Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes

Matthieu Deschamps,[1,2,3] Guillaume Laval,[1,2] Maud Fagny,[1,2,3] Yuval Itan,[4] Laurent Abel,[4,5,6] Jean-Laurent Casanova,[4,5,6,7,8] Etienne Patin,[1,2] and Lluis Quintana-Murci[1,2,*]

Human genes governing innate immunity provide a valuable tool for the study of the selective pressure imposed by microorganisms on host genomes. A comprehensive, genome-wide study of how selective constraints and adaptations have driven the evolution of innate immunity genes is missing. Using full-genome sequence variation from the 1000 Genomes Project, we first show that innate immunity genes have globally evolved under stronger purifying selection than the remainder of protein-coding genes. We identify a gene set under the strongest selective constraints, mutations in which are likely to predispose individuals to life-threatening disease, as illustrated by *STAT1* and *TRAF3*. We then evaluate the occurrence of local adaptation and detect 57 high-scoring signals of positive selection at innate immunity genes, variation in which has been associated with susceptibility to common infectious or autoimmune diseases. Furthermore, we show that most adaptations targeting coding variation have occurred in the last 6,000–13,000 years, the period at which populations shifted from hunting and gathering to farming. Finally, we show that innate immunity genes present higher Neandertal introgression than the remainder of the coding genome. Notably, among the genes presenting the highest Neandertal ancestry, we find the *TLR6-TLR1-TLR10* cluster, which also contains functional adaptive variation in Europeans. This study identifies highly constrained genes that fulfill essential, non-redundant functions in host survival and reveals others that are more permissive to change—containing variation acquired from archaic hominins or adaptive variants in specific populations—improving our understanding of the relative biological importance of innate immunity pathways in natural conditions.

Introduction

The burden of infectious diseases has been massive throughout human history, particularly before the advent of hygiene, vaccines, antiseptics, and antibiotics, when human populations were ravaged by illnesses that resulted in high childhood mortality and short life expectancy.[1] In light of this, and given that the human genetic makeup strongly influences an individual's susceptibility to infectious disease and the resulting clinical outcome,[2,3] natural selection imposed by pathogens is expected to have profoundly affected the patterns of variability of the human genome.[4–7] Indeed, interspecies analyses and within-species studies in humans have established that purifying and positive selection have been pervasive among both genes and functions related to immunity and host defense.[5,8–14] Furthermore, pathogen pressure is increasingly recognized as the underlying cause of such selection signatures, with many immunity-related genes presenting patterns of variation that strongly correlate with pathogen diversity.[15]

Over recent decades, the dissection of the form and intensity of selection in the human genome has established the value of population genetics as a complement to clinical and epidemiological genetic studies, in delineating the biological relevance of immunity genes in natura and in predicting their involvement in disease.[2,4,7,16] Genes evolving under strong purifying selection are predicted to include those involved in essential mechanisms of host defense, variation in which should lead to severe disorders.[16] This prediction is supported by genome-wide data, because Mendelian disease genes are enriched in signals of purifying selection.[8,9,17] Conversely, genes evolving adaptively—through positive or balancing selection (e.g., *HBB* [MIM: 141900], *DARC* [MIM: 613665], *FUT2* [MIM: 182100], the *HLA* locus genes, ABO blood group genes, and *TRIM5* [MIM: 608487])—are usually more permissive to functional variation, which can exert a protective effect against infections.[2,4,7,18] These signals of adaptive evolution in immune-related genes, tending to be recent and population specific, further emphasize the important role of pathogens in local adaptation.

Besides the occurrence of novel mutations, functional variants transmitted through admixture represent another potential source of adaptive variation. Recent data provided evidence that 1%–6% of modern Eurasian genomes were inherited from ancient hominins, such as Neandertals or Denisovans,[19–21] with specific genomic regions presenting up to 64% of Neandertal ancestry.[22] In the context of immunity, there is increasing evidence to suggest that modern humans have acquired advantageous variation through admixture with ancient hominins, as

[1]Unit of Human Evolutionary Genetics, Institut Pasteur, 75015 Paris, France; [2]CNRS URA3012, 75015 Paris, France; [3]Université Pierre et Marie Curie, Cellule Pasteur UPMC, 75015 Paris, France; [4]St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, The Rockefeller University, New York, NY 10065, USA; [5]Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM U.1163, 75015 Paris, France; [6]Imagine Institute, Paris Descartes University, 75015 Paris, France; [7]Howard Hughes Medical Institute, New York, NY 10065, USA; [8]Pediatric Hematology-Immunology Unit, Necker Hospital for Sick Children, 75015 Paris, France
*Correspondence: quintana@pasteur.fr
http://dx.doi.org/10.1016/j.ajhg.2015.11.014.

documented by candidate gene approaches for *HLA* class I genes, *STAT2* (MIM: 600556), or the *OAS* gene cluster (MIM: 164350, 603350).[23–25]

Among the two arms that form the immune system, innate immunity constitutes the front line of host defense and provides a valuable model for the study of the selective pressure imposed by microorganisms—pathogenic and symbiotic—on host genomes.[2,26] Innate immunity relies on receptors that sense conserved microbial patterns or molecules and activate signaling pathways that involve the coordinated action of an array of downstream molecules, including adaptors, regulators, transcription factors, and effector molecules, all of which are required for the eradication of pathogens and to maintain homeostasis.[27] Importantly, unlike adaptive immunity whose parameter variation is mostly somatic and presents limited heritability, variation of innate immunity is germline encoded and thus needs to be best-adapted to ensure host survival.[28,29] Population genetic studies have shown that the impact of selection on some families of innate immune receptors and downstream signaling molecules (e.g., Toll-like receptors, interferons, or antimicrobial peptides) varies considerably,[14,30–43] helping to delineate the relative functional importance of different immune pathways.[2,26] However, these studies have focused on specific candidate genes or gene families. A comprehensive, genome-wide view of how selection has driven the evolution of innate immunity in humans is thus missing.

Here, we took advantage of population whole-genome sequence data to increase our understanding of the degree of essentiality and adaptability of the different genes governing innate immunity and thus, to provide novel insights into their respective biological relevance in host survival. To do so, we first created a hand-curated list of more than 1,500 genes belonging to the different modules constituting the innate immune system in humans (Material and Methods). We then analyzed their patterns of population genetic variability, which we compared to the remainder of the genome, using the 1000 Genomes Project dataset,[44] allowing us to evaluate the occurrence and intensity of constraint and adaptation to geographic and environmental pressures with an unprecedented level of resolution. Finally, we estimated the time range at which the bulk of genetic adaptation involving innate immunity has occurred and evaluated the extent to which human populations have acquired innate immunity genetic variation through admixture with Neandertals.

Material and Methods

Hand-Curated List of Innate Immunity Genes

We created a hand-curated list of innate immunity genes (IIGs) by combining two public databases, Gene Ontology (GO)[45] and InnateDB,[46] as well as by incorporating missing entries. Specifically, we used the GO term "innate immune response" (GO: 0045087) to extract 1,309 entries corresponding to 884 unique annotations (last access January 2015). We removed all non-human taxon entries, non-SwissProt reviewed proteins, entries without gene symbol or not approved by the HUGO Gene Nomenclature Committee, as well as those encoding for HLA proteins and immunoglobulins. This yielded a final set of 806 GO genes. For InnateDB, we retrieved 2,158 entries, corresponding to 989 unique annotations (last access January 2015). Similarly to GO, we removed entries without approved HUGO names, *HLA* genes, and miRNAs, and obtained a final set of 905 InnateDB genes.

When manually reviewing these two gene lists, we noticed the presence of proteins belonging, based on structural homology, to gene families that are commonly accepted to play a role in innate immune processes (e.g., Nod-like receptors), even if the involvement of some of their individual members in innate immunity remains unclear. Because GO and InnateDB did not systematically use this "family-based criteria," we manually did so for gene families in which some of their individual members were missing (e.g., we added 28 TRIM proteins and 24 C-type lectins). In addition, we noticed the absence of several well-described or recently identified molecules, including some nucleic acid sensors such as *ABCF1* (MIM: 603429), *DHX15* (MIM: 603403), *DHX33* (MIM: 614405), and *PYHIN1* (MIM: 612677). By applying these inclusion criteria, we also retrieved some of the filtered GO and InnateDB entries that were initially removed because they were present under a non-approved HUGO symbol. This was the case for the interferons *IFNL1* (MIM: 607403), *IFNL2* (MIM: 607401), and *IFNL3* (MIM: 607402), which were annotated in InnateDB as *IL29*, *IL28A*, and *IL28B*, respectively. We acknowledge that some molecules that we manually added (which were absent from the lists that were downloaded from the databases at the time of the study) have now been included in the corresponding websites. Our manual inclusion of additional genes, based on current knowledge of gene families and functions related to innate immunity (e.g., missing chemokines, defensins, and caspases; see Table S1), was an attempt to update existing databases. Overall, we manually added a set of 187 genes, making a final dataset of 1,553 genes that constituted the basis of all subsequent analyses (Table S1).

We classified the 1,553 genes according to their main known (or inferred) function into nine different categories. These include sensors (n = 274), adaptors (n = 46), signal transducers (n = 245), transcription factors (n = 93), effector molecules (n = 284), and secondary receptors (n = 70). We also included regulators of the signaling pathways (n = 310) and accessory molecules (n = 68) necessary for an efficient immune response. This classification was based on the functional information available for each of these genes in InnateDB, UniProt, and/or the corresponding publications. Out of the 1,553 genes, 163 remained unclassified, because their reported molecular description did not allow us to include them in any of the categories above and were thus grouped into a final category termed as "uncharacterized."

Whole-Genome Sequence Datasets

Depending on the nature of the analyses performed, we used the high-coverage (~57×) exome sequencing data and/or the low-coverage (2–6×) sequencing data of the 1000 Genomes Project, which are available for 1,092 individuals from 14 populations from Europe, East Asia, sub-Saharan Africa, and the Americas.[44]

Assessing the Action of Purifying Selection

Quantification of the Extent of Purifying Selection

To estimate the strength of purifying selection, we used SnIPRE,[47] which relies on the comparison of polymorphism and divergence

at synonymous and non-synonymous sites (i.e., McDonald-Kreitman contingency table). This method uses a generalized linear mixed model to model the genome-wide variability among categories of mutations and estimates two population genetics parameters for each gene: γ, the population selection coefficient, and f, the proportion of non-synonymous mutations that are not deleterious. We focused our analyses on f, to quantify the strength of purifying selection: a low f value indicates that a large proportion of non-synonymous alleles were deleterious and have been removed from the population. We retrieved the alignment of the human genome (hg19 release) and the chimpanzee genome (PanTro3 release), used as an outgroup, provided by the UCSC Genome Browser, corresponding to ~2.5 Gb of aligned sequences. All regions of the human genome that are deleted or have no homology with the chimpanzee were excluded from the analysis.

We identified 33.5 million single bases that were different between the two species, which were then functionally annotated with SnpEff,[48] using the GRCh37.65 build. We obtained 200,676 non-synonymous or synonymous divergent differences between humans and chimpanzees. We next retrieved all human variants that have been identified by the 1000 Genomes Project high-coverage exome dataset. We kept 445,401 variants that were annotated as non-synonymous or synonymous, were outside of gaps in the human-chimpanzee alignment, and were polymorphic in at least one human population. Variants with a fixed alternate allele in the 1000 Genomes Project dataset (i.e., reference allele is absent from the sample) were added to fixed differences between human and chimpanzee. We excluded from human-chimpanzee fixed differences 16,345 positions that were actually polymorphic in humans or chimpanzees, using the dbSNP136 chimpanzee database. We retrieved all human CDS with length >68 bp and considered the longest transcript available for each gene. We deduced from the genetic code the number of synonymous and non-synonymous sites in the 22,571 transcripts obtained, accounting for gaps in the human-chimpanzee alignments. Finally, we excluded all transcripts that had a length <50 bp after accounting for these gaps, had no divergent nor polymorphic mutations, had no HUGO-approved gene symbol, or was not a SwissProt "reviewed" protein. HGNC-approved gene symbols were retrieved with the R BioConductor biomaRt package (v.2.22.0), and SwissProt protein status were retrieved with the R BioConductor UniProt.ws package (v.2.6.2). SnIPRE[47] was then used to estimate the f parameter for 17,967 genes, which included a final set of 1,492 IIGs, assuming human and chimpanzee sample sizes of 1,092 and 10, respectively.

