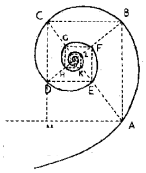
UNIVERSITÀ DEGLI STUDI DI MILANO



SCUOLA DI DOTTORATO IN MEDICINA MOLECOLARE

CICLO XXVII
Anno Accademico 2013/2014

TESI DI DOTTORATO DI RICERCA

MED26

HOW NATURAL SELECTION SHAPED DIVERSITY AT IMMUNE RESPONSE GENES AND AUTOIMMUNE RISK ALLELES DURING MAMMALIAN EVOLUTION

Dottorando : Diego Forni
Matricola N° R09723

TUTORE : Prof. Giacomo P. Comi

COORDINATORE DEL DOTTORATO: Prof. Mario Clerici

SOMMARIO

INTRODUZIONE

La diversità genetica è generata da una combinazione di differenti processi evolutivi, che includono principalmente mutazioni, deriva genetica, migrazioni e selezione naturale.
E' noto che la selezione naturale influenza uno specifico locus\variante, mentre gli effetti demografici hanno effetti su tutti i loci allo stesso modo; inoltre ci si aspetta che la selezione sia focalizzata su posizioni genomiche che hanno un ruolo funzionale. Le varianti target di selezione possono inoltre non avere solamente un ruolo funzionale, ma spesso possono anche correlare con predisposizione o protezione contro specifiche malattie; possono quindi essere studiate in screening per studi di associazione a malattie o infezioni: infatti varianti genetiche che sono vantaggiose tendono a crescere in frequenza in una popolazione, mentre mutazioni deleterie vengono eliminate.
Due differenti approcci vengono applicati per identificare l'azione della selezione naturale: un approccio inter-specie e un approccio intra-specie. Il confronto tra geni ortologhi permette di valutare l'azione della selezione naturale lungo tempi evoluzionistici lunghi; dall'altra parte l'analisi della variabilità genetica nelle popolazioni umane può mettere in evidenza processi adattativi più recenti.
I geni coinvolti nella risposta immunitaria sono tra i più studiati dal punto di vista evolutivo: è infatti stabilito che le infezioni hanno esercitato la pressione selettiva più forte nell'uomo e probabilmente anche nelle altre specie viventi. Le interazioni tra ospite e patogeno hanno modulato la diversità genetica nel corso del tempo da entrambi i lati: questa continua corsa alle armi tra ospite e patogeno crea una competizione tra geni che si adattano e contro adattano uno contro l'altro.
E' perciò importante valutare il livello di variabilità genetica che determina tratti fenotipici vantaggiosi, per identificare quelle regioni genomiche che sono responsabili della diversità e dell'adattamento.
Il mio primo studio è focalizzato sulle molecole coinvolte nella regolazione dell'attivazione delle cellule T. L'attivazione dei linfociti T è un fenomeno complesso che è mediato dall'interazione di diverse proteine espresse sulla superficie dei linfociti T e delle cellule che presentano l'antigene. Differenti patogeni hanno sviluppato strategie specifiche che hanno come target queste molecole; questi geni sono stati quindi ingaggiati in un conflitto costante con un vasto numero di patogeni e giocano un ruolo importante durante le infezioni; inoltre la variabilità genetica in questi loci può avere un potenziale impatto nello sviluppo di condizioni infiammatorie o autoimmuni.
Un secondo studio ha riguardato le molecole coinvolte nel pathway di processamento e presentazione dell'antigene. Qualunque sia la natura della molecola che deve essere presentata sulla superficie delle cellule che presentano l'antigene, la dimensione limitata della tasca rende impossibile la presenza di macromolecole: solamente i frammenti che derivano dalla loro lisi possono essere presentati. Questo repertorio antigenico è generato dal pathway di processamento e presentazione dell'antigene.

Un ulteriore studio ha riguardato il sistema di contatto e le molecole coinvolte in esso. In particolare questo pathway rappresenta un collegamento tra la cascata di coagulazione e la risposta immunitaria, due sistemi centrali nella sopravvivenza dell'ospite in presenza di danni ai tessuti e infezioni.
Infine, gli ultimi casi studiati hanno riguardato molecole che sono deputate a riconoscere gli acidi nucleici e ad attivare la risposta immunitaria contro virus e batteri, quali le molecole RIG-I-like e le molecole AIM-2 like.

SCOPO DEL LAVORO

Lo scopo di tutti questi studi è stato investigare la storia evolutiva dei geni coinvolti nei differenti patwhay, sia a livello intra-specifico che inter-specifico, e di estrarre queste informazioni per fornire nuovi sviluppi sul ruolo funzionale che queste molecole potrebbero avere nelle malattie che colpiscono l'uomo. Inoltre ho voluto valutare la storia di alcuni alleli di rischio per malattie autoimmuni e di come si sono diffuse nelle popolazioni umane.

MATERIALI E METODI

Le sequenze codificanti dei geni analizzati per le diverse specie sono state ricavate dal database Ensembl e allineate usando il tool The RevTrans 2.0. Per la ricerca di selezione positiva ho utilizzato il software PAML con differenti modelli evolutivi; siti sotto selezione sono stati identificati usando i metodi Bayes empirical Bayes e Mixed Effects Model of Evolution.
I dati dei genotipi di differenti popolazioni umane dei geni di interesse e di altri 1200 geni casuali di controllo sono stati ricavati dal database 1000 Genomi; in particolare sono state analizzate tre popolazioni: Europea, Africana e Est Asiatica. Questi dati sono stati utilizzati per calcolare diversi parametri di diversità nucleotidica, cosi come alcune statistiche basate sullo spettro di frequenza dei siti. I dati dei geni di controllo sono stati usati per calcolare distribuzioni empiriche di questi parametri.
F_{ST}, una misura di differenziazione genetica tra popolazioni, e il DIND test, un test basato sull'omozigosità degli aplotipi, sono stati calcolati per tutti gli SNPs analizzati.
La significatività statistica di questi test è stata ottenuta da distribuzioni empiriche di questi valori calcolati su tutte le varianti localizzate nei geni di controllo.

RISULTATI E DISCUSSIONE

Nella mie analisi ho trovato che i geni coinvolti nella regolazione dell'attivazione delle cellule T hanno rappresentato dei target di selezione sia durante l'evoluzione dei mammiferi che durante la recente storia evolutiva delle popolazioni umane; ho inoltre evidenziato che la varianti di questi geni che sono correlate a malattie nell'uomo sono dei target preferenziali delle selezione effettuata dai patogeni.

Questi risultati hanno mostrato che un allele può diffondersi in una popolazione perchè conferisce maggiore protezione contro alcuni agenti infettivi, indicando che l'adattamento alle infezioni è una possibile spiegazione per il mantenimento di alcuni alleli di rischio per malattie autoimmuni. Questi risultati inoltre supportano l'idea che l'adattamento dell'uomo verso un ambiente con un ridotta presenza di patogeni ha determinato la diffusione di alcuni alleli di rischio per malattie autoimmuni o infiammatorie.

L'analisi degli eventi selettivi per le molecole coinvolte nel pathway di processamento e presentazione dell'antigene ha rivelato la presenza di un numero elevato di geni sottoposti a selezione positiva durante la storia evolutiva dei mammiferi.

I dati hanno inoltre evidenziato la presenza di una continua pressione selettiva che agisce su differenti scale di tempo evolutive per alcuni di questi geni,e anche messo in evidenza come la maggior parte degli eventi selettivi che riguardano la specie umana sono localizzati in regioni regolatorie e possono avere un ruolo nella modificazione dei tratti fenotipici umani. L'analisi evolutiva dei geni coinvolti nel sistema del contatto ha indicato che KNG dei mammiferi è stato target di una lunga e forte pressione selettiva. In particolare i risultati supportano la teoria che KNG possa avere un ruolo centrale nel modulare la risposta immunitaria e mostrano come esso sia un target di differenti specie di patogeni. Per concludere, lo studio di due differenti famiglie di recettori di acidi nucleici ha mostrato che una porzione di questi geni sono coinvolti nel conflitto ospite patogeno, che porta allo scenario di corsa alle armi tra le due specie, e i risultati hanno evidenziato informazioni funzionali riguardo a specifiche varianti che possono influenzare specifici fenotipi.

CONCLUSIONI

I risultati di tutti i vari studi hanno mostrato come la selezione naturale abbiamo agito sulla variabilità in differenti processi coinvolti nella risposta immunitaria e siti target di selezione riguardano posizioni di fondamentale importanza per la funzione proteica.

Questi nuovi dati hanno generato nuove ipotesi che possono essere testate sperimentalmente, riguardanti il ruolo di siti specifici o regioni che modulano fenotipi negli esseri umani; questi risultati suggeriscono anche che è necessaria cautela nell'estrarre i risultati di uno specifico esperimento in organismi modello, perché una parte considerevole della diversità genetica in queste molecole non è dovuta a processi neutrali ma in risposta ad eventi adattativi.

