeti et al. 2007) from the HGDP Selection Browser (http://hgdp.uchicago.edu/cgi-bin/gbrowse/HGDP/) (Pickrell et al. 2009). Although no exceptional value was observed for CEU and AS, the two tests were significant for HGDP–CEPH populations living in America. Specifically, the iHS test was significant over a large genomic portion encompassing IFIH1 and several flanking genes, whereas the XP-EHH test showed significant P values over a narrower region encompassing IFIH1 and partially extending into the FAP gene, with a peak being observed within IFIH1 (fig. 5).

As mentioned above, two nonsynonymous variants in IFIH1, Arg843His (rs3747517) and Ala946Thr (rs1990760) located in exons 13 and 15 (fig. 1), have been associated with T1D and MS (Smyth et al. 2006; Enevold et al. 2009; Liu et al. 2009; Nejentsev et al. 2009; Shigemoto et al. 2009; Jermendy et al. 2010). For YRI, CEU, and AS, we retrieved genotype information for the two variants from HapMap, whereas the two SNPs were typed in SAM. Additionally, we resequenced exons 3–8 in the four human populations. We noticed that one variant located in exon 7 (rs10930046) is responsible for the His460Arg substitution. The derived histidine allele is fixed in Europeans and displays extremely high frequency in Asians; conversely, SAM mainly carry the ancestral Arg allele that is present at intermediate frequency in YRI. The His460Arg variant is located in the helicase domain and the ancestral arginine residue is completely conserved across vertebrates, including birds and fishes (supplementary fig. S1, Supplementary Material online). Inclusion of this variant in the haplotype network indicated that the derived His460 allele is located on the branch leading to the major haplogroups in clade B, with all chromosomes in clade A carrying the ancestral allele (variant 75 in fig. 3). This is consistent with the possibility that this variant represents the selection target in CEU and AS. We next performed a second network analysis by including the two nonsynonymous variants in exons 13 and 15 (Arg843His and Ala946Thr). As shown in figure 3, all haplotypes in clade A carry Arg843 and Ala946 alleles with the exception of one YRI chromosome (carrying His843, possibly due to a recurrent mutation, not shown). Conversely, the situation for chromosomes in clade B is represented in figure 3 (small insert) and indicates that CEU and AS haplotypes are split into two main subgroups depending on the allelic status at codon positions 843 and 946. Notably, variants 31 and 32 in the network (fig. 3), which are specific to South American chromosomes, are located 3 bp apart from each other within a CNS (supplementary fig. S2, Supplementary Material online).

**Discussion**

Results herein demonstrate that distinct variants/haplotypes in IFIH1 gene have been subjected to natural
selection in human populations and suggest that two distantly related haplotype clades have originated from distinct ancestral hominid populations or have been maintained by balancing selection. Several authors have previously proposed that ancestral African populations were structured (reviewed in Garrigan and Hammer 2006) and signatures of ancient population admixture have been identified at specific human loci (Satta and Takahata 2004; Barreiro et al. 2005; Garrigan et al. 2005; Cox et al. 2008; Kim and Satta 2008). A recent study found strong evidence of ancient admixture in both Europeans and West Africans, with contributions to the modern gene pool of about 5% (Plagnol and Wall 2006). A hallmark of ancient population structure is the presence of highly differentiated haplotypes with very little evidence of recombination between lineages (Wall 2000). This is the result of mutations arising in isolated populations and prevented from recombining with one another until after the admixture event. Ancient population structure is also expected to result in unusually deep TMRCA estimates (reviewed in Garrigan and Hammer 2006). Our analysis of IFIH1 in the YRI sample indicated that the pattern of LD is consistent with the ancestral African population being subdivided. As it is evident from the median-joining network, one single chromosome (the isolated African haplotype in clade B) might result from recombination between the two lineages and $S^*$, a measure of LD based on the number of congruent (or almost congruent mutations), yielded a significantly high score, supporting the notion whereby the two major clades result from admixture of isolated populations. Yet, selective events or complex evolutionary scenarios may also determine unusual LD patterns and long-standing balancing selection results in deep coalescence through the maintenance of distinct lineages (Charlesworth 2006). Recently, the MWUhigh test has been applied at the genome-wide level to identify targets of balancing selection (Andres et al. 2009; Nielsen et al. 2009). The test is based on the concept whereby balancing selection skews the allele frequency spectrum toward intermediate frequency alleles, a finding that is not expected in a situation of ancestral admixture. The MWUhigh test calculated on the folded spectrum yielded a significant result for YRI ($P = 0.021$; see Materials and Methods) but not for the other populations (not shown), suggesting a role for balancing selection in maintaining the two major haplotype clades. Therefore, although unusual LD patterns (as measured by $S^*$) are not expected in a situation of balancing selection, an excess of intermediate frequency alleles (as obtained by the MWUhigh test) is not consistent with a scenario of ancient population structure. Nonetheless, complex evolutionary scenarios following admixture or balancing selection may affect both allele frequency and LD patterns. Also, these two possibilities are not necessarily mutually exclusive as a haplotype introduced by admixture may have higher chances to be detected in modern populations when subjected to a selective event (that opposes its chances of being lost by drift) (Hawks et al. 2008).