Statistical Analyses

We estimated the enrichments of IIGs among genes evolving under purifying selection by measuring the odds ratios (OR) of purifying selection. This OR measures the relative proportion of IIGs among genes with purifying selection signals and is defined as follows:

$$OR = \left[\frac{P(IIG \mid SEL)}{P(\overline{IIG} \mid SEL)}\right]\left[\frac{P(\overline{IIG} \mid \overline{SEL})}{P(IIG \mid \overline{SEL})}\right],$$

with IIG and $\overline{IIG}$ denoting genes being or not innate immunity genes, respectively, SEL and $\overline{SEL}$ being "with" and "without purifying selection signals," respectively. If purifying selection preferentially targets IIGs, we expect proportionally more IIGs in the tail of the f distribution (OR > 1). Otherwise, we expect the same amount of IIGs in the tail of the f distribution as in the remainder of the genome (OR $\approx$ 1). Note that all statistical tests comparing f distributions among gene classes (e.g., IIGs against the rest of human genes) were performed (1) assuming statistical independence between genes (i.e., very weak correlation between f values of neighboring genes, $R^2 = 0.016384$) and (2) correcting by any potential differences in the distributions of gene length and number of SNPs per gene observed among classes. Because the f value is potentially dependent on gene length and/or the number of SNPs per gene, we corrected for these potential confounders by performing 10^5 resamplings where these distributions (gene length and number of SNPs per gene) were matched between the tested class (e.g., IIGs) and every resampled set of genes.

Prediction of the Functional Impact of Mutations

To evaluate the fitness status of variants at IIGs, we used the Combined Annotation Dependent Depletion (CADD) algorithm.[49] We downloaded the PHRED-scaled C-score calculation for the 39,701,210 variants (SNPs and indels) from the 1000 Genomes Project and filtered out mutations that were excluded from the analyses of purifying selection. We then compared the number of SNPs in IIGs (33,867) and in the remainder of protein-coding sequences (399,784) having a PHRED-scaled score $\geq$ 15. We considered this value as the limit above which mutations are probably damaging, because this score corresponds to the median value for all possible canonical splice site changes and non-synonymous variants.[49]

Selective Constraints on Genes Associated with Primary Immunodeficiencies

To assess to degree of purifying selection on IIGs associated with primary immunodeficiencies (PID), we retrieved all known PID genes from the database compiled by the Expert Committee of the International Union of Immunological Societies[50] and identified those corresponding to IIGs. We compared the distributions of f values between the 1,373 non-PID-associated IIGs and the 119 PID-associated IIGs for which the f parameter was available, considering their mode of inheritance. We considered only genes for which the mode of inheritance was autosomal dominant (AD) or autosomal recessive (AR) in a given subgroup of PIDs. Genes associated to an AD form in a subgroup of PID and to an AR form in another subgroup of PID were included in both AD and AR categories.

Protein-Protein Interaction Network Analysis

We reconstructed the protein-protein interaction network by retrieving the interactions from the BioGRID database v.3.2.105.[51] We retrieved protein Ensembl IDs via BioMart and considered only non-redundant direct physical interactions to compute degree centrality with the NetworkAnalyzer plugin[52] in Cytoscape.[53] Ubiquitin C and amyloid precursor protein were removed from further analysis because they display outlier degree centralities. We transformed the degree centrality to $\log_{10}$(1+degree centrality) to reduce the skewness of the distribution and used a Pearson correlation test to evaluate its relationship with the SnIPRE f parameter. We computed this correlation for the 1,114 IIGs and for the remaining 8,557 protein-coding sequences for which both degree centrality and f could be determined. We compared these correlations by using a linear model to estimate the effect of innate immunity in the relationship between f and degree centrality. For the network representation, we used our IIG list as input for Cytoscape and retrieved interactions among innate immunity proteins with the MiMI plugin.[54]

Transcription, Signal Transduction, and Innate Immunity

We retrieved the list of genes coding for proteins involved in transcription from the Gene Ontology "transcription, DNA-templated"

entry (GO: 0006351). From this list of 2,643 genes, we extracted the 2,337 genes for which *f* values were calculated via SnIPRE. We considered genes at the intersection of this GO list and our set of IIGs as involved in both innate immune response and transcription. We then compared the distribution of *f* values between this group of genes involved in both innate immunity and transcription with that of genes involved only in transcription processes. Because our set of IIGs also includes entries from InnateDB, we performed the same analyses by restricting the comparisons only between the two Gene Ontology terms "transcription, DNA-templated" and "innate immune response." The same rationale was applied to the comparisons involving signal transducers; we retrieved a list of 1,875 genes from the Gene Ontology "intracellular signal transduction" entry (GO: 0035556) and extracted the 1,695 genes for which SnIPRE *f* values were available.

Genome-Wide Detection of Positive Selection

Detection of Positive Selection via a Composite Statistics

We combined, for each SNP, the set of statistics used in previous studies,[5,55] based on haplotype homozygosity (iHS,[13] ΔiHH,[55] and XP-EHH[12]) or the degree of population differentiation (ΔDAF[55] and F_{ST}[56]). In addition, we incorporated the DIND statistics,[14] which has been found to be powerful to detect positive selection using low-coverage sequencing data.[57] For statistics based on haplotype homozygosity, we used the phased data of each population of the 1000 Genomes Project and sliding windows of 100 kb centered on each SNP. This procedure does not alter the power to detect selection and ensures each statistics to be computed using equivalent regions, in terms of recombination rate, coverage, and allele frequency spectrum.[57] Because some of these statistics require the ancestral/derived state of mutations, we retained sliding windows for which the ancestral/derived state of the core SNPs was unambiguously determined, i.e., 97% of the mutations of the 1000 Genomes dataset. Finally, we aimed to minimize the false positive rate, by excluding windows in which the core SNP had a derived allele frequency (DAF) below 0.2, because the power to detect selection at this allele frequency is limited.[57]

All neutrality statistics were then combined into a Fisher's combined score (F_{CS})

$$F_{CS} = -2\sum_{i=1}^{K} \ln(p_i)$$

where K is the number of combined statistics and p_i the empirical p value for the i^{th} statistics, i.e., the genomic rank of this i^{th} statistics divided by the total number of unique values obtained for this statistics in the entire genome (values exactly equal get the same rank and same p value). When p_i values tend to be small, the F_{CS} tends to be large. Under neutrality, F_{CS} has a chi-square distribution with $2K$ degrees of freedom. However, because the assumption of dependency among p_i is violated, we used the genomic distribution of F_{CS} in an empirical genome-wide test of selection, where the candidate SNPs with signals of selection are the ones exhibiting the 1% highest F_{CS} values, as reported for other statistics.[13,57] Note that the F_{CS} is computed for each population separately.

Statistical Analyses

Enrichments in positive selection signals among specific SNP classes (e.g., genic, located in IIGs, etc.) were tested as previously described.[8,13,57] Specifically, we used a logistic regression, generating an OR for the effect of positive selection. For a given SNP class, the OR is defined as follows:

$$OR = \left[\frac{P(class \mid SEL)}{P(\overline{class} \mid SEL)}\right]\left[\frac{P(\overline{class} \mid \overline{SEL})}{P(class \mid \overline{SEL})}\right],$$

with *class* being the SNP class, e.g., *class* and $\overline{class}$ being genic and non-genic SNPs, respectively, *SEL* and $\overline{SEL}$ being "with" and "without positive selection signal" respectively, i.e., SNP with an extreme F_{CS} value. For example, if positive selection has preferentially occurred in genic regions, an OR > 1 would be expected, reflecting an enrichment in genic SNPs among SNPs with extreme F_{CS} values (e.g., OR = 1.25 when there are 20% true and 80% false positive among genic SNPs with extreme F_{CS} values[57]). Otherwise (i.e., 100% of false positives among genic SNP outliers), we would expect an OR ~1, indicating that the proportion of genic SNPs among outliers is not greater than the expected proportion of genic SNPs among all SNPs (~38% for the 1000 Genomes Project datasets[57]). The p values of enrichment analysis were obtained from 10,000 independent resamplings, taking into account linkage disequilibrium (LD) between SNPs.[57] For each resampling, we drew non-overlapping regions of 500 consecutive SNPs and arbitrarily assigned them to a given class, until we reached the number of SNPs observed in this SNP class. We considered the remaining SNPs to be out of the given class and calculated the OR for each resampling.[57]

Identification of Candidate Regions

To identify candidate gene regions evolving adaptively, we used a conservative approach based on the degree of clustering of SNPs with extreme F_{CS} values (i.e., the 1% top F_{CS} values).[13,57] We used sliding windows of 100 kb centered on each SNP that contain at least 100 variants. We computed, for each 100 kb window, the proportion of extreme F_{CS} values and grouped these windows into 75 bins of equal sizes based on the total number of SNPs observed. Finally, we considered the 1% of windows with the highest proportion of extreme F_{CS} values in each bin as being under positive selection. A gene is thus considered to be a target of positive selection if it contains at least one window falling into this criterion.

Assessing the Power to Detect Selection

To evaluate the power to detect positive selection using the F_{CS}, we used cosi2[58] to simulate DNA regions according to realistic, accepted scenarios of human demography, as previously used for the 1000 Genomes Project dataset (for details on the parameters of the demographic model used, see Grossman et al.[5]). We simulated 60 unrelated individuals in each population sample, matching the 1000 Genomes Project dataset. We simulated 200-kb DNA regions with recombination rates sampled from the HapMap recombination map to generate realistic recombination patterns including local hotspots.[59] We simulated neutrally evolving regions and positive selection assuming the hard sweep model.[60] Specifically, a single new advantageous mutation with frequency 1/2N was inserted into the middle of the sequence in a specific population (YRI, CEU, or CHB) at a specific time t, with a population genetics selection parameter $2Ns = 100$ (selection coefficient $s = 0.01$, N = 10,000). We simulated different models of hard sweeps, by specifying various ages t of the selected allele (5 kya, 10 kya, 20 kya, and 30 kya) and various p_{sel}, i.e., the frequency of the selected allele in the current generation (0.2, 0.4, 0.6, 0.8, and 1.0). We simulated 1,000 neutral-evolving regions and 100 regions for each combination of selection parameters (t, s, and p_{sel}).

Because we used 100-kb windows centered on each SNP in the real data, and to avoid any truncation of these windows, we trimmed all simulated SNPs located at less than 50 kb of the edges of the 200 kb simulated regions. We computed the F_{CS} for each retained SNP located in the 100 kb in the middle of the 200 kb simulated region. We normalized iHS and DIND via the same method as previously described.[13,57] The empirical p value p_i used in the computation of F_{CS} was determined for each population separately, using all neutral simulations. We detected simulated regions under positive selection on the basis of the proportion of extreme F_{CS} values. The power to detect positive selection was then computed as the proportion of regions simulated under positive selection effectively detected by our statistics (i.e., the percentage of simulations presenting proportions of extreme F_{CS} values above the neutral threshold defined for a FPR of 1%).

Annotation using GWAS Hits and Immune-Related eQTLs

For the 57 IIGs presenting signatures of positive selection (i.e., innate immunity genes carrying at least one SNP whose window has a proportion of outlier F_{CS} among the 1% of genome-wide windows), we explored their involvement in diseases or traits by using hits of genome-wide association studies (GWASs) and expression quantitative trait loci (eQTLs) data. For the GWAS analyses, we used the data from the 02/06/2015 version of the NHGRI database and only GWAS signals with p values lower than 5×10^{-8} were considered. We used two approaches: the first (gene-based) approach relies on the fact that the tested IIG is the reported gene of a GWAS hit. The second (SNP-based) approach considers equivalence or strong LD between candidate SNPs for positive selection and SNPs reported as best GWAS hits. For this second LD-aware approach, we selected all outlier SNPs (i.e., F_{CS} among the 1% of genome-wide windows) in the genomic region of the tested IIG. We then retrieved all SNPs in strong LD ($r^2 > 0.8$) with any of these candidate SNPs, using the correlation coefficient implemented in Plink[61] on the unphased 1000 Genomes Project data for the relevant population (i.e., that where the F_{CS} signal was maximal). We finally checked whether some of our candidate SNPs for positive selection, or any SNP in LD with them, were among GWAS best signals. For the eQTL data, we used the same SNP-based approach to identify candidate positively selected SNPs that have been previously associated with the expression of surrounding genes in purified, stimulated monocytes (i.e., 21,516 eQTLs from Fairfax et al.[62]). We considered only the best p value obtained across stimulation conditions (FDR < 0.05).