ABSTRACT

INTRODUCTION

Genetic diversity is generated by a combination of different evolutionary processes, including mutation, genetic drift, migration, and natural selection.It is well known that natural selection acts on a specific locus\variant, whereas demographic effects act on all loci in the same way; also the selection is expected to be focused on genomic positions that have a functional role. Importantly, the selected variants targeted by selection may not only have a functional role but can correlate with predisposition or protection to some specific diseases. They can therefore be prioritized in screenings for association with diseases and infections; indeed, genetic variants that are advantageous tend to increase in frequency in the population, while deleterious mutations tend to be eliminated. To identify selection, intra and inter species approaches are usually applied; comparing orthologous genes among different species is a successful approach to detect positive selection acting over long evolutionary timescales; on the other hand, comparing genetic variation within human populations may underline more recent adaptive events. Genes related to immune system are among the most studied genes from an evolutionary point of view: it is now established that infections have been acting as a major selective pressure on humans and, most likely, on all living organisms. Thus, the interactions between hosts and pathogens have shaped the genetic diversity over time on both sides: moreover the continuous arm race between hosts and pathogens creates a competition of co-evolving genes that develop adaptations and counter-adaptations against each other. Therefore, it is important to evaluate the level of genetic variation that determines advantageous phenotypic traits to identify genomic regions/positions underlying diversity and adaptation. My first study was focused on molecules involved in the regulation of T-cell activation. The activation of T lymphocytes is a complex phenomenon that is mediated by the interaction of a number of proteins expressed on the surface of T lymphocytes and antigen presenting cells (APC).Several pathogens have evolved strategies that specifically target these genes to either invade the host or to reduce the response of the immune system. Thus, on one hand,these genes have been engaged in a constant conflict with a large number of pathogens and play a fundamental role during infections; on the other hand, genetic variation at these loci has a potential impact on the development of autoimmune and inflammatory conditions. In the second study I focused on molecules involved in the antigen processing and presentation pathway (APP).Whatever the nature of the presenting molecule, the limited dimension of its cleft makes it impossible for macro molecules to be presented: only fragments (antigens) derived from the lysis of such molecules can be nested in the cleft. This antigenic repertoire is generated by the antigen processing and presentation pathway. Another study focused on the contact system and the molecules involved in this pathway. In particular this pathway represents a link between the coagulation and inflammatory responses, two systems central to host survival in the face of tissue damage and infection.

Finally, the last molecules analyzed were the proteins responsible for nucleic acids

recognition and the activation of the immune response against virus and bacteria, as the RIG-I like proteins and the AIM-2 like proteins

AIM OF THE WORK

The aim of all the studies was to investigate the evolutionary history of the genes involved in different pathways, both at the inter- and intra- specific level, and to exploit this information to provide novel insight into the functional role of these molecules in human health and disease. Also i wanted to evaluate the history of autoimmune risk alleles and how they spread in human populations.

MATERIALS AND METHODS

Mammalian coding sequences were retrieved from the Ensembl website and aligned using the The RevTrans 2.0 utility. For detecting the action of positive selection I used the PAML software with different evolutionary models. Sites under selection were identified using Bayes empirical Bayes and Mixed Effects Model of Evolution analyses.Genotype Data from the Pilot 1 phase of the 1000 Genomes Project were retrieved from the dedicated website for the genes analyzed and for 1,200 randomly selected RefSeq genes (control set) for three populations with different ancestry: European, African and East Asian. These data were used to calculate nucleotide diversity parameters as well as some site frequency spectrum-based statistics. Data from the control gene set were used to calculate empirical distributions of these parameters.
F_{ST}, a measure of population genetic differentiation, and the DIND test, a test based on haplotype homozygosity, were calculated for all SNPs analyzed. I calculated statistical significance of these tests by obtaining an empirical distribution for variants located within the control genes.

RESULTS AND DISCUSSION

In my analyses I found that genes involved in the regulation of T cell activation have represented selection targets both along mammalian evolution and during the history of human populations; I also found that variants in these genes related to human diseases to be preferential targets of pathogen-driven selection. These results showed that an allele can spread in a population because it confers higher protection against some infectious agent, indicating adaptation to infection as the underlying explanation for the maintenance of a set of autoimmune risk alleles. These result has a relevance for the hygiene hypothesis, and support the idea that human adaptation to an environment with reduced presence of pathogens has determined the spread of some risk alleles for autoimmune and inflammatory diseases.

We then presented a comprehensive analysis of the selective events acting on the antigen processing and presentation pathway across different evolutionary timescales, revealing a high proportion of genes under positive selection in mammalian species.

Data also indicate a continuum in selective pressure acting on different timescales for some of these genes analyzed, and we also demonstrated that the selected

variants in human populations were always located within regions with regulatory function and can have a role in modulating human phenotypes.

The evolutionary analysis we performed about contact system genes indicated that mammalian kininogen has been a target of long-lasting and strong selective pressures. In particular our results reinforced the possibility that kininogen plays a central role in the modulation of immune response and is a target of different pathogen species. Finally our study of two different families of nucleic acid receptors showed that a proportion of these genes have been engaged in host-virus genetic conflict leading to a continuous host–pathogen arms race scenario, and again our results provide functional information about variants that might affect immunologic phenotypes.

CONCLUSIONS

Results in all these studies showed how natural selection shaped diversity in different pathway involved in the immune response, and selected sites involve positions of fundamental importance to the protein function. These novel data give rise to a number of experimentally testable hypothesis concerning the role of specific sites or regions as modulators of immunological phenotypes; they also suggest caution when extrapolating results from specific experiments in model organisms, as a considerable portion of genetic diversity in these molecules has accumulated not as a result of neutral processes but in response to adaptive events.

SYMBOL LIST

SNP: single nucleotide polymorphism
MAF: minor allele frequency
DAF: derived allele frequency
θ_W: an estimate of the expected per site heterozigosity
π: the average number of pairwise sequence nucleotide differences among haplotypes
D_T or D: Tajima's D
D* and F*: Fu and Li's D* and F*
HKA test: Hdson-Kreitman-Aguadè test
MLHKA: maximum likelihood HKA test
F_{ST}: fixation index
DIND: Derived Intra-allelic Nucleotide Diversity
iHS: integrated haplotype score
EHH: extended haplotype homozygosity
τ: Kendall's rank correlation coefficient
dN: the observed number of nonsynonymous substitutions per nonsynonymous site
dS: the observed number of synonymous substitutions per synonymous site
ω: dN/dS
LD: linkage disequilibrium
LRT: likelihood ratio tests
XP-EHH: cross-population extended haplotype heterozygosity test

INDEX

1 INTRODUCTION

1.1 Genetic variability in human populations

Genetic diversity is the result of many events acting during the evolutionary history of a species; adaptive and demographic events like migrations, genetic drift, mutation , and selection are the main factors responsible for genetic variability during different time-scales.

Different processes, like genetic drift and demography, can modify frequencies of different alleles from one generation to another by chance alone; this means that particular alleles can be maintained in a population by chance. This variability is also affecting human populations, and in particular different studies have demonstrated that most of the differentiation that we find in our species results from migration events and is consistent with population expansions and bottlenecks [1].

It is known that modern humans appeared in Africa about 200,000 years ago and approximately 60,000 years ago they started to spread out, colonizing Europe and East Asia [2]. A study by [1] indicated that there is a strong correlation between geographic distance between populations and genetic variability within a population; indeed populations closer to Africa show more variability than populations further from Africa, reflecting the route of migrations of ancient humans.

The allele pool of populations reflects this geographic scenario, in fact most of the alleles\haplotypes of Non-African populations are a subset of the African ones. This observation is crucial in order to identify genetic adaptation, because it must be considered when evaluating whether a specific variant is influenced by the action of natural selection. Demographic events influence variability in all genes in the same way, whereas selection specifically targets defined regions; so deviation from the general behavior of genetic variation can be an indication of the action of

natural selection. It should also be considered that, during their evolutionary history, humans have adapted to the environment, and this adaptation is driven by different forces, with pathogens being among the most important one [3].

1.2 Pathogens and natural selection

Infections caused by different types of pathogens are recognized as one of the major selective pressures for humans and, possibly, for all living organisms [3]. Interactions of pathogens with hosts result in a situation that is defined as arms race, in which there is a continuous selective pressure on hosts to generate resistance against pathogens, and at the same time pathogens try to develop new strategies to evade host defenses for a successful infection. The result of this arms race scenario is a great shape in the genetic diversity both in the hosts and in the pathogens genomes, determining fast evolutionary rates. This constant genotype turnover in the interacting species is generally referred to as a “Red Queen” scenario, from the character in Lewis Carroll’s novel who says: “It takes all the running you can do, to keep in the same place”.

This hypothesis has been supported by the description of rapid rates of evolution in genes involved in host-pathogen conflicts and, more recently, the development of experimental evolution approaches has allowed its formal testing [4, 5].

Linked to this evolutionary scenario, another hypothesis has been formulated: the so called “ hygiene hypothesis” [6]. This hypothesis describes a situation in which big changes in the environment and more hygienic conditions have reduced the exposure to antigens, generating an unbalanced immune response. This unbalanced situation can lead to the development and widespread occurrence of chronic inflammatory conditions.

Given these two premises, it is obvious to expect that genes involved in immune response commonly targeted by natural selection; in fact, different studies [7-9] have demonstrated that immune-related genes are preferential targets of natural selection, reflecting the dynamic nature of this system.
The recent availability of large-scale genomic data can be very helpful to study how genetic variation is distributed among different human populations; moreover, inter-species comparative analyses can be also provide pivotal information on the genetic and immunologic determinants underlying pathogen-driven selective scenarios. Inter-species evolutionary guided approaches can shed light into into host-pathogen interactions, and delineate the basis of host range and disease appearance.

1.3 Signatures of natural selection (intra-species level)

Most changes that affect genome variability are a consequence of random genetic drift and demography rather than adaptive evolution; so distinguishing the action of selection and demographic events is very important, albeit difficult.
In principle, natural selection acts on variants in the genome, influencing their segregation in different environmental conditions and without differences whether the mutation is an old or a new one. Variants in the genome usually do not segregate independently due to linkage disequilibrium (LD), the non random association of alleles at two or more loci. Natural selection therefore influences not only the selection target, but also its surrounding variants and the expected result is not only a change in the frequency of the selected allele, but also in all variants linked to it.
There are different ways in which natural selection can act. Purifying (or negative) selection tends to eliminate new deleterious mutations or keeps them at low frequency. Positive selection is the situation in which a new

mutation rises in frequency in a population due to its selective advantage; and in case of balancing selection the variability within a population is maintained, usually due to an advantage of the heterozygous individuals.