Several evidences suggest that different alleles in IFIH1 have been the target of directional selection in CEU, AS, and SAM. Specifically, our data suggest that clade B haplotypes have reached a high frequency in Europe and Asia as a result of directional selection. In the CEU sample, a significant reduction of nucleotide diversity is observed and the CLR test indicated that a sweep model fits the data for CEU, SAM, and AS. In this latter population, a small number of chromosomes is also observed in clade A; given the deep divergence of the two clades, this results in a high level of polymorphism (as assessed by the MLHKA test) and in a significant value for $D^s$, a situation that is not generally consistent with a simple model of directional selection. Yet, both tests are not devised to incorporate population structure (or preexisting balancing selection) in $P$ value calculation but rather rely on the null hypothesis of neutral evolution and perform comparison with other loci (MLHKA) or exploit coalescent simulation in a panmictic population ($D^s$). Similarly, complex evolutionary scenarios, such as the succession of balancing and directional selection regimes, might result in genetic diversity patterns that are difficult to reconcile with simple expectations. These same observations hold for the SAM sample, but in this case, directional selection resulted in increased frequency of population-specific haplotypes. Consistently, we observed high $F_{ST}$ between Asian and American populations, strongly suggesting that local selective pressures resulted in the spread of different alleles. The significant XP-EHH test is in line with this observation and suggests a relatively recent selective sweep event in SAM.

IFIH1 plays a central role in antiviral response and rs10439256 was previously indicated as being targeted by virus-driven selective pressure (Fumagalli et al. 2010). In the populations we analyzed, this variant is in strong LD with the His460Arg SNP (supplementary table S4, Supplementary Material online), which affects a highly conserved residue located in the helicase domain. Notably, the helicase domain is directly involved in viral RNA binding (Takeuchi and Akira 2008), suggesting that variants in this region alter the specificity of IFIH1 against one or more viral species. Therefore, it is tempting to speculate that the derived 460His allele has been the target of positive selection in European and Asian populations, possibly because it confers increased resistance to one or more viruses in these geographic areas. This would also imply that the Arg460His polymorphism represents a susceptibility/protective variant against diverse viral infections. With respect to South American populations, we noticed that two variants (rs12474958 and rs12478730, positions 31 and 32 in fig. 3) define all SAM chromosomes in clade A and occur within a CNS. Whether these variants affect IFIH1 regulation remains to be evaluated as well as the possibility that they represent the selection target in this population. In this respect, it is worth mentioning that selective sweeps typically affect large genomic regions due to genetic hitchhiking; thus, inference on the real selection target is difficult unless functional information is available.
The structure and distribution of IFIH1 haplotypes described herein may harbor consequences for association studies, especially when populations of non-European descent are being analyzed. The derived allele of rs1990760 was shown to predispose to T1D in Caucasians (Smyth et al. 2006; Liu et al. 2009; Nejentsev et al. 2009; Jermendy et al. 2010). As shown in figure 3, all CEU chromosomes belong to clade B, and both the ancestral and risk alleles of rs1990760 occur on the same haplotype background in this population. This is not the case for YRI and AS given that chromosomes harboring the ancestral allele are represented both in clades A and B. If, as expected, haplotypes in the two clades are functionally different, association studies in non-European populations using rs1990760 might be hindered by haplotype heterogeneity for the protective allele.

Given that T1D is characterized by a juvenile onset and is a potentially lethal disease, it is unclear why alleles that predispose to this condition have not been eliminated by natural selection.

It was recently suggested that a portion of risk alleles for autoimmune diseases, including T1D, have conferred a selective advantage against ancestral infections and may therefore have been maintained in human populations as a result of balancing or positive selection (Fumagalli, Pozzoli, et al. 2009; Barreiro and Quintana-Murci 2010). Data herein suggest that this is not the case for genetic variants in IFIH1 conferring susceptibility to T1D. Specifically, our results do not support the possibility that the Ala946Thr polymorphism associated with T1D represents the selected variant, whereas the association of the putative selection target (His460Arg) with T1D has been excluded (Nejentsev et al. 2009). Rather, \( F_{ST} \) and haplotype analyses suggest that the predisposing 946Thr and 843His alleles arose on clade B haplotypes and behaved as neutral or almost neutral variants. As a selective sweep favoring the raise of a selected allele may result in the parallel increase of linked neutral and mildly deleterious alleles, the T1D risk SNPs may have hitchhiked with the selected variant(s) in IFIH1 so as to increase in frequency in non-African populations. This observation supports the idea that a portion of susceptibility alleles for autoimmune conditions segregated as neutral or mildly deleterious variants in human populations for a long time because environmental conditions in preindustrialized societies did not allow the development of autoimmune diseases (Sironi and Clerici 2010).

**Supplementary Material**

Supplementary tables S1–S4 and figures S1–S2 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

M.C. is supported by grants from Istituto Superiore di Sanita’ ’Programma Nazionale di Ricerca sull’ AIDS,” the EM-PRO and AVIP EC WP6 Projects, the nGIn EC WP7 Project, the Japan Health Science Foundation, 2008 Ricerca Finalizzata (Italian Ministry of Health), 2008 Ricerca Corrente (Italian Ministry of Health), Progetto FIRB RETI: Rete Italiana Chimica Farmaceutica CHEM-PROFARMA-NET (RBP05NWWC), and Fondazione CARIPLO. M.S. is a member of the Doctorate School in Molecular Medicine, University of Milan.

**References**