ABC Estimation of the Age of Selection

We used an approximate Bayesian computation (ABC) approach[63] to estimate the posterior probability of the age of selection of candidate mutations according to the model described above, i.e., a new advantageous mutation, occurring at a frequency of 1/2N, in a specific population, at a specific time *t*. We simulated 200-kb regions with a selected SNP located in the center of the sequence, according to the demographic model and recombination patterns described above. To generate a set of 2×10^5 simulations, we used uniform prior distributions for the age and intensity of selection and for the current frequency of the selected allele: the age of selection (t) varies from 0 to 62,500 years, the intensity of selection (s) varies from 0.002 to 0.05, and the current frequency of the selected allele (p_{sel}) from 0.2 to 1.0. Note that the prior distributions of the age and intensity of selection do not remain uniform because some parameter vectors (t, s, p_{sel}) are unlikely (e.g., ancient selective events of strong intensity cannot generate a frequency of the selected allele equal to 0.2). Finally, we also simulated the low-coverage nature of the data (5×) by randomly drawing limited numbers of reads.[57]

We initially used a set of summary statistics (Θ_S, Θ_π, Tajima's *D*, Fay and Wu's *H*, iHS, and F_{ST}) that are informative for estimating age of selection in an ABC framework,[64] to which we incorporated the DIND statistics.[14,57] As performed for iHS and DIND, Θ_S, Θ_π, Tajima's *D*, and Fay and Wu's *H* were computed in a window of 100 kb around the selected mutation. We tested the performance of various sets of summary statistics with different ABC methods—ridge and neuralnet—implemented in the "abc" R package. We validated the performance of the ABC methods by using simulated datasets as if they were true empirical data, for which parameter values to estimate are known. This procedure allowed us to compare the estimated to the true values via various accuracy indices: the prediction error *PE* (i.e., the mean square error [MSE], divided by the prior variance of the parameter), the relative estimation bias *rEB* (i.e., the bias expressed a proportion of the true value, also known as relative error), and the coverage of the 95% credible interval (95% COV) (i.e., the percent of times where the true value was found within the 95% credible interval). Let θ_k and $\widehat{\theta_k}$ be the true and the ABC estimated values of the parameter θ in the k^{th} simulated dataset:

$$PE = \frac{\frac{1}{S}\sum_1^S\left(\widehat{\theta_k} - \theta_k\right)^2}{var(\theta)},$$

$$rEB = \frac{1}{S}\sum_1^S\left(\widehat{\theta_k} - \theta_k\right)\Big/\theta_k,$$

$$95\%COV = \frac{1}{S}\sum_1^S 1(q_1 < \theta_k < q_2),$$

where S is the number of simulated data, $1(C)$ the indicative function (equal to 1 when C is true, 0 otherwise), and q_1 and q_2 the two percentiles of the posterior distribution of $\widehat{\theta_k}$ (q_1 and q_2 were adjusted to obtain 95% *COV* approximately equal to 0.95). These accuracy indices were computed using S equal to 300 simulated data. Note that the posterior distributions were obtained by retaining the 1,000 "best" simulations in the ABC procedure.

We next aimed to improve the ABC estimations by adding more summary statistics. We first used the summary statistics described above ("set 1 of summary statistics") as well as their corresponding average and proportion of 1% top values computed over 100 kb around each candidate variant ("set 2 of summary statistics"). Furthermore, we aimed to boost the ABC estimations by including arithmetic transformations of the used summary statistics.[65] Specifically, we applied to the set 2 of summary statistics the following transformation, $T(S_i S_{j\geq i})$ with S_i and S_j the i^{th} and j^{th} summary statistics. This procedure generated a new set of summary statistics ("set 3 of summary statistics") that were used in the neuralnet ABC method, which will be referred to as the boosted-neuralnet ABC method.

Analysis of Neandertal Ancestry

To investigate introgression from archaic hominins to modern humans at IIGs, we used the probabilities of Neandertal ancestry calculated for each SNP of the 1000 Genomes Project dataset.[22] These probabilities were obtained using a conditional random field method, which takes into account the allelic state at a SNP

in non-African, Neandertal, and Yoruba individuals, the relative sequence divergence between these individuals and the consistency of haplotype lengths with estimated time of interbreeding with archaic humans.[22] We downloaded the inferred Neandertal ancestry at each allele and used the reported combined results across the CEU, GBR, FIN, IBS, and TSI populations as representing Europeans (EUR) and CHB, CHS, and JPT as representing East Asians (ASN).

We calculated the average introgression score for each protein-coding gene (i.e., removing open reading frames and genes encoding for putative proteins) as the average of the marginal probabilities of Neandertal ancestry for all bases of the gene. We then compared the distributions of the average introgression scores for our set of IIGs and the remainder of protein-coding genes. p values between distributions were obtained from 10^6 independent resamplings, taking into account the genomic correlation of average introgression scores. To this end, we retrieved genomic regions showing high probability to be introgressed from Neandertal for each population, defined as runs of SNPs that have a probability of Neandertal ancestry > 0.9. We calculated the median length of contigs (called ml_c) obtained by constructing a tiling path across confidently inferred Neandertal haplotypes in each population as described in Sankararaman et al.[22] We divided the genome in adjacent windows of length ml_c and assessed the distribution of the number of protein-coding genes and IIGs by window. Resamplings were carried out taking into account the distribution of number of IIGs by window. We used the resampled datasets to obtain the expected distribution of Mann-Whitney U under the null hypothesis and calculate the empirical p values.

Finally, we determined the 5% of genes harboring the highest probability of Neandertal ancestry at the genome-wide level and searched for those that corresponded to IIGs. These latter analyses were restricted to the CEU and CHB populations, because they were those used to detect positive selection in modern humans. For the haplotype analyses, we retrieved confidently inferred Neandertal haplotypes, i.e., runs of SNPs that present a probability of Neandertal ancestry > 0.9 (see Sankararaman et al.[22]).

Results

Building of the Innate Immunity Gene List

We established a list of innate immunity genes (IIGs) by combining two publicly available databases, Gene Ontology (GO) and InnateDB,[45,46] which we manually curated according to a number of criteria listed in Material and Methods. This yielded a curated set of 806 genes from GO and 905 from InnateDB, 345 of which were overlapping (Figure S1A). Furthermore, we incorporated an additional set of 187 genes that were missing from these datasets at the time of the study, making a final list of 1,553 IIGs (Table S1). We then classified all IIGs according to their main known (or inferred) function into nine different categories, ranging from sensors of microorganisms or danger signals to adaptor and effector molecules, and we also included regulators of the signaling pathways and accessory molecules necessary for an efficient immune response (Figure S1B).

Pervasive Signatures of Purifying Selection at Innate Immunity Genes

To define the degree of selective constraint at IIGs, we first investigated the extent to which purifying selection has acted on the different categories of IIGs since the divergence of human and chimpanzee lineages. To do so, we used the exome dataset from the 1000 Genomes Project[44] and merged all individuals into a single group to focus on the human lineage. For all protein-coding sequences, where we applied the same filtering criteria as those applied to IIGs (Material and Methods), we estimated the f parameter via SnIPRE,[47] which estimates the degree of selective constraints at each gene by using polymorphism and divergence data at non-synonymous and synonymous sites. The lower the f value, the stronger the deficit of non-synonymous mutations compared to synonymous variants, highlighting strong evolutionary constraints (Table S1).

We found that the distribution of the f parameter for IIGs was significantly skewed toward lower values (Figure 1A; resampling $p < 4.7 \times 10^{-3}$ considering gene length and number of SNPs per gene). Similar results were obtained when performing the analyses with the GO and InnateDB gene lists separately (resampling $p < 7.6 \times 10^{-4}$ and $p < 2.7 \times 10^{-3}$, respectively), indicating that our inclusion criteria of IIGs have no major impact on our conclusions. This suggests that genes involved in innate immune processes eliminate proportionally more non-synonymous variants than the remainder of protein-coding genes. When restricting the analyses to genes presenting the lowest f values at the genome-wide level, we observed a systematic enrichment in IIGs using different percentiles (Figure 1B). For example, when focusing on genes displaying the 1% lowest f values, we observed a strong, significant enrichment in IIGs (OR = 2.42, resampling $p < 4 \times 10^{-5}$), corresponding to the set of IIGs that have evolved under the strongest degree of purifying selection (Table 1). Such a significant enrichment was also observed when considering the different populations separately (OR > 1.82, resampling $p < 9.3 \times 10^{-4}$), consistent with the high correlations observed between population-specific and species-wide f values (Figure S2). This suggests that the degree of purifying selection on IIGs is similar across populations and independent of recent variation in environmental pressures.

To test the functional impact of variation at IIGs on protein structure or function, we first evaluated the deleteriousness of exonic variants by using the PHRED-scaled C-scores provided by CADD.[49] Interestingly, the proportion of variants with a scaled C score ≥ 15 (i.e., among the ~3% most deleterious mutations of the genome) was lower in IIGs compared to non-IIGs (0.425 and 0.463, respectively; $p = 2.1 \times 10^{-41}$; Figure S3A), a pattern consistently observed along the site frequency spectrum, with the exception of high-frequency alleles (Figure S3B). We next tested whether IIGs known to be involved in severe

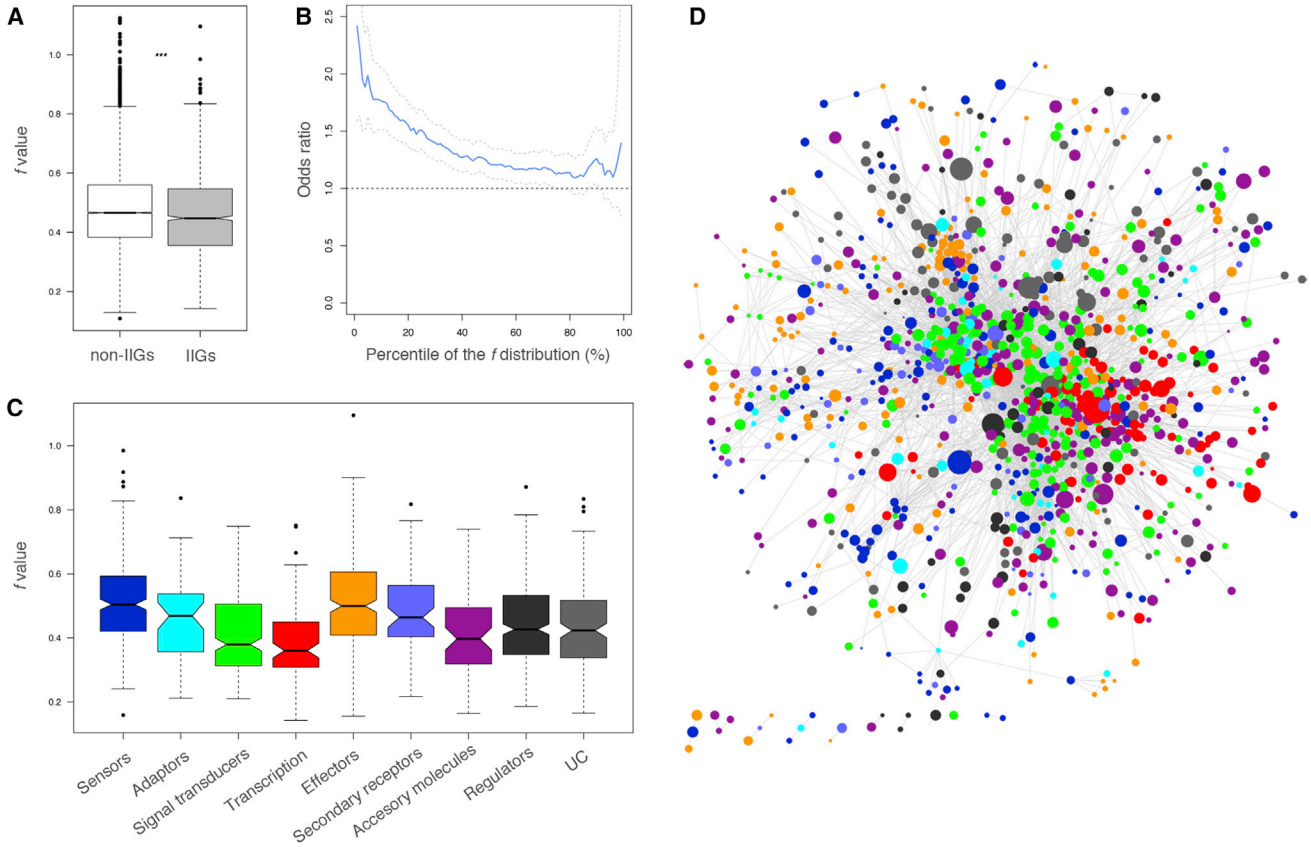

Figure 1. Varying Degrees of Selective Constraints Targeting Innate Immunity Genes
(A) Strength of purifying selection acting on innate immunity genes and the remainder of protein-coding genes, as measured by the *f* value. We tested the significance of the observed difference by means of 10^5 resamplings taking into account gene length and number of SNPs per gene in the two tested gene sets (***$p < 4.7 \times 10^{-3}$).
(B) Enrichment of innate immunity genes among the most constrained genes at the genome-wide level, as assessed by odds ratios (ORs). We calculated ORs for increasing percentiles of the *f* distribution, with a pace of 1%. The 95% confidence intervals of ORs were calculated via the Fisher's exact test.
(C) Strength of purifying selection acting on the different functional categories of innate immunity genes, as measured by the *f* value (UC stands for unclassified).
(D) Innate immunity protein interaction network. Only innate immunity proteins interacting with a molecular partner also involved in this cellular process are represented. Node sizes are negatively correlated to *f* values, i.e., large nodes represent low *f* values, indicating stronger action of purifying selection. Color codes are the same as those used in (C).

diseases such as primary immunodeficiencies (PID) were under stronger selective constraints than the remainder of IIGs. Although no significant differences were observed between the global distributions of the *f* parameter, IIGs associated with autosomal-dominant forms of PID presented significantly lower *f* values (resampling $p < 1.5 \times 10^{-2}$) than the remainder of IIGs (Figure S4).