Under neutral evolution the only factors that influence genetic variability are mutation rate population size and demographic events. Different types of selection act in different ways on the frequencies of alleles: for example, when a mutation is favoured by selection in a population, it will rise to high frequencies, and also all variants that are neutrally evolving and are in LD with it will rise to high frequencies.

This situation is described as a selective sweep and results in an overall decrease of genetic variation at the selected site, as well as at the surrounding region (Fig. 1.1).

Over time, new mutations will arise, but will initially be present at low frequencies. In case of selection, these derived alleles (new alleles that are not present in the closest related species) linked to the selected variant will also rise in frequency (Fig. 1.1). Therefore, another signature of a selective sweep, together with low levels of genetic variation, is an excess of high-frequency derived alleles [10].

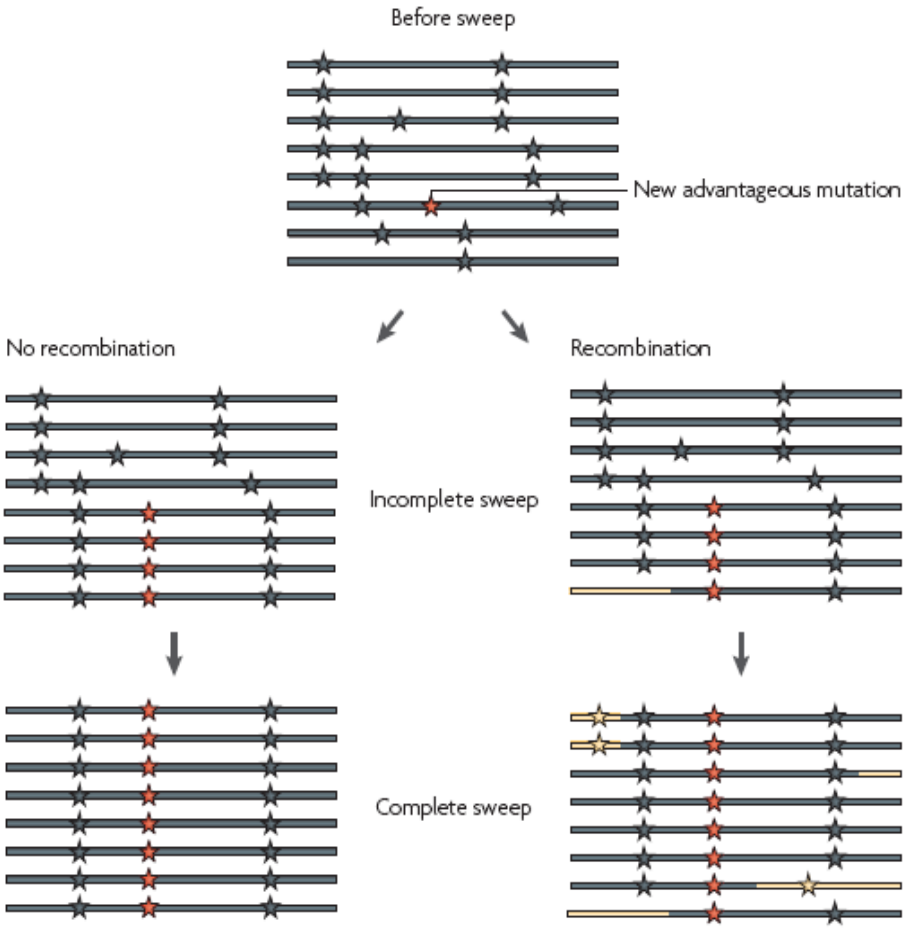


Figure 1.1. Schematic representation of a selective sweep. Horizontal lines and stars indicate haplotypes
and derived allele of SNPs, respectively; red colour indicates a new advantageous mutation.Chromosomal segments that are in linkage disequilibrium with the advantageous mutation are coloured yellow. Figure taken from [10]

Another selective scenario that can cause an increase of low-frequency variants is purifying selection. New mutations with a deleterious effect are generally maintained at low frequencies and as a consequence the frequency of derived alleles is expected to be low.

Finally, Balancing selection increases the proportion of variants at intermediate frequencies, and it tends to maintain variation at one or more sites. Moreover, variants that are linked to the selection targets are maintained together with the selected alleles, and the result is an excess of nucleotide diversity in the region (Fig.1.2) [11]. It should be noticed that opposite to a selective sweep, where the region influenced by selection is usually large, in the case of balancing selection the target tends to be small, especially when the event is a long-standing.

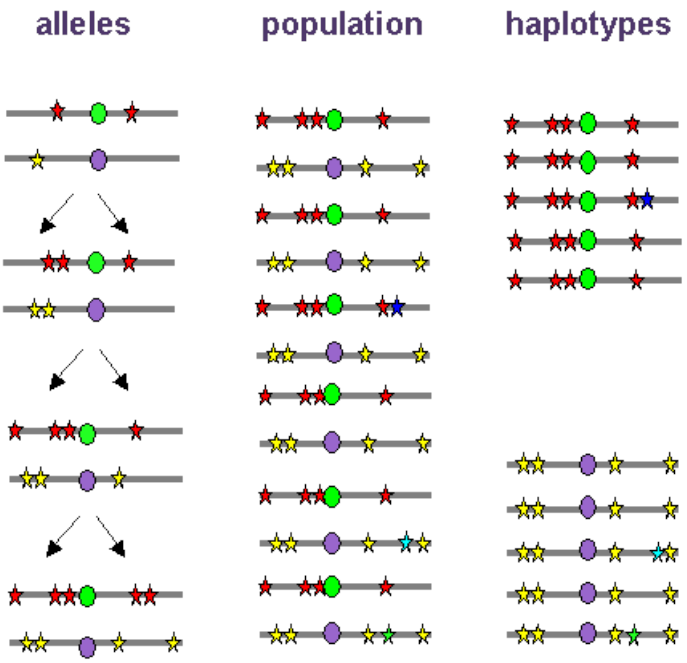


Figure 1.2. Schematic representation of the action of balancing selection. Horizontal lines indicate haplotypes; stars represent neutral SNP alleles. The two alleles of a SNP under balancing selection are shown with violet and green circles. Neutral SNP alleles and new mutations arising over time are maintained together with the two selected alleles.

The most commonly applied model of evolution to distinguish neutral variation from variation targeted by natural selection is the neutral model [12], that makes relation between the rate of mutation and evolutionary parameters.

This model assumes that most of the variation within a species is not due to natural selection but is caused by random drift of neutral alleles; specifically, polymorphisms within a species and the frequency of alleles are related to the mutation rate and population size.

This model also specifies the relationship between two different species; in particular the rate at which differences accumulate as two species diverge is the same as the rate at which neutral mutations rise in frequency in each species.

Given this premise, many statistical tests have been designed to detect natural selection using this model of evolution; these and are defined as tests of the null hypothesis of neutrality, or neutrality tests.

1.4 Neutrality tests

1.4.1. Allele frequency spectrum statistics

Numerous population genetics tests have been developed to identify and distinguish the action of different types of selections looking at the spectrum of allele frequencies in one or more populations. The most widely used test was proposed by Tajima [13] and it is based on the comparison of two parameters: θ_w*, an estimate of the expected heterozigosity per site* [14]*, and* π [15]*, the average number of pairwise sequence nucleotide differences. D is the standardized difference between* π *and* θ_w*; under neutral selection these estimates are expected to be equal, so the value for D under neutral*

evolution is 0. Since π depends on the frequency of alleles, D will be positive in the presence of many intermediate alleles and negative when an excess of rare alleles is observed. In the case of selective sweep, for example, the presence of a large proportion of low-frequency variants generate negative D values. The same result of low values can occur under purifying selection.

Conversely, balancing selection causes high positive values of this statistic due to the presence of different alleles in the population with intermediate frequency.

Fay and Wu [16] proposed a test to distinguish among negative values of Tajima's D generated by the two different selection scenarios. In fact the action of positive selection can drive derived alleles in the affected region to high frequencies. So they proposed a test that compares two estimates: π and a measure of diversity that takes into account whether the selected allele is the derived one (θ_H). In the case of a recent selective sweep the excess of high-frequency derived mutations results in a negative value for H [16].

Fu and Li [17] also developed neutrality tests based on the allele frequency spectrum. These tests are conceptually similar to Tajima's D, but also include information about the genealogy of the haplotypes. Mutations are classified as external and internal: the former are the ones which occurred on the external branches and the latter are mutations that occurred on the internal branches of the genealogy. In case of natural selection the number of external mutations is likely to deviate from its neutral expectation, while the number of internal mutations is less affected.

All these tests can be biased by the fact that demographic events can also change the site frequency spectrum. For example, a population bottleneck is expected to result in an increase of intermediate frequency alleles, while population expansion results in a higher proportion of low frequency variants.

As said above, different human populations are known to have experienced diverse demographic scenarios (Fig 1.3), so this information must be considered, when it is necessary to establish statistical significance of these tests. two common approaches are applied. The first is based on simulations of different scenarios using a coalescent approach.

The coalescent is a process in which going backwards in time the genealogies of two alleles merge at a common ancestor; the scheme is to consider the ancestral history of a gene/ region by modelling time intervals between each coalescent event going back in time.

Coalescent simulations are very common approaches to simulate neutral genealogies and are useful in making predictions regarding the evolutionary history of a gene/region. Therefore, a large number of coalescent scenarios can be simulated and the statistics of interest calculated for each of the simulations. The distribution of values for the statistics obtained through the simulation process can give information about the statistical significance of the values of the region analysed.

With this approach it is also possible to integrate coalescent information with demographic scenarios and to simulate different demographic events for different populations.

*T*he second approach is based on the empirical comparison with data calculated for a large number of genomic regions analysed in the same population as the locus of interest. The idea is that demography affects all loci equally, while selection is locus-specific. Therefore, the test is calculated for the gene/ region analysed and for a set of random regions; significance is evaluated by comparing the distribution of the values of the test in that population.

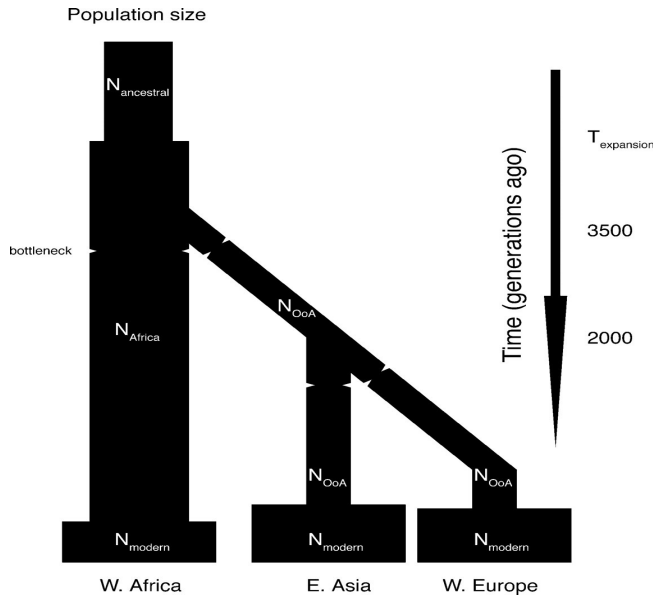


Figure 1.3. Demographic model for three major human populations. $N_{ancestral}$: ancestral population size; N_{Africa}: African population size; N_{OoA}: non-African population size; $T_{expansion}$: Time of ancestral population expansion. Bottlenecks are indicated by constrictions. Figure taken from [18]

1.4.2. Tests based on differences among populations

One of the most widely used approaches for detecting natural selection is to study differences in allele frequencies between populations. Levels of genetic differentiation among populations can be modified by the action of positive selection; adaptations to the local environment or mutations that arise in specific geographical regions might be the cause of these changes. Conversely, balancing selection may result in different scenarios. In the case of two populations being more similar than expected, the explanation is that the selected allele and all linked variants tend to be maintained at detectable levels in different populations. However, if balancing selection is restricted to one specific geographic location then it can increase genetic

differentiation between or among populations.

The most commonly used statistical measure of population differentiation was developed by Wright [19] and is known as the fixation index (F_{ST}). The main problem in using F_{ST} for the detection of natural selection is that differentiation among populations is influenced by different demographic events, for example the amount of gene flow and the rate of genetic drift, and these factors make it difficult to discard demographic scenarios that could account for the observed high or low F_{ST} values. One possibility is to use information of population differentiation from a large number of genetic loci to create an empirical F_{ST} distribution. Another aspect to be taken into account is that F_{ST} is not independent on the allele frequency, so when evaluating the test, the frequency of the analysed variantsshould be considered.

1.4.3. Extended haplotype tests

Another approach for detecting recent selective sweeps is based on the analysis of haplotypes and their extension along a genomic region [20, 21]. These tests, called extended haplotype tests, exploit linkage disequilibrium information of local genomic regions; they are based on the fact that under positive selection an advantageous mutation rises quickly in frequency. If the frequency increase of the selected allele occurs faster than the rate of recombination, an extended region containing all the variants linked to the selection target, will also rise in frequency; this phenomenon originates an extended region of LD. Therefore, a haplotype at high frequency with high homozygosity that extends over large regions is a sign of an incomplete selective sweep.

The most common statistic used for evaluating the presence of large homogeneous genomic regions is the iHS (integrated haplotype score) test [20]; it compares the homozygosity of a specific haplotype against all other haplotypes present in that region.

A more recently developed test for detecting selective sweeps is the DIND (Derived Intra-allelic Nucleotide Diversity*)* test [22]. It is also based on the homozygosity of haplotypes, but it distinguishes between haplotypes that carry the derived allele and the ancestral allele of the candidate variant. The main idea is that if a new variant is selected, it increases in frequency in the population with the same result as seen before, but this will not happen for the ancestral allele. The DIND test compares the homozygosity of haplotypes carrying the ancestral allele against the homozygosity of haplotypes carrying the derived allele of that specific variant. High value of this test together with high frequency of the derived allele is usually an indication of positive selection.

The cross-population extended haplotype heterozygosity test (XP-EHH) [21, 23] is another approach to identifying signals of positive selection by comparing the lengths of haplotypes from two populations. This method identifies selective sweeps if the selected allele has approached or achieved fixation in one population but remains polymorphic in the general human population [21].

1.4.4. Variation and divergence based tests

A different approach to search for selection is to use information about intra-species variation and inter-species divergence. Under neutrality, these two parameters are expected to be proportional to the neutral mutation rate; so, one possibility to detect natural selection is to test whether this expectation is verified. The HKA (Hudson-Kreitman-Aguadè) test [24] is based on an assumption that under neutrality the ratio of polymorphism to divergence is the same for at least two genes analyzed. A gene of interest is compared to a putatively neutral locus, and differences in the ratio of polymorphism to divergence between these are taken as evidence of selection.

1.4.5. Environmental variables as selective pressure

Another possibility to identify the action of natural selection on a specific gene/region is to search for variants that show differentiated allele frequencies among populations that live in different environments, and to evaluate whether a correlation exists between environmental variables and genetic information.. This type of differentiation can be generated if selective pressures exerted by the environmental variables are different among geographic locations. Caution is needed when using this approach due to the fact that the major contribution of genetic diversity across geographic locations is accounted for by demographic event. Therefore, it is important to disentangle neutral events from selective ones.

1.5 Detecting natural selection at the inter-species level

Genetic comparisons among different species can highlight selective events that have been ongoing over long time scales. These approaches aim to evaluate evolutionary patterns using extant genetic diversity and phylogenetic relationships among species. The basis is to compare orthologous coding sequences, and to analyze at each site of the sequence whether all possible substitutions would be nonsynonymous (changing the amino acid) or synonymous (not changing the amino acid). The number of nonsynonymous differences per nonsynonymous site (dN) and the number of synonymous differences per synonymous site (dS) are then calculated.
In case of neutral evolution, the rate at which nonsynonymous changes accumulate in a protein is expected to be comparable to the rate of synonymous changes, so dN/dS (ω) will be equal to 1. In mammalian species, most amino acid replacements are deleterious and tend to be eliminated by purifying selection; this occurs at the majority of sites and, consequently, for the majority of genes dN/dS is lower than 1. Conversely, a specific selective pressure may favor amino acid replacements (positive

or diversifying selection): in this case dN/dS will be higher than 1.

Often positive selection is focused on few sites with specific roles in the function or stability in an otherwise selectively constrained protein. Also, the selective pressure might act on a limited number of lineages in a phylogeny, and not with the same strength on all lineages, and in this case is defined as episodic selection.

Thus, evolutionary analyses use different methods to detect genes, sites or lineages that show evidence of positive selection. Also most of these methods depend on the number of species being analyzed and on their phylogenetic relatedness,. Also sequence errors and poor alignment quality ,as well as the unrecognized action of recombination, can lead to an overestimation of the action of positive selection [25, 26],. These factors must therefore be accounted for.

1.5.1. Selection at a specific site

The most widely used models to detect selection at specific sites are the site models implemented in the PAML (Phylogenetic Analysis by Maximum Likelihood) package [27, 28]. They are used to infer positive selection and to identify the sites under selection. These models allow ω to vary from site to site, assuming a constant rate at synonymous sites.

The analysis is based on fitting the data from a multi-species alignment and from a phylogenetic tree to different types of models that either allow (consider positive selection) or do not allow (consider only neutral evolution) a class of codons to evolve with ω>1.

To determine whether the neutral model can be rejected in favor of the positive selection model likelihood ratio tests (LRT) are then used. Also, if the null hypothesis of neutral selection is rejected in favor of a model that allows the action on positive selection on that gene, a Bayes empirical Bayes approach (BEB) can be used to detect specific sites targeted by

selection [29]. In particular, the BEB analysis calculates the posterior probability that each site of the alignment belongs to the class of codons that is positively selected.

This approach assumes that natural selection is acting with the same strength and in the same direction for all lineages analyzed in the tree. As this is often not what happens, a different approach can be applied: the mixed effects model of evolution (MEME) [30] allows the distribution of ω to vary from site to site and from branch to branch and, therefore, has greater power to detect episodic selection.

1.5.2. Selection of lineages under positive selection

Another possibility when searching for positive selection at the inter-species level is to evaluate how different lineages in a phylogeny are evolving. The PAML software implements methods to detect selection along specific branches through the so called branch-site models [31, 32]. These require that different branches of the phylogeny are dived into two classes: foreground and background branches. A likelihood ratio test is again applied to compare a model that allows positive selection on a class of codons only for the foreground branches with a model that does not allow such selection [32].

The main problem in this type of approach is how to define which lineages are foreground branches: the best solution is to have a priori information based on biological evidence about the genes analyzed. If this information is not known, it is necessary to consider each branch of the tree as foreground.

There are other methods that evaluate the presence of positive selection at branch level and do not require information about the evolutionary history of genes analyzed. For example, the branch site-random effects likelihood (BS-REL) method simulates three different evolutionary scenarios (neutral,

purifying and positive selection) for all branches in the phylogenetic tree, and each branch is considered independently from the others; in this way the algorithm can identify branches that are influenced by positive selection [33].

1.6 Aim of the thesis

The aim of the work was to investigate the evolutionary history of candidate genes involved in different pathways with a central role in the immune response. I analyzed at the most important molecules that modulate T cell activation, proteins responsible for the generation of the antigenic repertoire presented by MHC molecules, and molecules involved in coagulation pathway.

I also analyzed the evolutionary patterns of different nucleic acid receptors, as the AIM2-like receptors and the RIG-I-like receptors.