We next assessed whether the global signal of strong purifying selection at IIGs differed among genes with distinct functional roles in innate immunity. A strongly significant difference was detected (Kruskal-Wallis rank sum test $p < 2.2 \times 10^{-16}$; Figure 1C). Molecules involved in signal transduction and transcription were those presenting the strongest selective constraints. Such constraints could indicate the additional involvement of these genes in functions other than innate immunity. To test this hypothesis, we compared the *f* values of the "signal transduction" and "transcription" groups of IIGs with those of signal transducers and transcription factors that are not part of innate immunity processes. Innate immunity molecules involved in these processes presented significantly lower *f* values than their respective comparison groups (Figure S5), suggesting that their involvement in innate immunity has further constrained their evolution. We also observed that sensor and effector molecules presented the greatest range of *f* values (Figure 1C), indicating that the degree of constraint affecting these categories varies among their members. For example, when comparing the *f* values among the different sub-families of receptors, we found that the family of cytosolic nucleic acid sensors (CNASs) displays the strongest deficit of nonsynonymous mutations (Kruskal-Wallis rank sum test $p = 2.9 \times 10^{-3}$), whereas RIG-I-like receptors (RLRs) were those evolving under the most relaxed selective constraints (Figure S6).

Table 1. Innate Immunity Genes Presenting the Strongest Signatures of Purifying Selection at the Genome-wide Level

Functional Category	Genes
Sensors	*DHX9*
Adaptors	*CNKSR2, CTNNB1, SRC, SYK*
Signal transducers	*CAMK2B, MTOR, TRAF3*
Transcription	*GATA3, HNRNPL, SMARCA2, SMARCA4, STAT1*
Effectors	*AGO1, AGO2, AGO3*
Secondary receptors	*EGFR*
Accessory molecules	*HSP90AB1, UBC*
Regulators	*CYLD, HDAC1, KHSRP, MID2, USP7*
UC	*ACTB, ACTG1, CYFIP2, DOCK1, FSCN1, ITPR1, NCKAP1, TUBB4B*

Finally, we tested whether the varying degree of selective constraints detected at the different IIGs could be partly explained by their localization in the protein-protein interaction network (PIN). We therefore reconstructed the innate immunity PIN using the protein interactions from the BioGRID database.[51] As previously observed in other protein interaction networks,[42,66,67] we observed a negative correlation between degree centrality and f values, with genes located in the center of the network (mostly signal transducers and transcription factors) presenting the strongest selective constraints (Figure 1D). Interestingly, a stronger negative correlation was observed for IIG products than for the remainder of protein-coding genes ($R = -0.341$ and -0.186, respectively; $p = 3.56 \times 10^{-4}$), suggesting a crucial role of network topology in driving the specific evolution of genes involved in innate immunity.

Identification of Regions Presenting High-Confidence Signals of Positive Selection

We next searched for IIGs that present signals of recent, population-specific positive selection, because they should contain variation that has contributed to human adaptation to varying environments. To do so, we restricted our analyses to (1) one population per geographic region—Yoruba from Nigeria (YRI), Northern Europeans (CEU), Han Chinese from Beijing (CHB)—and (2) the low-coverage dataset of the 1000 Genomes Project, because we needed to go beyond exonic regions to compute statistics based on extended haplotype homozygosity (Material and Methods). The low coverage (~5×) of this dataset has been shown to have little impact on the power of some statistics to detect positive selection.[57]

To detect signals of local adaptation, we used a composite method, the Fisher's combined score (F_{CS}), because using composite methods has been shown to increase power and minimize the detection of false positive signals.[5,55,68] We combined statistics based on haplotype homozygosity (iHS, ΔiHH, and XP-EHH), intra-allelic haplotype diversity (DIND), and population differentiation (ΔDAF and F_{ST}). We first assessed the power of F_{CS} by conducting a simulation-based study (Figure 2), which considered accepted demographic scenarios for the populations studied. We found that the power of F_{CS} was comparable to that of the composite of multiple signals (CMS) test.[5] Furthermore, the power of F_{CS}, which is not affected by the inclusion or removal of highly correlated statistics (Figure S7), declines with the age of selection, particularly among non-Africans (Figures 2 and S8). This indicates that F_{CS} tends to favor the detection of recent events of positive selection (i.e., <30 kya).

At the genome-wide level, among SNPs displaying multiple selection signals, we found a significant enrichment in genic with respect to non-genic regions ($OR > 1$, $p < 10^{-4}$; Table S2), as previously reported.[5,57] That no significant enrichment in SNPs located in IIGs was observed among SNPs with multiple selection signals, nor among the different categories of innate immunity genes (data not shown), suggests that positive selection has not targeted IIGs to a greater extent than the remainder of the genome (Table S2). However, we identified a set of IIGs presenting strong signatures of positive selection, by looking at gene regions with a high clustering of SNPs presenting selection signals.[5,13,57] Specifically, we searched for 100-kb windows with the highest (top 1%) proportions of SNPs with extreme F_{CS} values, and found 1,110 (YRI), 670 (CEU), and 1,229 (CHB) of such sliding windows, corresponding to 21, 16, and 22 genes, respectively (Table 2). We retrieved several already reported signals of positive selection, including *TLR6-TLR1-TLR10* (MIM: 605403, 601194, 606270), *IL4* (MIM: 147780), *IFIH1* (MIM: 606951), *CD36* (MIM: 173510), and *CEACAM1* (MIM: 109770),[5,8,14,36,38,69,70] but also a number of previously uncharacterized hits (Table 2).

To fine-map the candidate variants underlying the positive selection signals, we merged the SNPs from the low-coverage and exome high-coverage datasets for each of the 57 candidate genes, considering also their flanking regions (1 Mb upstream and downstream). The incorporation of the exome data allows detection of variants that have failed to pass the quality-control filters and are missing in the low-coverage data, e.g., the known functional non-synonymous SNP rs5743618 in *TLR1* (GenBank: NM_003263.3; c.1805G>T [p.Ser602Ile]).[14] We re-computed the F_{CS} statistics on this merged dataset and determined the variants that exhibited the strongest selection signals (1% variants with highest F_{CS}; dark blue dots in Figure 3). Focusing on coding variation, we identified 13 high-scoring variants (12 non-synonymous variants and 1 stop mutation) in 11 genes (Table 2 and Figure 3), some of which have been previously identified as adaptive mutations, e.g., rs5743618 in *TLR1*[14] (GenBank: NM_003263.3; c.1805G>T [p.Ser602Ile]), rs10930046 in *IFIH1*[36,38] (GenBank: NM_022168.3;

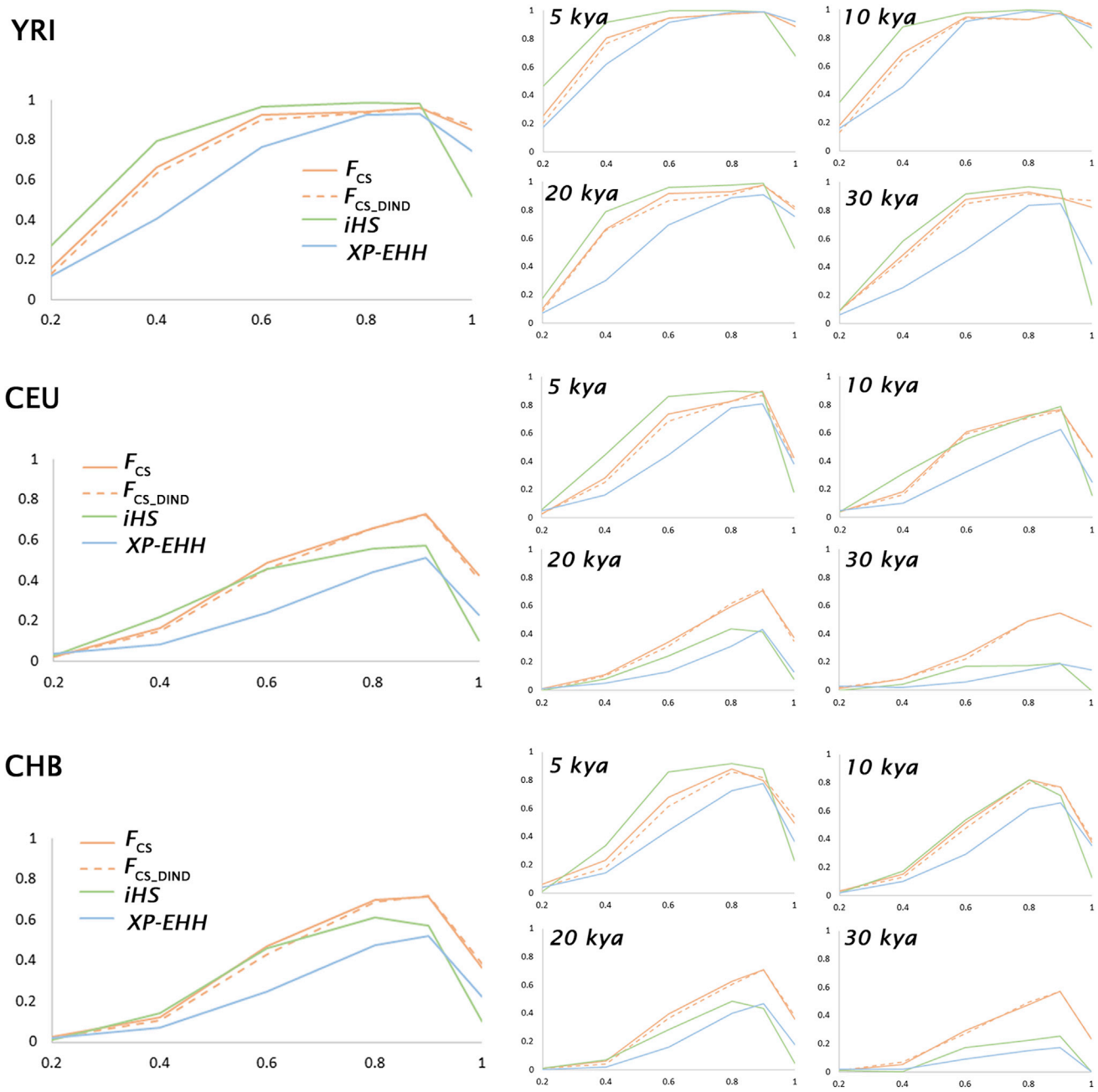

Figure 2. Power of the Fisher's Combined Score to Detect Positive Selection

We simulated 200-kb DNA regions according to accepted scenarios of human demography for West African (YRI), European (CEU), and East Asian (CHB) samples (see Material and Methods and Grossman et al.[5]). We simulated positive selection models, in each population separately, using various ages (t) of the selected allele (5 kya, 10 kya, 20 kya, and 30 kya) and current frequencies (p_{sel}) of the selected allele (0.2, 0.4, 0.6, 0.8, and 1.0), and setting the selection coefficient s to be equal to 0.01 (100 datasets for each parameter combination, see Material and Methods). For each population, we plot the power (i.e., the proportion of simulated regions under positive selection effectively detected) obtained with the F_{CS} as well as, for comparison, F_{CS_DIND} (i.e., F_{CS} removing the DIND statistics), iHS, and XP-EHH (see Material and Methods, FPR of 1%). For a detailed comparison of the differences in power of the F_{CS} with respect to different combinations of neutrality statistics, see Figure S7. Left panels show, for each population, the power obtained with ages of selection t uniformly distributed from 5 kya to 30 kya. Smaller right panels display, for each population, the power obtained with ages of positive selection of 5 kya, 10 kya, 20 kya, and 30 kya, respectively. The x axis represents the current frequency of the selected allele p_{sel}.

c.1379A>G [p.His460Arg]), and rs3211938 in *CD36*[70] (GenBank: NM_000072.3; c.975T>G [p.Tyr325Ter]).