The analyses were carried out both at the inter- and intra-specific level, with the aim to provide novel insights into the functional roles of these molecules in human health and disease. The final goal was to identify variants to be prioritized in screenings for association with autoimmune diseases and infections.

2 METHODS

2.1 Inter-specific analyses

Inter-species analyses were performed using the coding sequences of the analyzed genes; in particular, mammalian sequences for all genes of interest were retrieved from the Ensembl and NCBI databases (http://www.ensembl.org/index.html; http://www.ncbi.nlm.nih.gov/); the list of species analyzed for each gene varies depending on availability and other factors (e.g. gene lossees, reliability of orthology).

DNA alignments of orthologous sequences were performed using the RevTrans 2.0 utility [34], which uses the protein sequence alignment as a scaffold for constructing the corresponding DNA multiple alignment. One of the most common problems in these types of analyses is reconstructing the correct alignment and using the correct sequences: errors in this two steps can generate false positive results due to the incorrect evaluation of the number of nonsynonymous and synonymous substitutions. Therefore, all the alignments were generated using a software that maintains the codon reading frame and were also manually checked and edited to remove uncertainties.

Another element to consider is that variability generated by recombination can be mistaken as positive selection [25]. Thus, alignments should be screened for recombination before running positive selection tests. The Genetic Algorithm for Recombination Detection (GARD) [35] from HYPHY package [36] was used to screen the alignments for recombination; GARD uses phylogenetic incongruence among fragments of a sequence alignment to detect recombination events; the application of a genetic algorithm allows searching for multiple breakpoints and the probability that each breakpoint is due to recombination is assessed through Kishino-Hasegawa tests [37].

To evaluate the dN/dS (ω) ratio along gene alignments the Single Likelihood Ancestor Counting (SLAC) method [38] present in the HYPHY package was used.

The site models implemented in PAML [28] can detect positive selection, in particular if this selective pressure is affecting only few sites in a gene. These models consider the dN/dS (ω) ratio for any codon in the gene as a random variable from a statistical distribution, thus allowing ω to vary from site to site, assuming a constant rate at synonymous sites. To detect selection, site models that allow (M2a, M8) or disallow (M1a, M7) a class of sites to evolve with ω >1 were fitted to the data using two different codon frequency model: the F3x4 and the F61 model, which weight in different ways the frequency of each nucleotide in the data analyzed.

The nested models (M1a vs M2 and M7 vs M8) are compared through likelihood-ratio tests (degree of freedom= 2) to asses statistical significance.

Positively selected sites were identified using the Bayes Empirical Bayes (BEB) analysis (with a cut-off of 0.90), which calculates the posterior probability that each codon is from the site class of positive selection (under model M8) [39].

For the identification of specific positively selected sites the Mixed Effects Model of Evolution (MEME) from HYPHY (with the default cutoff of 0.1) [30] was also applied. MEME allows the distribution of ω to vary from site to site and from branch to branch at a site, therefore allowing the detection of both pervasive and episodic positive selection.

Positive selection can act on all lineages of the tree, but also on specific branches; to explore possible variations in selective pressure among different lineages, other models from the PAML package, called the free-ratio models, were used. The M0 model assumes all branches to have the same ω, whereas M1 allows each branch to have its own ω [40]. The models are compared through likelihood-ratio tests (degree of freedom=

total number of branches -1). In order to identify specific branches with a proportion of sites evolving with ω>1, I used BS-REL [33]. This method implements branch-site models that simultaneously allow ω to vary across branches of the tree and sites within the alignment. BS-REL requires no prior knowledge about which lineages are of interest and uses sequential likelihood ratio tests to identify significant branches. Branches identified using this approach were cross-validated using the branch-site likelihood ratio tests from PAML (the so-called modified model A and model MA1, “test 2”) [32]. In this test, branches are divided a priori into foreground (those to be analyzed for positive selection) and background lineages, and a likelihood ratio test is applied to compare a model that allows positive selection on the foreground lineages with a model that does not allow such positive selection. A false discovery rate correction was applied to account for multiple hypothesis testing (i.e. I corrected for the number of tested lineages), as suggested [41].

Using the MA model has the advantage to identify specific sites positively selected on the foreground branches (although it has limited statistical power [32]), because it also implements a BEB analysis analogous to that described above to calculate the posterior probabilities that each site belongs to the site class of positive selection on those lineages.

2.2 Protein stability analyses

To evaluate the effect of different mutations on protein stability, analysis was carried out using three different methods: FoldX 3.0 [42], PoPMuSiC and I-Mutant 2.0 [43]. In FoldX and I-Mutant the ΔΔG values are calculated as follows: $\Delta\Delta G = \Delta G_{mutant} - \Delta G_{wild\text{-}type}$. In FoldX and I-Mutant ΔΔG values > 0 kcal/mol are an indication of substitutions that decrease protein stability, whereas in PoPMuSiC ΔΔG values > 0 kcal/mol are evidence of

substitutions increasing protein stability. So, to obtain comparable results, PoPMuSiC ΔΔG values were multiplied by -1.

In the analysis carried out with FoldX 3D, the three-dimensional structure of the protein was repaired using the <RepairPDB> command and substitutions were introduced using the <BuildModel> command with <numberOfRuns> set to 5 and <VdWdesign> set to 0. Temperature (298K), ionic strength (0.05 M) and pH (7) were set to default values and the force-field predicted the water molecules on the protein surface.

2.3 Docking and secondary structure analyses

To analyze the interaction between molecules a docking analysis was performed. In case of absence of known crystal protein structure in public databases, secondary structure prediction was performed using PSIPRED [44], and structure was obtained using QUARK [45], a server for *ab initio* protein folding and protein structure prediction. Variants were generated through MODELLER v9.11, using the *ab initio* prediction as template. RosettaDock [46] and ClusPro [47] were used for docking calculations.

Protein 3D structures were derived from the Protein Data Bank (PDB). Sites were mapped into structures using PyMol (The PyMOL Molecular Graphics System, Version 1.5.0.2 Schrödinger, LLC)

2.4 Population genetics-*phylogenetics analysis*

A different way to analyze the action of positive selection is to integrate divergence information with polymorphism data to detect fine-scale differences in selective pressure within a gene. Wilson et al [48] developed a population genetics-phylogenetics method that evaluates the distribution of a selection coefficient within a species and compares it with the

distribution generated by an inter-species analysis, and model the results of this genetic variation in a selection coefficient γ.

This method, known as gammaMap [48], assigns the selection coefficient γ to 12 different categories of selective effects, ranging from strongly beneficial to effectively inviable;

GammaMap allows to estimate the frequency of γ along all codons by applying a sliding windows approach (thus providing information about the whole gene region) and assigns posterior probabilities for each selection coefficient at each site, allowing a site specific analysis.

For the gammaMap analysis I assumed θ (neutral mutation rate per site), k (transitions/transversions ratio), and T (branch length) to vary among genes following log-normal distributions. For each gene I set the neutral frequencies of non-STOP codons (1/61) and the probability that adjacent codons share the same selection coefficient (p=0.02). For selection coefficients a uniform Dirichlet distribution with the same prior weight for each selection class was considered. For each gene I run 10,000 iterations with thinning interval of 10 iterations.

2.5 Genome projects

In the last few years the interest in analyzing genome data from a large number of human individuals of distinct ethnic groups has become widespread. Different project with different methodological approaches have addressed this issue.

2.5.1. HapMap Project

The International HapMap Project is a multi-country collaboration program involving centers from all over the world (www.hapmap.org). The aim of this project was to generate an haplotype map of the human genome, that

describes the patterns of similarity of human genetic variation in a great number of individuals of different ethnicity. The data produced by the project are genotypes of the 270 individual samples and the frequencies of 1,500,000 SNPs from four populations with African, Asian, and European ancestry.

2.5.2. Human Genome Diversity Project

The Human Genome Diversity Project (HGDP) is another international project that aims to understand the diversity of different human populations. Its goal is to collect genetic information from different population groups throughout the world and to create a database of human genetic diversity. The data generated are more than 650,000 SNPs in about 1,000 individuals from more than 50 distinct ethnic groups.

2.5.3. 1000 Genomes Project

The 1000 Genomes Project aimed to characterize human genome sequence variation with the final aim to investigate the relationship between genotypes and phenotypes.

The first phase of the project used different strategies of genome-wide sequencing with high-throughput platforms. This phase was based of low-coverage whole-genome sequencing of 179 individuals from four populations with different ancestries: African (Yoruba, YRI), Asian (Chinese plus Japanese, AS), and European (CEU); and produced genotype datafrom approximately 15,000,000 SNPs.

2.6 Population genetic analyses

2.6.1. 1000 Genomes data analyses

A set of programs was developed to retrieve genotypes from the 1000 Genomes Pilot Project MySQL database and to analyse them according to selected regions/populations. These programs were developed in C++ using the GeCo++ [49] and the libsequence [50] libraries.

In order to obtain a control set of random genes to use as a reference set and to define empirical distribution of all statistics applied, I initially selected 1,200 genes by random sampling of those included in the RefSeq list. For these genes I retrieved orthologous regions in the chimpanzee, orangutan or macaque genomes (outgroups) using the LiftOver tool; genes showing less than 80% human-outgroup aligning bases were discarded. This resulted in a final set of 987 genes. These data were used to calculate θ_W [14], π [15], as well as Tajima's D [13], Fu and Li's D* and F* [51], and normalized Fay and Wu's H [16, 52] over each entire gene regions.

Normalized Fay and Wu's H was also calculated in 5kb sliding windows moving with a step of 500 bp.

F_{ST} [19] and the DIND test [22] were calculated for all SNPs mapping to the control gene sets and to the genes of interest. Because F_{ST} values are not independent of allele frequencies, I binned variants based on their minor allele frequency (MAF) and calculated percentiles for each MAF class. As for the DIND test, I calculated statistical significance by obtaining an empirical distribution of DIND-DAF (derived allele frequency) value pairs for variants located within control genes. Specifically, DIND values were calculated for all SNPs using a constant number of 40 flanking variants (20 up- and down-stream). The distributions of DIND-DAF pairs for African, Asian and European populations was binned in DAF intervals and for each

class percentiles were calculated. Due to the nature of low-coverage data, for low DAF values most $i\pi_D$ resulted equal to 0 (i.e. the 95th percentile could not be calculated); thus, I did not calculated DIND in these ranges and consequently selection acting on low frequency derived alleles can not bed detected.

The XP-EHH and iHS tests were calculated as previously described [20, 23]; specifically, the two statistics were calculated for all tested SNPs using information from 200 kb flanking regions (100 kb 5' and 3'). To obtain empirical distributions, I randomly selected 100 genic SNPs and calculated their values for all SNPs in their 200 kb flanking regions.

2.6.2. Sanger-resequenced data analyses

Haplotypes were inferred from Sanger resequencing data using PHASE version 2.1 [53, 54]. Median-joining networks to infer haplotype genealogy were constructed using NETWORK 4.5 [55]. Estimates of the time to the most recent common ancestor (TMRCA) was obtained using different methods: i) a phylogeny based approach implemented in NETWORK 4.5 using a mutation rate based on the number of fixed differences between chimpanzee and humans [55]; ii) GENETREE, which is based on a maximum-likelihood coalescent method [56, 57] assuming an infinite-site model without recombination; haplotypes and sites that violate these assumptions were removed; iii) a previously described method [58] that calculates the average pairwise difference between all chromosomes and the most recent common ancestor; this value was converted into years on the basis of mutation rate retrieved as above.

I based calculations on the assumption that the divergence between human and chimpanzee occurred 6 MY ago [59] and that the generation time is 25 years.

An approach based on coalescent simulations was applied to Sanger

sequencing data. In particular, calibrated coalescent simulations with different demographic parameters were performed using the cosi package [18]. Simulations were conditioned on mutation and recombination rates. Estimates of the population recombination rate parameter ρ were obtained from resequencing data with the use of the Web application MAXDIP [60].
For Sanger-resequenced regions the percentile ranks of θ_W and π were obtained from the distribution of the same parameters calculated for 5Kb windows deriving from 238 human genes resequenced by NIEHS (National Institute of Environmental Health Sciences) SNPs Program. The maximum-likelihood-ratio HKA test was performed using the MLHKA software [61].

2.6.3. Detection of pathogen-driven selection

HGDP-CEPH panel data derive from a previous work [62]. The approach used to identify variants selected by different pathogen species can be briefly described as follows: it is based on calculating the Kendall's correlation coefficient (τ) between allele frequencies of HGDP-CEPH SNPs [62] and pathogen diversity in the countries where populations included in the Panel live. In order to account for demographic events, each SNP is then assigned a percentile rank in the distribution of τ values calculated for all SNPs having a minor allele frequency (MAF) similar (in the 1% range) to that of the SNP being analyzed [63-66]. I considered a SNP to be significantly associated with pathogen diversity if it displayed a significant correlation coefficient and a τ rank higher than 0.95.
F_{ST} was calculated for all HGDP-CEPH variants among continental groups; F_{ST} distributions were calculated for the MAF-matched SNP classes described above; outliers are defined as variants with an F_{ST} higher than the 95th percentile in the distribution of SNPs in the same MAF class.
I also applied a resampling approach in order to evaluate whether the

identification of a set of pathogen-selected SNPs is expected by chance. In particular, I retrieved all GWAS SNPs associated with any trait or disease and collapsed SNPs in tight linkage disequilibrium (r^2 >0.8) into single loci. By performing 10,000 re-samplings of a set of randomly selected SNPs I calculated an empirical probability of obtaining variants significantly associated with pathogen diversity.

2.6.4. Comparison of Neandertal and Denisova data

Alignments of Neandertal [67] and Denisova [68] sequence reads to the human reference genome were retrieved from the UCSC website (http://genome.ucsc.edu/). I only considered reads with an alignment quality higher than 60 and 25 for Neandertal and Denisova, respectively. Positions where different reads, either for the same or for distinct individuals, carried different bases were discarded.

3 RESULTS AND DISCUSSION

3.1 A 175 million year history of T cell regulatory molecules reveals widespread selection, with adaptive evolution of disease alleles

Immunity
Article

A 175 Million Year History of T Cell Regulatory Molecules Reveals Widespread Selection, with Adaptive Evolution of Disease Alleles

Diego Forni,[1] Rachele Cagliani,[1] Uberto Pozzoli,[1] Marta Colleoni,[1] Stefania Riva,[1] Mara Biasin,[2] Giulia Filippi,[3] Luca De Gioia,[3] Federica Gnudi,[2] Giacomo P. Comi,[4] Nereo Bresolin,[1,4] Mario Clerici,[5,6] and Manuela Sironi[1,*]
[1]Scientific Institute IRCCS E. Medea, 23842 Bosisio Parini (LC), Italy
[2]Department of Biomedical and Clinical Sciences, University of Milan, 20157 Milan, Italy
[3]Department of Biotechnology and Biosciences, University of Milan-Bicocca, 20126 Milan, Italy
[4]Dino Ferrari Centre, Department of Physiopathology and Transplantation, University of Milan, Fondazione Ca' Granda IRCCS Ospedale Maggiore Policlinico, 20122 Milan, Italy
[5]Chair of Immunology, Department of Physiopathology and Transplantation, University of Milan, 20090 Milano, Italy
[6]Don C. Gnocchi Foundation ONLUS, IRCCS, 20148 Milan, Italy
*Correspondence: manuela.sironi@bp.lnf.it
http://dx.doi.org/10.1016/j.immuni.2013.04.008

SUMMARY

T cell activation plays a central role in immune response and in the maintenance of self-tolerance. We analyzed the evolutionary history of T cell regulatory molecules. Nine genes involved in triggering T cell activation or in regulating the ensuing response evolved adaptively in mammals. Several positively selected sites overlap with positions interacting with the binding partner or with cellular components. Population genetic analysis in humans revealed a complex scenario of local (*FASLG*, *CD40LG*, *HAVCR2*) and worldwide (*FAS*, *ICOSLG*) adaptation and *H. sapiens*-to-Neandertal gene flow (gene transfer between populations). Disease variants in these genes are preferential targets of pathogen-driven selection, and a Crohn's disease risk polymorphism targeted by bacterial-driven selection modulates the expression of *ICOSLG* in response to a bacterial superantigen. Therefore, we used evolutionary information to generate experimentally testable hypotheses concerning the function of specific genetic variants and indicate that adaptation to infection underlies the maintenance of autoimmune risk alleles.

INTRODUCTION

T lymphocytes play a central role in the elicitation of effective immune responses and in the maintenance of immune homeostasis. A pivotal role in T lymphocyte activation is played by the B7 family of costimulatory molecules. The best-described B7 molecules, B7-1 (CD80) and B7-2 (CD86), are expressed on antigen-presenting cells (APCs) and engage proteins belonging to the CD28 family on the surface of T lymphocytes (Chen, 2004). Binding of either CD28 or CTLA4 by CD80 or CD86 results in immune activation or in the dampening of immune responses, respectively (Chen, 2004). Among other members of the B7 family, CD274 binds the programmed death 1 (*PDCD1*) receptor and plays an important role in tolerizing and destroying self-antigen-specific cells (Nurieva et al., 2009). PD-1 can bind a second protein, PD-L2 (*PDCD1LG2*), on non-T cells (Nurieva et al., 2009); this interaction suppresses human T cell activation.

Apoptosis can also be mediated by the interaction between FAS and FASL (Strasser et al., 2009) and by the ligation of Galectin-9 (*LGALS9*) by TIM-3 (*HAVCR2*). This latter belongs to the T cell immunoglobulin mucin (TIM) family and is involved in the inhibition of Th1 cell type immune responses (Sakuishi et al., 2011). The interaction between Galectin-9 and TIM-3 is believed to suppress the differentiation of naive T cells toward the Th17 cell lineage and, instead, to stimulate their differentiation into regulatory T (Treg) cells (Sakuishi et al., 2011). Finally, the interaction between ICOS and ICOSL plays an important role in cell-cell signaling and regulation of cell proliferation (Nurieva et al., 2009), whereas the interaction between CD40 and CD40L is central in lymphocyte activation, as well as in the regulation of B cell function (Elgueta et al., 2009).

Although they represent only a subset of players in the complex process of T cell regulation, the above-described proteins have a central function in the elicitation of effective immune responses; as a consequence, several pathogens have evolved strategies that specifically target T cell regulatory molecules to facilitate the establishment of infection (Khan et al., 2012). Therefore, on one hand T cell regulatory molecule genes have been engaged in a constant conflict with pathogens and play a fundamental role during infections; on the other, genetic variation within these loci has a potential impact on the development of immunodeficiencies, autoimmune and inflammatory conditions, and cancer.

RESULTS

Widespread Positive Selection during Mammalian Evolution

We analyzed the following genes: *CD28, CD80, CD86, CD274, CTLA4, PDCD1, PDCD1LG2, ICOS, ICOSLG, FAS, FASLG,*

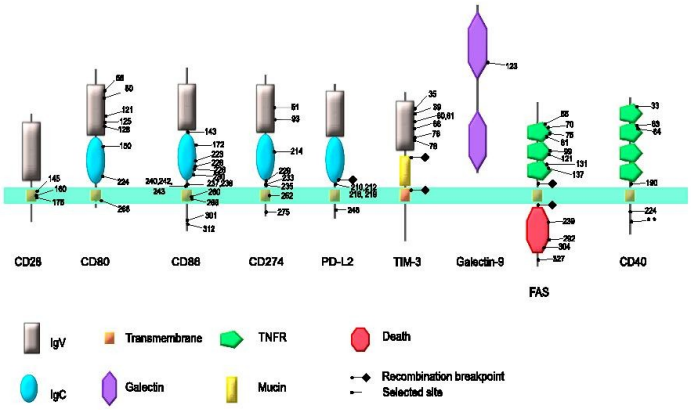


Figure 1. Schematic Domain Representation of the Proteins Encoded by the Nine Positively Selected Genes
Selected sites identified through both BEB and MEME are shown. Positions refer to the human sequence and asterisks represent gaps in the human sequence that correspond to positively selected residues in the alignment. The location of recombination breakpoints is also shown.