Finally, we explored the involvement of the 57 candidate genes in diseases or traits by using GWAS and eQTL data. We found that 27 of them have been associated, to different extents, with common diseases, including susceptibility to infections or autoimmune disorders (enrichment resampling p = 3.77×10^{-4}, 3.71×10^{-2}, and 0.058 in YRI, CEU, and CHB populations, respectively, compared to all IIGs; Table S3). For 13 of these genes, we

Table 2. Innate Immunity Genes Showing the Strongest Signatures of Positive Selection

Population	Innate Immunity Genes[a]
YRI	*VSP45*, *CD1D*, *FCER1A*, *LTBP1*, *CCDC88A*, *LY75*, *IFIH1*, *LTF*, *CCR2*, *CD80*, *MAPK10*, *CD36*,[b] *ZFPM2*, *TRIM55*, *CHUK*, *DAK*, *POLR3B*, *HIF1A*, *CEACAM1*, *TNRC6B*, *MYH9*
CEU	*CCDC88A*, *TLR10*, *TLR1*, *TLR6*, *MAP3K1*, *IL4*, *IRGM*, *TRIM27*, *EYA4*, *ARPC1A*, *ZC3HAV1*, *SRPK2*, *SMARCA2*, *SIRT1*, *DUOX1*, *ADAM10*
CHB	*ARHGEF2*, *ADAM15*, *LYST*, *PELI1*, *ACTR2*, *MERTK*, *ERBB4*, *SP100*, *RAF1*, *LRRFIP2*, *CLEC3B*, *RHOA*, *GAB1*, *FER*, *ITPR3*, *EGFR*, *BLK*, *NRG1*, *SIRT1*, *OTUB1*, *ARHGEF7*, *PIAS4*

[a]These genes overlap with at least one 100-kb window with the 1% highest proportions of outlier SNPs with the highest F_{CS} values. Note that larger genes have a higher probability to be attributed to a selection signal (regardless of whether this is a true or false signal), with respect to small genes. Genes that are underlined contain at least one non-synonymous mutation with an outlier F_{CS} value: *IFIH1*: rs10930046 (GenBank: NM_022168.3; c.1379A>G [p.His460Arg]); *LTF*: rs60938611 (GenBank: NM_001199149.1; c.446C>T [p.Ala149Val]); *ZFPM2*: rs11993776 (GenBank: NM_012082.3; c.1208C>G [p.Ala403Gly]); *DAK*: rs2260655 (GenBank: NM_015533.3; c.553G>A [p.Ala185Thr]); *TLR1*: rs5743618 (GenBank: NM_003263.3; c.1805G>T [p.Ser602Ile]) and rs4833095 (GenBank: NM_003263.3; c.743A>G [p.Asn248Ser]); *TLR6*: rs5743810 (GenBank: NM_006068.2; c.745T>C [p.Ser249Pro]); *MAP3K1*: rs702689 (GenBank: NM_005921.1; c.2416G>A [p.Asp806Asn]); *MERTK*: rs7604639 (GenBank: NM_006343.2; c.1397G>A [p.Arg466Lys]) and rs2230515 (GenBank: NM_006343.2; c.1552A>G [p.Ile518Val]); *CLEC3B*: rs13963 (GenBank: NM_003278.2; c.316G>A [p.Gly106Ser]); and *NRG1*: rs3924999 (GenBank: NM_001159995.1; c.50G>A [p.Arg17Gln]).
[b]*CD36* contains an outlier premature stop mutation (rs3211938, GenBank: NM_000072.3; c.975T>G [p.Tyr325Ter]).

also identified a strong correlation between our candidate SNPs for positive selection and GWAS best hits (Table S4). In addition, we found that SNPs at 30 of the 57 genes were significantly associated with the expression of surrounding genes (Table S5), based on eQTL data from monocytes activated with various immune stimuli.[62] We therefore provide a list of high-confidence genes and mutations that might have conferred a selective advantage for local adaptation to specific human populations.

Estimating the Age of Genetic Adaptations Targeting Innate Immunity

We next aimed to estimate the age, *t*, at which positive selection has targeted the high-scoring coding variants described above (Table 2), using an ABC framework.[63] We first checked the accuracy of the ABC estimations and tested the performance of various sets of summary statistics (Material and Methods) with different ABC methods (i.e., ridge and neuralnet), using simulated datasets (Table S6). As previously reported,[64] we observed some overestimations of *t*, e.g., the best relative estimation bias (rEB) being around 0.3 when using the "set 2 of summary statistics." We therefore aimed to improve our ABC estimations and found that the best estimations were obtained when including arithmetic transformations of the summary statistics.[65] Specifically, when implementing the boosted-neuralnet method, we obtained the greatest accuracy, i.e., lowest relative estimation bias (rEB) and lowest prediction error (PE) (see "set 3 of summary statistics" in Table S6). We also noticed that the estimations of *t* are more accurate for more recent events of positive selection, as expected when considering that the power to detect selection decreases with the age of selection (Figure S9 and Table S7).

We used the boosted-neuralnet method to estimate the age of selection for the C/T-13910 polymorphism (rs4988235, GenBank: NM_005915.5; c.1917+326C>T) in the *LCT* (MIM: 603202) region, the most iconic case of positive selection in Europeans associated with lactase-persistence in adulthood.[71–73] We found an estimated age (7,100 years; 95% CI: 3,500–11,000; Table S8 and Figure S10) in good agreement with previous reports (8,000 years in Tishkoff et al.,[73] 7,400 years in Itan et al.,[72] and 11,200 in Peter et al.[64]). We therefore applied this procedure to estimate the selection ages for the 13 high-scoring coding variants at IIGs. In all cases, selection events were dated at ~6,000–13,000 years ago (Table S8 and Figures S11–S13) with a few exceptions. The most recent selection events were estimated at less than 3,900 years, for *CD36* in Africans and *NRG1* (MIM: 142445) in Asians, and the oldest was found at 35,500 years for *CLEC3B* (MIM: 187520) in Asians.

Investigating Neandertal Ancestry of Innate Immunity Genes

Recent studies of individual loci have shown that adaptive immunity genes such as *HLA* or innate immunity genes such as *STAT2* and *OAS* carry haplotypes in modern humans that appear to have introgressed from archaic populations.[23–25] In light of this, we evaluated, at the genome-wide level, the extent to which modern humans have acquired variation at IIGs via ancient admixture. We first assessed the degree of Neandertal ancestry among IIGs as a whole, taking advantage of the Neandertal introgression map.[22] We found that IIGs have a higher average introgression score when compared to the remainder of the coding genome, in both Europeans and Asians ($p = 8 \times 10^{-6}$ and $p = 2 \times 10^{-6}$, respectively; Figure 4A). Notably, these results were also significant when considering the different European and Asian subpopulations individually ($p \leq 1.8 \times 10^{-5}$ in all subpopulations). This result cannot be accounted for by the strong selective constraint detected at IIGs, because this selective regime has been associated to a decrease in Neandertal ancestry.[22]

Next, we determined the 5% of genes genome-wide harboring the highest probability of Neandertal ancestry in each population and searched for those corresponding to IIGs. Out of the sets of 810 genes presenting the highest introgression scores, we found 76 and 78 IIGs in Europeans and Asians, respectively, 28 of which were shared between the two groups (Table S9). Among these genes, we found the *OAS* gene cluster, as previously documented by a candidate gene study.[25] Importantly, we detected additional regions involved in innate immunity, including genes encoding receptors such as *NLRC5* (MIM: 613537) in Asians, transcription factors such as *IRF6* (MIM: 607199) in Asians,

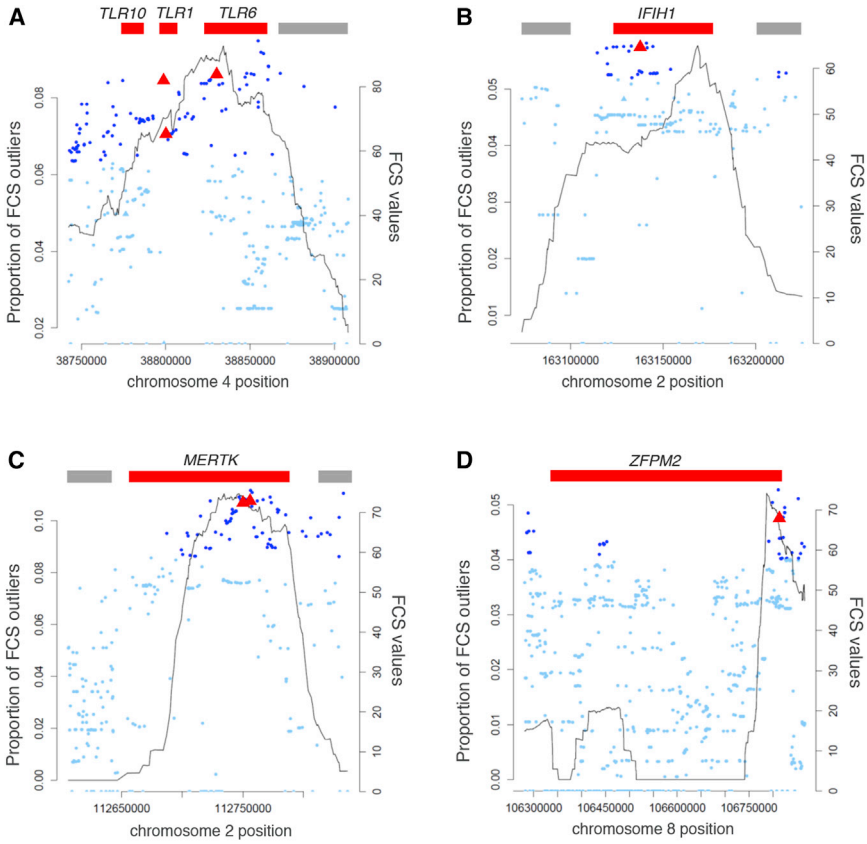

Figure 3. Innate Immunity Genes Presenting High-Confidence Signals of Geographic Adaptation

Four examples of innate immunity genes presenting strong signals of positive selection, including (A) the *TLR6-1-10* gene cluster in CEU, (B) *IFIH1* in YRI, (C) *MERTK* in CHB, and (D) *ZFPM2* in YRI. The black curves delineate the proportions of outlier SNPs (i.e., SNPs with the 1% highest F_{CS} values of the genome), within 100-kb regions, at the genome-wide level, using the low-coverage 1000 Genomes Project dataset (see Material and Methods for details). Blue dots represent the F_{CS} value of each SNP, calculated using the merged dataset (both high- and low-coverage) for the fine mapping of putative adaptive mutations. Dark blue dots indicate SNPs with the 1% highest F_{CS} values of the genome, within which non-synonymous variants are represented by red triangles. The remaining variants are plotted in light blue, where triangles represent non-synonymous mutations.

and effector molecules such as the *IFITM1-3* (MIM: 604456, 605578, 605579) gene family in Europeans and some type I IFNs in Asians.