CD40, CD40LG, HAVCR2, and *LGALS9* (Table S1 available online). To investigate their evolutionary history in mammals, we retrieved coding sequence information for all available species (a list of the 39 species is available in the Supplemental Experimental Procedures; see also Table S2). Multiple alignments were analyzed to calculate the average ratio of dN (nonsynonymous substitution rate) to dS (synonymous substitution rate). For most protein-coding genes, fewer nonsynonymous replacements were observed compared to synonymous substitutions ($dN/dS < 1$), because alterations in amino acid sequences are more likely to be detrimental to protein function and are selected against (purifying selection). Indeed, analysis of T cell regulatory molecules revealed that all genes evolved under purifying selection ($dN/dS < 1$; Table S3). Nonetheless, positive selection ($dN/dS > 1$) might act on a few sites within a gene that is globally subject to purifying selection. To test this possibility, we applied maximum-likelihood analyses (Yang, 2007). Specifically, we used the codeml program to compare two models of gene evolution that allow a class of codons to evolve with $dN/dS > 1$ against models of neutral evolution (which disallow $dN/dS > 1$). Results indicated that for 9 out of 15 genes studied, the neutral evolution models were rejected in favor of the positive selection models (Table S2). Because recombination might yield false positive results when likelihood-ratio tests are applied (Anisimova et al., 2003), we screened the nine alignments for evidence of recombination via GARD (genetic algorithm recombination detection) (Kosakovsky Pond et al., 2006). *PDCD1LG2*, *FAS*, and *HAVCR2* displayed at least one recombination breakpoint (Figure 1). Thus, we took into account both the breakpoint locations and the domain structure for these T cell regulatory molecules and performed the likelihood-ratio tests on gene subregions (Figure 1). After multiple test correction (Anisimova and Liberles, 2012), positive selection was supported by all tests for at least one domain in each of the three genes (Table S2).

In order to identify specific sites subject to diversifying selection, we applied two methods: the Bayes empirical Bayes (BEB) analysis (Anisimova et al., 2002) and the mixed effects model of evolution (MEME) (Murrell et al., 2012). Only sites detected by both methods were considered (Figure 1).

We next mapped sites targeted by diversifying selection onto available crystal and domain structures. Analysis of CD80 (Figure 2A) indicated that three positively selected sites are located at the CTLA4 binding interface, with one position directly interacting with the MYPPPY conserved sequence of CTLA4; additional selected positions involve (Val56) or closely flank (Asp80) sites that allow interaction between the two CD80 molecules forming the homodimer (Figure 2A; Ikemizu et al., 2000; Stamper et al., 2001). Analysis of the single selected site in the galectin domain of Galectin-9 indicated that it is not directly involved in oligosaccharide binding (Figure S1; Nagae et al., 2006). As for the IgV domain of TIM-3, it also acts as a receptor for phosphatidylserine (PtdSer) (DeKruyff et al., 2010), for the alarmin HMGB1 (Chiba et al., 2012), and for an unknown protein or glycan (Cao et al., 2007). In analogy to other TIM family members, PtdSer is accommodated within a pocket of the TIM-3 IgV domain (metal ion-dependent ligand binding site, MILIBS) created by the CC′ and FG loops (DeKruyff et al., 2010); PtdSer binding alters the structure of the tip of the CC′ loop (Figure 2C; DeKruyff et al., 2010), and we found the two residues at the tip of this loop to be subject to positive selection (Figure 2C). Mapping of positively selected sites in the FAS death domain indicated that two out of the three are located at the primary contact

interface with FADD (Figure S1; Scott et al., 2009). As for the tumor necrosis factor receptor (TNFR) domain, residues involved in binding to FASL have been identified by mutagenesis (Bajorath, 1999; Starling et al., 1998). Although no position directly involved in ligand-receptor interaction was among the positively selected sites we identified, these latter tend to cluster within a relatively short stretch where the majority of FAS-FASL interactions occur (Figure S1). Likewise, one of the selected sites in the TNFR domain of CD40 is located at the binding interface with CD40L (Figure 2D; An et al., 2011). Analysis of sites targeted by diversifying selection in the cytoplasmic tail of CD40 indicated that they occur within regions serving as binding sites for Ku, JAK3, and TRAF6 (Morio et al., 1999; Pullen et al., 1999); conversely, residues involved in the binding of TRAF1-3 are highly conserved (Figure 2D; Lu et al., 2003; Pullen et al., 1999). Finally, in the case of CD274, one of the positively selected sites in the IgV domain flanks residues forming the PD-1 binding interface (Figure 2B; Lin et al., 2008). In the case of CD86, several positions subject to diversifying selection were identified in the juxtamembrane (JM) and transmembrane (TM) regions (Figure 1), which interact with the MIR2 immunomodulator encoded by KSHV (Kaposi sarcoma-associated herpesvirus) (Kajikawa et al., 2012). Therefore, we performed ab initio structure prediction and in silico docking analysis to study the interaction between MIR2 and CD86. The structure of the CD86 JM region was predicted to be random coil, preventing docking analysis; conversely, the TM region was amenable to docking calculation: this revealed that the two CD86 positively selected sites, Val260 and Trp286, are located at the contact interface of the two proteins and can interact, mainly via hydrophobic contacts, with residues of MIR2 (Figure 2E). To verify the importance of the amino acids in positions 260 and 268, we created four different in silico CD86 variants (Val260Ala and Trp268Ala, or Val260Gly and Trp268Arg, these latter observed in some nonprimate mammals, Figure 2E): none of them was able to bind MIR2 with the same orientation as the human protein (Figure 2E). This analysis confirms a crucial role of Val260 and Trp268 in the correct positioning of the two proteins during the interaction. The computed binding poses of the variant CD86 molecules with MIR2 are unlikely to occur in vivo, because of the presence of the phospholipid bilayer (which is not explicitly considered in the docking experiments). Therefore, these results suggest that variation at the positively selected sites renders interaction between CD86 and MIR2 unlikely or very weak.

Local Adaptation, Balancing Selection, and Possible Gene Flow among Archaic Hominins

We next investigated whether natural selection has affected genetic diversity at T cell regulatory molecule genes in human populations. We exploited data from the 1000 Genomes pilot project deriving from the low-coverage whole-genome sequencing of 179 individuals with different ancestry: Europeans (CEU), Yoruba from Nigeria (YRI), and East Asians (AS; Japanese plus Chinese) (1000 Genomes Project Consortium et al., 2010). We calculated the following parameters: (1) θ_W (Watterson, 1975) and π (Nei and Li, 1979) describe genetic diversity and were estimated for all T cell regulatory molecule genes (Figure S2); (2) site frequency spectrum (SFS)-based statistics such as Tajima's D (Tajima, 1989), normalized Fay and Wu's H (Fay and Wu, 2000), and Fu and Li's F* and D* (Fu and Li, 1993) were also calculated over all genes (Table S4); (3) F_{ST}, a measure of population genetic differentiation, was obtained for all single-nucleotide polymorphisms (SNPs) within T cell regulatory molecule genes (Figure S2); and (4) tests based on haplotype homozygosity (derived intra-allelic nucleotide diversity [DIND] and integrated haplotype score [iHS]) (Barreiro et al., 2009; Voight et al., 2006) were calculated or retrieved for all SNPs in the genes under analysis (Figure S2). The statistical significance of all tests was obtained by deriving empirical distributions of the same parameters calculated for a randomly selected set of ~1,000 genes (see Supplemental Experimental Procedures).

SNP genotype data from the Human Genome Diversity Panel (HGDP) (Li et al., 2008) were also used to calculate F_{ST} values among continental groups, haplotype homozygosity (iHS) (Sabeti et al., 2007), and cross population extended haplotype homozygosity (XP-EHH) (Voight et al., 2006). Finally, measures of the selective pressure exerted by viruses, bacteria, protozoa, and helminths were obtained (see Experimental Procedures). Data from the 1000 Genomes Project and from the HGDP resource were integrated and we considered genes to represent selection targets if they showed significant results in the same population for at least two tests based on different features (e.g., F_{ST} and haplotype-based tests), as previously suggested (Manry et al., 2011).

Six genes satisfied these criteria, although closer inspection of *CD80* genetic diversity indicated that the nearby gene might represent the selection target. *CD80* showed two variants with a significant DIND test (Figure S2) in YRI. The rationale behind the DIND test is that a derived allele under positive selection will display lower nucleotide diversity at linked sites than expected from its frequency in the population. Thus, the ratio of intra-allelic diversity associated with the ancestral and derived alleles ($i\pi_A/i\pi_D$) was analyzed against the frequency of the derived allele (DAF) (Figure S2); given a DAF interval, a high value of $i\pi_A/i\pi_D$ suggests positive selection. One of the *CD80* variants with a significant DIND test (rs6810215, DIND = 23.24, DAF = 0.48) also had extremely high F_{ST} values (Figure S2). As noted above, F_{ST} is a measure of population genetic differentiation: under selective neutrality F_{ST} is mainly determined by demographic history and drift, but natural selection may drive allele frequencies to differ more than expected. Because selective sweeps may involve long genomic regions, we extended the analysis to flanking regions: a nonsynonymous variant in *TIMMDC1* (rs57168946) was in linkage disequilibrium (LD) with the variants in *CD80* (Figure S3) and showed a highly significant DIND test (DIND = 79, DAF = 0.48), suggesting that it represents the selection target.