Remarkably, two regions that we found to present high Neandertal ancestry—the *TLR6-TLR1-TLR10* gene cluster and *SIRT1* (MIM: 607199)—were also part of our high-confidence genes under positive selection in Europeans and Asians, respectively (Table 2). For these regions, we identified the SNPs that were most probably introgressed from Neandertal[22] (Figures 4B and 4C) and determined whether they were correlated to candidate SNPs for positive selection. In *SIRT1*, introgressed haplotypes were not specifically carrying any of our candidate SNPs, suggesting that variation identified as under positive selection in modern humans has not been acquired through archaic admixture. For *TLR6-TLR1-TLR10*, haplotypes of inferred Neandertal origin (16% in CEU and 49% in CHB) were tagged by SNPs detected as targets of positive selection in Europeans, including the non-synonymous SNP rs4833095 at *TLR1* (GenBank: NM_003263.3; c.743A>G [p.Asn248Ser]), which is in partial LD ($r^2 = 0.657$) with the functional rs5743618 variant (GenBank: NM_003263.3; c.1805G>T [p.Ser602Ile]). However, the rs4833095 allele most probably introgressed from Neandertal is not the putatively selected derived allele (associated with protection against asthma, allergy, and hay fever)[74–76] but the ancestral, rare allele. These patterns suggest a much more complex history than a single adaptive mutation transmitted to Eurasians on a Neandertal haplotype background (Figures 4B and 4C).

Discussion

In this study, we have taken advantage of whole-genome sequence datasets to provide a comprehensive assessment of how selection, in its different forms and intensities, has driven the evolution of innate immunity genes in humans. We must point out that any definition of "innate immunity" is arbitrary and not exempt of ambiguity. Herein, our definition of innate immunity includes intrinsic, non-hematopoietic immunity, in addition to the traditional definition of innate immunity as hematopoietic and non-adaptive. Bearing this in mind, there are several important insights that can be drawn from our study.

First, we show that innate immunity genes evolve under stronger selective constraints than the remainder of protein-coding genes, indicating that the purge of deleterious mutations has been particularly important in this gene class. This observation is consistent with pathogens being one of the most important long-time threats to human survival.[4,15] Furthermore, innate immunity is germline encoded, unlike adaptive immunity whose variation is mostly somatic,[28,29] and ensures the sensing of pathogens and the maintenance of homeostasis with symbiotic microbiota.[27] Consequently, any mutation disturbing these processes would be deleterious and rapidly eliminated from the population. That the strength of selective constraints varies considerably among functional categories, as well as among the different members within each, informs us about the degree of redundancy or essentiality of the corresponding genes. For example, among sensors, the relaxed constraints of cytosolic RLRs attest to higher

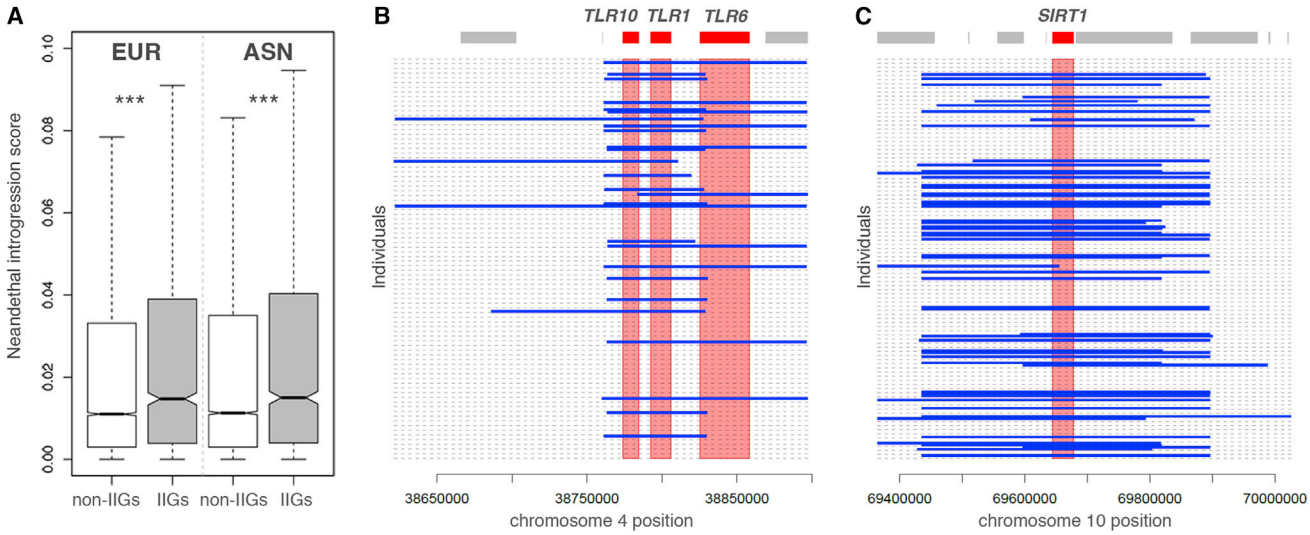

Figure 4. Neandertal Ancestry of Innate Immunity Genes

(A) Comparison of the average introgression scores of innate immunity genes (IIGs) with respect to the remainder of protein-coding genes (non-IIGs) in European (EUR) and East Asian (ASN) populations. ***$p < 0.001$ (see Material and Methods).

(B and C) Haplotypes of Neandertal ancestry in (B) CEU individuals at the *TLR6-TLR1-TLR10* gene cluster and (C) CHB individuals at the *SIRT1* locus. Confidently inferred haplotypes of Neandertal ancestry, defined as long runs of SNPs that present a probability of Neandertal ancestry > 0.9,[22] are indicated in blue in each diploid individual from the 1000 Genomes Project. Red shadows highlight genomic regions containing innate immunity genes.

immunological redundancy,[39] whereas the strong purifying selection detected for the CNASs, similar to that of endosomal TLRs,[14] suggests that variation at these molecules might be strongly deleterious for the host.

That genes evolving under strong purifying selection are likely to fulfill essential, non-redundant functions in host defense is supported by the observation that innate immunity genes associated with autosomal-dominant forms of primary immunodeficiencies present the strongest evolutionary constraints (Figure S4). This is well illustrated by the cases of *STAT1* (MIM: 600555) and *TRAF3* (MIM: 601896),[2] which are among the 1% most constrained of the genome (Table 1). Indeed, gain- and loss-of-function mutations at STAT1 have been associated with a range of immunological and clinical phenotypes, including life-threatening and mild bacterial and viral diseases, Mendelian susceptibility to mycobacterial disease, chronic mucocutaneous candidiasis, and autoimmunity.[77] Similarly, deficiency in TRAF3 has been associated with herpes simplex virus 1 encephalitis, a devastating infection of the central nervous system.[78] These examples support the notion that the genes we report as targeted by strong purifying selection are of major biological relevance in host survival. Given the pleiotropic functions of many innate immunity genes, it is conceivable that some of these genes might be involved in mechanisms that can go beyond immunity, including housekeeping functions. Regardless of the breadth of their biological functions, mutations in highly constrained IIGs are likely to predispose individuals to life-threatening disease; combining next-generation sequencing and evolutionary data in clinical studies should facilitate the discovery of novel genetic etiologies of severe, infectious disease phenotypes.

Second, our study shows that innate immunity genes have not undergone hard sweeps to a greater extent than the remainder of the genome, supporting the notion that polygenic adaptation has been pervasive among functions related to innate immunity.[79] However, we identify 57 genes presenting high-confidence signals of selective sweeps in specific populations (Table 2). Most of the signals detected in Europeans and Asians are accounted for by nearly complete sweeps (mean selected allele frequency of 67% [min 41%; max 88%] and 81% [min 31%; max 99%], respectively), whereas those identified in Africans are largely explained by partial sweeps (mean 41% [min 22%; max 92%]). The general dearth of nearly complete sweep signals in Africans is consistent with genome-wide patterns,[12,68,80] suggesting that increased drift and geographic differences in selection pressures might have inflated the number of nearly complete sweep signals among non-Africans.[80,81]

Our age estimations show that most adaptations targeting coding variation at innate immunity genes have occurred in the last 6,000–13,000 years. Some over-representation of "young" positive selection events occurring in the last 30,000 years is expected, because our simulations show that the power of F_{CS} tends to linearly decline as the age of selection increases (Figures 2 and S8). However, this decline was found to be moderate in African populations, suggesting that their enrichment in recent selection signals cannot be explained only by the power of F_{CS}. Despite the possible underestimation of older selection events, our estimated age ranges correspond well

with the transition from food collection (hunting/gathering) to food production (farming/herding) starting 10,000–13,000 years ago.[82] In this context, a recent study of West Eurasian ancient samples dating to between 8,500 and 3,000 years ago has detected strong signals of positive selection at loci related to pigmentation, diet, and immunity, supporting a scenario of Neolithic populations adapting to sedentary agricultural lifestyles.[83] The strongest selective signal they detect is the lactase persistence allele, the origin of which we estimate at 7,100 years, which reached appreciable frequencies in early farmers (~10%) around 4,300 years ago. The shift to agriculture was also accompanied by increased population density, food crises, and contacts with cattle and biological wastes, which modified human exposure to pathogens[84] and, as our results suggest, was associated to some degree of genetic adaptation.

Several of our high-scoring positively selected genes have been associated with common infectious, inflammatory, or autoimmune phenotypes (Table S3) and contain variants that have been previously detected as evolving adaptively in specific populations; e.g., *IFIH1* and *CD36* in Africans, the *TLR6-TLR1-TLR10* cluster in Europeans, and *BLK* (MIM: 191305) in Asians.[5,8,14,36,38,69,70] The case of the stop mutation Thr1264Gly at *CD36* (rs3211938, GenBank: NM_000072.3; c.975T>G [p.Tyr325Ter]) is particularly worth discussing. CD36 is an archetypal pattern recognition receptor that mediates cytoadherence of *Plasmodium falciparum*-parasitized erythrocytes.[85,86] Although the association between *CD36* and malaria remains complex,[70,87] the stop variant represents a well-supported case of positive selection,[70,88] which our analysis confirms, reaching a frequency of 29% in the Yoruba from Nigeria. It has been proposed that the high frequency of this variant in Nigeria results from a geographically confined selective event.[70] That we estimate the age of the stop variant at only 3,600 years (95% CI: 2,125–5,025 years) strongly supports the notion that the increase in frequency of this mutation restricted to west-central Africa represents a local, recent, and strong event of genetic adaptation. Investigating the association between *CD36* and malaria phenotypes specifically in Nigerian populations is now needed.

Importantly, our list of positively selected candidate genes includes hits that have not been previously detected as targets of selection but contain SNPs that are associated with immunity-related phenotypes (Tables S3 and S4). Notably, we have detected two high-scoring non-synonymous mutations in *MERTK* ([MIM: 604705], rs7604639, GenBank: NM_006343.2; c.1397G>A [p.Arg466Lys], and rs2230515, GenBank: NM_006343.2; c.1552A>G [p.Ile518Val]; Figure 3C), with a derived allele frequency of 79% in the Asian population. Interestingly, variation at *MERTK*, a member of the 3 TAM receptor tyrosine kinases that are involved in the regulation of inflammatory responses, has been associated with hepatitis C-induced liver fibrosis.[89] Likewise, our analysis suggests that positive selection has increased the frequency to 58% in Africans of a non-synonymous mutation in *ZFPM2* ([MIM: 603693], rs11993776, GenBank: NM_012082.3; c.1208C>G [p.Ala403Gly]; Figure 3D). ZFPM2 modulates the activity of GATA transcription factors at the *HAMP* (MIM: 606464) promoter, an antimicrobial peptide involved in the metabolism of iron, which is critical for *Mycobacterium tuberculosis* growth in macrophages.[90] Variation at *ZFPM2* has been recently suggested to be associated with susceptibility to tuberculosis in a South African admixed population.[91] Altogether, we provide a tractable list of high-scoring selected coding variants for experimental follow-up, which are likely to have played a dominant role in recent adaptations of human populations to their respective environments.

Finally, our study provides insight into the degree of introgression of innate immunity genes from archaic hominins. It has been shown that protein-coding genes are generally depleted in Neandertal ancestry, owing to the widespread effects of negative selection against Neandertal ancestry in gene regions.[22] Interestingly, we find that innate immunity genes present a higher average probability of Neandertal ancestry than the remainder of the coding genome. Among the genes presenting the highest Neandertal ancestry we find the IFITM1-3 proteins (Table S9), a family of restriction factors that restrict the replication of multiple viruses in vitro, including influenza A, dengue, and West Nile.[92] In particular, variation at *IFITM3* has been suggested to alter the morbidity and mortality associated with influenza infection in humans.[93] That innate immunity genes present evidence of both strong purifying selection and high Neandertal ancestry suggests either a weaker purge or a slightly stronger selective advantage of Neandertal alleles in innate immunity genes in Eurasian populations.

Our analyses suggest that neutral introgression is the most likely explanation for innate immunity genes presenting high Neandertal ancestry, because virtually none of them present any positive selection signal in modern humans. However, this lack of adaptive introgression signals could be also explained by the limited power of the F_{CS} statistics to detect old selection events and, most importantly, other selective regimes, such as polygenic adaptation or selection on standing variation. For example, alleles introgressed from archaic hominins probably had a substantial population frequency when positive selection started to act on them in modern populations. This emphasizes the need to extend our study by including, and developing, statistics empowered to detect more subtle models of positive selection.

Despite this potential limitation, the case of the *TLR6-TLR1-TLR10* cluster is particularly worth discussing. First, it presents high Neandertal introgression scores in both Europeans and Asians (Figure 4B and Table S9). Second, *TLR6-TLR1-TLR10* has been proposed to be a hotspot of positive selection in human and non-human primates.[94] Third, this gene region, here and elsewhere, is detected as a strongly supported case of local adaptation in

Europeans (Table 2).[14,41,95] Further support for the adaptive significance of *TLR6-TLR1-TLR10* comes from ancient DNA data, where this gene cluster appears to be among the strongest signals of selection detected.[83] Furthermore, three high-scoring adaptive non-synonymous mutations have been detected in this region (Figure 3A), one of which (rs5743618 in *TLR1*, GenBank: NM_003263.3; c.1805G>T [p.Ser602Ile]) appears to be the genuine target of selection; it remarkably impairs agonist-induced NF-κB activation by up to 60% and is linked to infectious disease phenotypes, such as leprosy.[14,41,96,97] However, this hyporesponsiveness allele is not present in the Neandertal genomes.[98] More generally, for SNPs most probably introgressed from Neandertal that show signatures of positive selection in Europeans, the alleles present in archaic hominins are rare and ancestral in modern humans, whereas positive selection has targeted the frequent, derived alleles. Altogether, although we provide compelling evidence supporting both high Neandertal ancestry and positive selection for functional mutations at *TLR6-TLR1-TLR10*, our analyses show that Neandertal introgression is probably not the source of such adaptive variation. Future studies should experimentally test whether the Neandertal-introgressed variation detected at this gene cluster has any functional impact on TLR-mediated immunity to infection.

In summary, our analyses have shown that the contemporary diversity of innate immunity genes in humans results from the intermingling of different demographic and selective events, including introgression from Neandertal, hard sweeps at some loci in specific human populations occurring mostly during the Neolithic transition, and continued selective constraints at other loci. In doing so, they increase our understanding of the degree of essentiality and adaptability of innate immunity genes, with several candidates for having played a dominant role in recent adaptations, and provide insight into the extent to which modern humans might have acquired variation at innate immunity genes through admixture with archaic hominins.

Supplemental Data

Supplemental Data include 13 figures and 9 tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2015.11.014.

Acknowledgments

The authors thank Matthew Albert and Jean-Marc Cavaillon for useful advice and discussions. This work was supported by the Institut Pasteur, the Centre National de la Recherche Scientifique (CNRS), and the Agence Nationale de la Recherche (ANR) grants: "DEMOCHIPS" ANR-12-BSV7-0012, "IEIHSEER" ANR-14-CE14-0008-02, and "TBPATHGEN" ANR-14-CE14-0007-02. The laboratory of L.Q.-M. has received funding from the French Government's Investissement d'Avenir program, Laboratoire d'Excellence "Integrative Biology of Emerging Infectious Diseases" (grant no. ANR-10-LABX-62-IBEID), and from the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013)/ERC Grant Agreement No. 281297.

Received: August 17, 2015
Accepted: November 6, 2015
Published: January 7, 2016

Web Resources

The URLs for data presented herein are as follows:

1000 Genomes, http://browser.1000genomes.org
CADD, http://cadd.gs.washington.edu/
CRAN – Package abc, https://cran.r-project.org/web/packages/abc/index.html
Datasets – Neandertal Introgression, http://genetics.med.harvard.edu/reichlab/Reich_Lab/Datasets_-_Neandertal_Introgression.html
Gene Ontology Consortium, http://geneontology.org/
GWAS Catalog, http://www.genome.gov/gwastudies/
InnateDB, http://www.innatedb.ca/
OMIM, http://www.omim.org/
PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/
RefSeq, http://www.ncbi.nlm.nih.gov/RefSeq
UCSC Genome Browser, http://genome.ucsc.edu

References

1. Casanova, J.L., and Abel, L. (2005). Inborn errors of immunity to infection: the rule rather than the exception. J. Exp. Med. *202*, 197–201.
2. Casanova, J.L., Abel, L., and Quintana-Murci, L. (2013). Immunology taught by human genetics. Cold Spring Harb. Symp. Quant. Biol. *78*, 157–172.
3. Chapman, S.J., and Hill, A.V. (2012). Human genetic susceptibility to infectious disease. Nat. Rev. Genet. *13*, 175–188.
4. 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.
5. Grossman, S.R., Andersen, K.G., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., Park, D.J., Griesemer, D., Karlsson, E.K., Wong, S.H., et al.; 1000 Genomes Project (2013). Identifying recent adaptations in large-scale genomic data. Cell *152*, 703–713.
6. Fumagalli, M., and Sironi, M. (2014). Human genome variability, natural selection and infectious diseases. Curr. Opin. Immunol. *30*, 9–16.
7. Karlsson, E.K., Kwiatkowski, D.P., and Sabeti, P.C. (2014). Natural selection and infectious disease in human populations. Nat. Rev. Genet. *15*, 379–393.
8. 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.
9. Bustamante, C.D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M.T., Glanowski, S., Tanenbaum, D.M., White, T.J., Sninsky, J.J., Hernandez, R.D., et al. (2005). Natural selection on protein-coding genes in the human genome. Nature *437*, 1153–1157.
10. Leffler, E.M., Gao, Z., Pfeifer, S., Ségurel, L., Auton, A., Venn, O., Bowden, R., Bontrop, R., Wall, J.D., Sella, G., et al. (2013). Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science *339*, 1578–1582.

11. 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.
12. 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.
13. 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.
14. Barreiro, L.B., Ben-Ali, M., Quach, H., Laval, G., Patin, E., Pickrell, J.K., Bouchier, C., Tichit, M., Neyrolles, O., Gicquel, B., et al. (2009). Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. PLoS Genet. *5*, e1000562.
15. 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.
16. Alcaïs, A., Quintana-Murci, L., Thaler, D.S., Schurr, E., Abel, L., and Casanova, J.L. (2010). Life-threatening infectious diseases of childhood: single-gene inborn errors of immunity? Ann. N Y Acad. Sci. *1214*, 18–33.
17. Blekhman, R., Man, O., Herrmann, L., Boyko, A.R., Indap, A., Kosiol, C., Bustamante, C.D., Teshima, K.M., and Przeworski, M. (2008). Natural selection on genes that underlie human disease susceptibility. Curr. Biol. *18*, 883–889.
18. Key, F.M., Teixeira, J.C., de Filippo, C., and Andrés, A.M. (2014). Advantageous diversity maintained by balancing selection in humans. Curr. Opin. Genet. Dev. *29*, 45–51.
19. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H., et al. (2010). A draft sequence of the Neandertal genome. Science *328*, 710–722.
20. Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., de Filippo, C., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. Science *338*, 222–226.
21. Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., Johnson, P.L., et al. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature *468*, 1053–1060.
22. Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., and Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. Nature *507*, 354–357.
23. Abi-Rached, L., Jobin, M.J., Kulkarni, S., McWhinnie, A., Dalva, K., Gragert, L., Babrzadeh, F., Gharizadeh, B., Luo, M., Plummer, F.A., et al. (2011). The shaping of modern human immune systems by multiregional admixture with archaic humans. Science *334*, 89–94.
24. Mendez, F.L., Watkins, J.C., and Hammer, M.F. (2012). A haplotype at STAT2 introgressed from neanderthals and serves as a candidate of positive selection in Papua New Guinea. Am. J. Hum. Genet. *91*, 265–274.
25. Mendez, F.L., Watkins, J.C., and Hammer, M.F. (2013). Neandertal origin of genetic variation at the cluster of OAS immunity genes. Mol. Biol. Evol. *30*, 798–801.
26. Quintana-Murci, L., and Clark, A.G. (2013). Population genetic tools for dissecting innate immunity in humans. Nat. Rev. Immunol. *13*, 280–293.
27. Takeuchi, O., and Akira, S. (2010). Pattern recognition receptors and inflammation. Cell *140*, 805–820.
28. Brodin, P., Jojic, V., Gao, T., Bhattacharya, S., Angel, C.J., Furman, D., Shen-Orr, S., Dekker, C.L., Swan, G.E., Butte, A.J., et al. (2015). Variation in the human immune system is largely driven by non-heritable influences. Cell *160*, 37–47.
29. Casanova, J.L., and Abel, L. (2015). Disentangling inborn and acquired immunity in human twins. Cell *160*, 13–15.
30. Mukherjee, S., Sarkar-Roy, N., Wagener, D.K., and Majumder, P.P. (2009). Signatures of natural selection are not uniform across genes of innate immune system, but purifying selection is the dominant signature. Proc. Natl. Acad. Sci. USA *106*, 7073–7078.
31. Wlasiuk, G., and Nachman, M.W. (2010). Adaptation and constraint at Toll-like receptors in primates. Mol. Biol. Evol. *27*, 2172–2186.
32. Ferrer-Admetlla, A., Bosch, E., Sikora, M., Marquès-Bonet, T., Ramírez-Soriano, A., Muntasell, A., Navarro, A., Lazarus, R., Calafell, F., Bertranpetit, J., and Casals, F. (2008). Balancing selection is the main force shaping the evolution of innate immunity genes. J. Immunol. *181*, 1315–1322.
33. Ferrer-Admetlla, A., Sikora, M., Laayouni, H., Esteve, A., Roubinet, F., Blancher, A., Calafell, F., Bertranpetit, J., and Casals, F. (2009). A natural history of FUT2 polymorphism in humans. Mol. Biol. Evol. *26*, 1993–2003.
34. Fornarino, S., Laval, G., Barreiro, L.B., Manry, J., Vasseur, E., and Quintana-Murci, L. (2011). Evolution of the TIR domain-containing adaptors in humans: swinging between constraint and adaptation. Mol. Biol. Evol. *28*, 3087–3097.
35. Ferwerda, B., McCall, M.B., Alonso, S., Giamarellos-Bourboulis, E.J., Mouktaroudi, M., Izagirre, N., Syafruddin, D., Kibiki, G., Cristea, T., Hijmans, A., et al. (2007). TLR4 polymorphisms, infectious diseases, and evolutionary pressure during migration of modern humans. Proc. Natl. Acad. Sci. USA *104*, 16645–16650.
36. Fumagalli, M., Cagliani, R., Riva, S., Pozzoli, U., Biasin, M., Piacentini, L., Comi, G.P., Bresolin, N., Clerici, M., and Sironi, M. (2010). Population genetics of IFIH1: ancient population structure, local selection, and implications for susceptibility to type 1 diabetes. Mol. Biol. Evol. *27*, 2555–2566.
37. Manry, J., Laval, G., Patin, E., Fornarino, S., Itan, Y., Fumagalli, M., Sironi, M., Tichit, M., Bouchier, C., Casanova, J.L., et al. (2011). Evolutionary genetic dissection of human interferons. J. Exp. Med. *208*, 2747–2759.
38. Vasseur, E., Patin, E., Laval, G., Pajon, S., Fornarino, S., Crouau-Roy, B., and Quintana-Murci, L. (2011). The selective footprints of viral pressures at the human RIG-I-like receptor family. Hum. Mol. Genet. *20*, 4462–4474.
39. Vasseur, E., Boniotto, M., Patin, E., Laval, G., Quach, H., Manry, J., Crouau-Roy, B., and Quintana-Murci, L. (2012). The evolutionary landscape of cytosolic microbial sensors in humans. Am. J. Hum. Genet. *91*, 27–37.
40. Hollox, E.J., and Armour, J.A. (2008). Directional and balancing selection in human beta-defensins. BMC Evol. Biol. *8*, 113.
41. Laayouni, H., Oosting, M., Luisi, P., Ioana, M., Alonso, S., Ricaño-Ponce, I., Trynka, G., Zhernakova, A., Plantinga, T.S., Cheng, S.C., et al. (2014). Convergent evolution in European and Roma populations reveals pressure exerted by plague on Toll-like receptors. Proc. Natl. Acad. Sci. USA *111*, 2668–2673.
42. Casals, F., Sikora, M., Laayouni, H., Montanucci, L., Muntasell, A., Lazarus, R., Calafell, F., Awadalla, P., Netea, M.G., and Bertranpetit, J. (2011). Genetic adaptation of the antibacterial human innate immunity network. BMC Evol. Biol. *11*, 202.

43. Cagliani, R., Forni, D., Biasin, M., Comabella, M., Guerini, F.R., Riva, S., Pozzoli, U., Agliardi, C., Caputo, D., Malhotra, S., et al. (2014). Ancient and recent selective pressures shaped genetic diversity at AIM2-like nucleic acid sensors. Genome Biol. Evol. *6*, 830–845.
44. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
45. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.
46. Breuer, K., Foroushani, A.K., Laird, M.R., Chen, C., Sribnaia, A., Lo, R., Winsor, G.L., Hancock, R.E., Brinkman, F.S., and Lynn, D.J. (2013). InnateDB: systems biology of innate immunity and beyond–recent updates and continuing curation. Nucleic Acids Res. *41*, D1228–D1233.
47. Eilertson, K.E., Booth, J.G., and Bustamante, C.D. (2012). SnIPRE: selection inference using a Poisson random effects model. PLoS Comput. Biol. *8*, e1002806.
48. Cingolani, P., Platts, A., Wang, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80–92.
49. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.
50. Al-Herz, W., Bousfiha, A., Casanova, J.L., Chatila, T., Conley, M.E., Cunningham-Rundles, C., Etzioni, A., Franco, J.L., Gaspar, H.B., Holland, S.M., et al. (2014). Primary immunodeficiency diseases: an update on the classification from the international union of immunological societies expert committee for primary immunodeficiency. Front. Immunol. *5*, 162.
51. Chatr-Aryamontri, A., Breitkreutz, B.J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., et al. (2015). The BioGRID interaction database: 2015 update. Nucleic Acids Res. *43*, D470–D478.
52. Assenov, Y., Ramírez, F., Schelhorn, S.E., Lengauer, T., and Albrecht, M. (2008). Computing topological parameters of biological networks. Bioinformatics *24*, 282–284.
53. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.
54. Gao, J., Ade, A.S., Tarcea, V.G., Weymouth, T.E., Mirel, B.R., Jagadish, H.V., and States, D.J. (2009). Integrating and annotating the interactome using the MiMI plugin for cytoscape. Bioinformatics *25*, 137–138.
55. Grossman, S.R., Shlyakhter, I., Karlsson, E.K., Byrne, E.H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., Zuk, O., et al. (2010). A composite of multiple signals distinguishes causal variants in regions of positive selection. Science *327*, 883–886.
56. Holsinger, K.E., and Weir, B.S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nat. Rev. Genet. *10*, 639–650.
57. Fagny, M., Patin, E., Enard, D., Barreiro, L.B., Quintana-Murci, L., and Laval, G. (2014). Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. Mol. Biol. Evol. *31*, 1850–1868.
58. Shlyakhter, I., Sabeti, P.C., and Schaffner, S.F. (2014). Cosi2: an efficient simulator of exact and approximate coalescent with selection. Bioinformatics *30*, 3427–3429.
59. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al.; International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.
60. 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.
61. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
62. Fairfax, B.P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R., McGee, C., and Knight, J.C. (2014). Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science *343*, 1246949.
63. Beaumont, M.A., Zhang, W., and Balding, D.J. (2002). Approximate Bayesian computation in population genetics. Genetics *162*, 2025–2035.
64. Peter, B.M., Huerta-Sanchez, E., and Nielsen, R. (2012). Distinguishing between selective sweeps from standing variation and from a de novo mutation. PLoS Genet. *8*, e1003011.
65. Aeschbacher, S., Beaumont, M.A., and Futschik, A. (2012). A novel approach for choosing summary statistics in approximate Bayesian computation. Genetics *192*, 1027–1047.
66. Hahn, M.W., and Kern, A.D. (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol. Biol. Evol. *22*, 803–806.
67. Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C., and Feldman, M.W. (2002). Evolutionary rate in the protein interaction network. Science *296*, 750–752.
68. Pybus, M., Luisi, P., Dall'Olio, G.M., Uzkudun, M., Laayouni, H., Bertranpetit, J., and Engelken, J. (2015). Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. Bioinformatics. Published online August 26, 2015. http://dx.doi.org/10.1093/bioinformatics/btv493.
69. Colonna, V., Ayub, Q., Chen, Y., Pagani, L., Luisi, P., Pybus, M., Garrison, E., Xue, Y., Tyler-Smith, C., Abecasis, G.R., et al.; 1000 Genomes Project Consortium (2014). Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. Genome Biol. *15*, R88.
70. Fry, A.E., Ghansa, A., Small, K.S., Palma, A., Auburn, S., Diakite, M., Green, A., Campino, S., Teo, Y.Y., Clark, T.G., et al. (2009). Positive selection of a CD36 nonsense variant in sub-Saharan Africa, but no association with severe malaria phenotypes. Hum. Mol. Genet. *18*, 2683–2692.
71. Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E., and Hirschhorn, J.N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. Am. J. Hum. Genet. *74*, 1111–1120.

72. Itan, Y., Powell, A., Beaumont, M.A., Burger, J., and Thomas, M.G. (2009). The origins of lactase persistence in Europe. PLoS Comput. Biol. *5*, e1000491.
73. Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M., et al. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. Nat. Genet. *39*, 31–40.
74. Bønnelykke, K., Matheson, M.C., Pers, T.H., Granell, R., Strachan, D.P., Alves, A.C., Linneberg, A., Curtin, J.A., Warrington, N.M., Standl, M., et al.; Australian Asthma Genetics Consortium (AAGC); EArly Genetics and Lifecourse Epidemiology (EAGLE) Consortium (2013). Meta-analysis of genome-wide association studies identifies ten loci influencing allergic sensitization. Nat. Genet. *45*, 902–906.
75. Ferreira, M.A., Matheson, M.C., Tang, C.S., Granell, R., Ang, W., Hui, J., Kiefer, A.K., Duffy, D.L., Baltic, S., Danoy, P., et al.; Australian Asthma Genetics Consortium Collaborators (2014). Genome-wide association analysis identifies 11 risk variants associated with the asthma with hay fever phenotype. J. Allergy Clin. Immunol. *133*, 1564–1571.
76. Hinds, D.A., McMahon, G., Kiefer, A.K., Do, C.B., Eriksson, N., Evans, D.M., St Pourcain, B., Ring, S.M., Mountain, J.L., Francke, U., et al. (2013). A genome-wide association meta-analysis of self-reported allergy identifies shared and allergy-specific susceptibility loci. Nat. Genet. *45*, 907–911.
77. Boisson-Dupuis, S., Kong, X.F., Okada, S., Cypowyj, S., Puel, A., Abel, L., and Casanova, J.L. (2012). Inborn errors of human STAT1: allelic heterogeneity governs the diversity of immunological and infectious phenotypes. Curr. Opin. Immunol. *24*, 364–378.
78. Pérez de Diego, R., Sancho-Shimizu, V., Lorenzo, L., Puel, A., Plancoulaine, S., Picard, C., Herman, M., Cardon, A., Durandy, A., Bustamante, J., et al. (2010). Human TRAF3 adaptor molecule deficiency leads to impaired Toll-like receptor 3 response and susceptibility to herpes simplex encephalitis. Immunity *33*, 400–411.
79. Daub, J.T., Hofer, T., Cutivet, E., Dupanloup, I., Quintana-Murci, L., Robinson-Rechavi, M., and Excoffier, L. (2013). Evidence for polygenic adaptation to pathogens in the human genome. Mol. Biol. Evol. *30*, 1544–1558.
80. Coop, G., Pickrell, J.K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R.M., Cavalli-Sforza, L.L., Feldman, M.W., and Pritchard, J.K. (2009). The role of geography in human adaptation. PLoS Genet. *5*, e1000500.
81. Messer, P.W., and Petrov, D.A. (2013). Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol. Evol. *28*, 659–669.
82. Diamond, J., and Bellwood, P. (2003). Farmers and their languages: the first expansions. Science *300*, 597–603.
83. Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. Nature. Published online November 23, 2015. http://dx.doi.org/10.1038/nature16152.
84. Wolfe, N.D., Dunavan, C.P., and Diamond, J. (2007). Origins of major human infectious diseases. Nature *447*, 279–283.
85. Newbold, C., Craig, A., Kyes, S., Rowe, A., Fernandez-Reyes, D., and Fagan, T. (1999). Cytoadherence, pathogenesis and the infected red cell surface in Plasmodium falciparum. Int. J. Parasitol. *29*, 927–937.
86. Hoebe, K., Georgel, P., Rutschmann, S., Du, X., Mudd, S., Crozat, K., Sovath, S., Shamel, L., Hartung, T., Zähringer, U., and Beutler, B. (2005). CD36 is a sensor of diacylglycerides. Nature *433*, 523–527.
87. Malaria Genomic Epidemiology Network; Malaria Genomic Epidemiology Network (2014). Reappraisal of known malaria resistance loci in a large multicenter study. Nat. Genet. *46*, 1197–1204.
88. Bhatia, G., Patterson, N., Pasaniuc, B., Zaitlen, N., Genovese, G., Pollack, S., Mallick, S., Myers, S., Tandon, A., Spencer, C., et al. (2011). Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. Am. J. Hum. Genet. *89*, 368–381.
89. Patin, E., Kutalik, Z., Guergnon, J., Bibert, S., Nalpas, B., Jouanguy, E., Munteanu, M., Bousquet, L., Argiro, L., Halfon, P., et al.; Swiss Hepatitis C Cohort Study Group; International Hepatitis C Genetics Consortium; French ANRS HC EP 26 Genoscan Study Group (2012). Genome-wide association study identifies variants associated with progression of liver fibrosis from HCV infection. Gastroenterology *143*, 1244–52.e1, 12.
90. Boelaert, J.R., Vandecasteele, S.J., Appelberg, R., and Gordeuk, V.R. (2007). The effect of the host's iron status on tuberculosis. J. Infect. Dis. *195*, 1745–1753.
91. Chimusa, E.R., Zaitlen, N., Daya, M., Möller, M., van Helden, P.D., Mulder, N.J., Price, A.L., and Hoal, E.G. (2014). Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. Hum. Mol. Genet. *23*, 796–809.
92. Brass, A.L., Huang, I.C., Benita, Y., John, S.P., Krishnan, M.N., Feeley, E.M., Ryan, B.J., Weyer, J.L., van der Weyden, L., Fikrig, E., et al. (2009). The IFITM proteins mediate cellular resistance to influenza A H1N1 virus, West Nile virus, and dengue virus. Cell *139*, 1243–1254.
93. Everitt, A.R., Clare, S., Pertel, T., John, S.P., Wash, R.S., Smith, S.E., Chin, C.R., Feeley, E.M., Sims, J.S., Adams, D.J., et al.; GenISIS Investigators; MOSAIC Investigators (2012). IFITM3 restricts the morbidity and mortality associated with influenza. Nature *484*, 519–523.
94. Enard, D., Depaulis, F., and Roest Crollius, H. (2010). Human and non-human primate genomes share hotspots of positive selection. PLoS Genet. *6*, e1000840.
95. 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.
96. Johnson, C.M., Lyle, E.A., Omueti, K.O., Stepensky, V.A., Yegin, O., Alpsoy, E., Hamann, L., Schumann, R.R., and Tapping, R.I. (2007). Cutting edge: A common polymorphism impairs cell surface trafficking and functional responses of TLR1 but protects against leprosy. J. Immunol. *178*, 7520–7524.
97. Misch, E.A., Macdonald, M., Ranjit, C., Sapkota, B.R., Wells, R.D., Siddiqui, M.R., Kaplan, G., and Hawn, T.R. (2008). Human TLR1 deficiency is associated with impaired mycobacterial signaling and protection from leprosy reversal reaction. PLoS Negl. Trop. Dis. *2*, e231.
98. Castellano, S., Parra, G., Sánchez-Quinto, F.A., Racimo, F., Kuhlwilm, M., Kircher, M., Sawyer, S., Fu, Q., Heinze, A., Nickel, B., et al. (2014). Patterns of coding variation in the complete exomes of three Neandertals. Proc. Natl. Acad. Sci. USA *111*, 6666–6671.