For *CD274* we identified five variants with significant DIND test results in AS (Figures 3A and S2). The variant showing the highest DIND (rs2890657) also showed an F_{ST} value higher than the 99th percentile in the YRI versus AS comparison (Figure S2). rs822339, which is in LD with rs2890657 (r^2 = 0.82 in AS), displayed a significant correlation with the diversity of bacteria (rank = 0.96). We performed haplotype analysis over an extended genomic interval encompassing *CD274* (Figure S3). Within this region, no SNP displayed a DIND test higher than rs2890657, and only one (rs10815233) had a higher F_{ST}. Overall, these data indicate that a homogeneous haplotype increased in

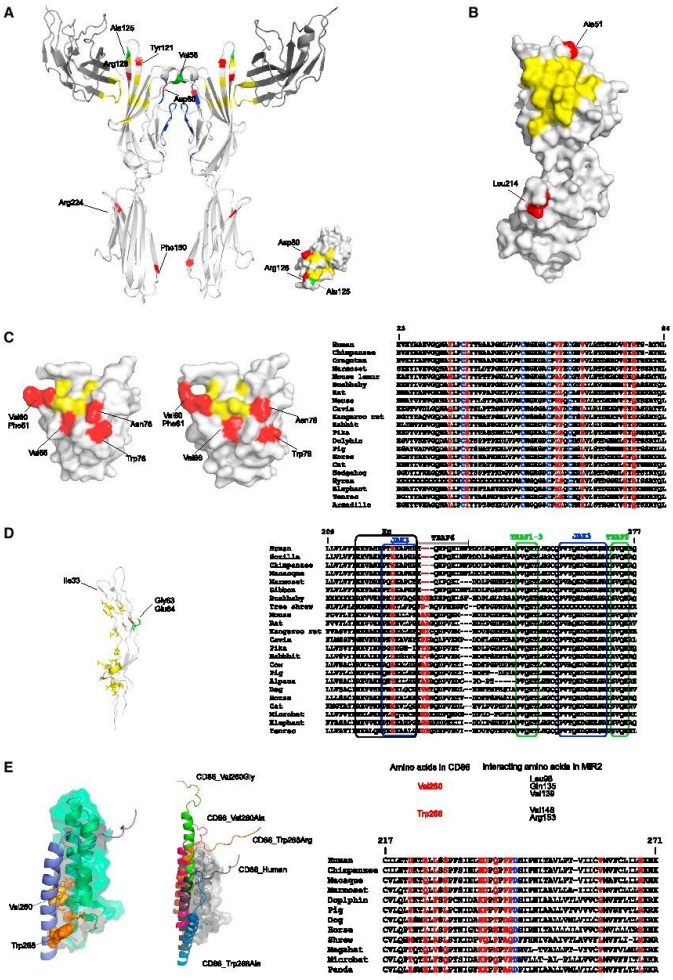


Figure 2. Analysis of Positively Selected Sites

In all panels positively selected sites are shown in red.

(A) Ribbon diagram of the extracellular regions of CD80 (light gray) and CTLA4 (dark gray). Residues involved in the CD80-CTLA4 interaction are represented in yellow on both proteins; positions involved in the interaction between the two CD80 monomers are in blue. Positions in green indicate positively selected sites that

(legend continued on next page)

frequency in AS (Figure S3), possibly because of the selective pressure exerted by bacterial pathogens, and suggest that rs2890657 represents a selection target. This SNP is located in a region characterized by H3K4Me1 and H3K27Ac histone marks and by the binding of RNA polymerase II, as determined by chromatin immunoprecipitation and sequencing (ChIP-seq) (Figure 3A).

CD40LG showed very low diversity in CEU and tended to display low π in AS (Figure S2); in this latter population, Tajima's D and Fay and Wu's H displayed significantly low values (Table S4). These features are suggestive of a selective sweep. In fact, the increase in frequency of a selected haplotype may result in a temporary reduction in the level of genetic variability and in a shift of the SFS, leading to a deficiency of intermediate frequency alleles (negative values of Tajima's D or Fu and Li's D* and F*). Also, a selective sweep may determine an excess of high-frequency derived alleles (negative values of Fay and Wu's H). F_{ST} analysis of *CD40LG* indicated that rs3092923 had a value higher than the 95th percentile in both YRI versus CEU and YRI versus AS comparisons (Figure S2). Another variant in *CD40LG* (rs3092936) was an outlier in the distribution of F_{ST} values calculated for continental groups in the HGDP, an effect mainly driven by populations in the Americas (Figure 3B). Therefore, we performed haplotype analysis over an extended region with the inclusion of the 1000 Genomes Main Project data, which comprise Native American populations (Figure S3). Two variants in the region, rs3092923 and rs3092921, had extremely high F_{ST} (YRI versus AS). The former was located at the end of intron 4, a region where a T-cell-lineage-specific enhancer has been described (Schmidl et al., 2009), and the region surrounding rs3092923 displayed H3K4me1 histone marks (Figure 3B). In Native Americans a second extended haplotype carried the derived allele of rs3092936 plus several additional derived alleles (Figures 3B and S3). These define a haplotype with a frequency of 27% in Native Americans and near absence in Africa, Europe, and Asia. To identify the putative selection target(s), we calculated AS versus American F_{ST} for all variants in the region: the originally identified rs3092936 together with nine additional SNPs had the highest F_{ST} (0.153) and the ten variants displayed the same DAF of 0.27 in Americans. Three of these SNPs fall within noncoding sequences that are highly conserved in mammals (Figure 3B).

For *HAVCR2*, we identified two variants with a significant DIND test in YRI (Figure S2); one of these SNPs (rs4704846) also displayed extreme F_{ST} values in the YRI versus AS comparison (Figure S2), whereas the other (rs11741184) had a significant iHS score (iHS = −2.037) in Africans. The iHS test is similar in concept to the DIND test and is based on haplotype homozygosity. Analysis over an extended region (Figure S3) indicated that no variant has higher DIND or F_{ST} than rs4704846. This variant falls in the 3′ UTR, within a predicted binding site for miR-379 (http://www.microrna.org/microrna), and replaces a G-C pair to a G-U wobble (Figure 3C). Two *HAVCR2* variants (rs61159436 and rs6886320) significant at the DIND test were also identified in AS (Figure S2), and in this population low values of Fu and Li's D* and F* were observed (Table S4). rs61159436 had the highest DIND test in the analyzed region and is located in the fifth intron of *HAVCR2* (Figure 3C).

FAS showed high diversity in AS and CEU populations (Figure S2), and several SFS-based statistics were significantly high in these same populations (high values of Tajima's D or Fu and Li's F* and D* suggest balancing selection) (Table S4). Additionally, we detected six variants with extreme F_{ST} values in the CEU versus YRI comparison (Figure S2) and located in the 5′ portion of intron 1 (Figure 4A); one variant in this region (rs7097467) was found to significantly correlate with helminth diversity (rank = 0.96). These features might suggest the action of balancing selection in the 5′ gene region. Signatures of balancing selection tend to be elusive, because no extended haplotype is expected under this selective regime. Therefore, we performed Sanger-resequencing and coalescent simulations to test for selection. We sequenced a ~5 kb region encompassing the *FAS* core promoter and first exon (Figure 4A). Results confirmed very high nucleotide diversity in CEU and AS, as well as high values for SFS statistics especially in CEU (Table 1). Under neutral evolution, the amount of within-species diversity correlates with between-species divergence, because both depend on the neutral mutation rate. This comparison is formalized in the Hudson-Kreitman-Aguadè (HKA) test (Hudson et al., 1987). We performed a maximum likelihood HKA test (MLHKA): for all populations, MLHKA rejected the null model of selective neutrality and an excess of polymorphism compared to divergence was detected (Table 1). Haplotype analysis identified two major clades (A and B) (Figure 4B), and calculation of the time to the most recent common ancestor (TMRCA) resulted in estimates ranging from 3.75 to 2.03 My (Table S5). Analysis of variants in the haplotype network indicated that rs1800682 (−670 A>G) separates the two major clades (Figure 4B); this variant affects a signal transducer and activator of transcription-1 (STAT1) binding site (Kanemitsu et al., 2002) and

also represent positions directly involved in CTLA4 binding or in CD80-CD80 interactions. The insert shows the CD80 surface involved in the interaction with CTLA4.

(B) Structure of CD274 extracellular domain; residues in yellow directly interact with PD-1. Residue 93 is not visible in this image (back surface).

(C) Surface of the TIM-3 IgV domain bound (left) or unbound (right) to PtdSer. Residues 35 and 39 are not visible in this model (back surface). Residues in yellow are involved in the binding of an unknown cellular component. The alignment for a portion of the IgV domain is also shown for a few representative mammalian species. Color codes are the same as in the surface model; four cysteines conserved in all TIM genes are shown in blue.

(D) Ribbon diagram of the TNFR domain of CD40 and alignment of the CD40 cytoplasmic domain. The yellow residues in the TNFR domain are involved in binding to CD40L. The green residue is positively selected and involved in CD40L binding. On the alignment, the binding sites for Ku, JAK3, and TRAFs are shown.

(E) Structures of CD86 and MIR2. Left: docked structures of CD86 (blue) and MIR2 (green). The human CD86 amino acids Val260 and Trp268, which are involved in protein-protein interaction, are shown as yellow and orange spheres, respectively. The residues of MIR2 that interact with these two amino acids are represented as sticks. Middle: docked structures of human CD86 (black) and four different variants: Val260Ala (fuchsia), Val260Gly (green), Trp268Ala (blue), and Trp268Arg (orange). Right: MIR2 interacting sites and alignment of the JM and TM regions of CD86 for a few representative mammals; the aspartic residue at position 244 is in blue.

See also Figure S1.

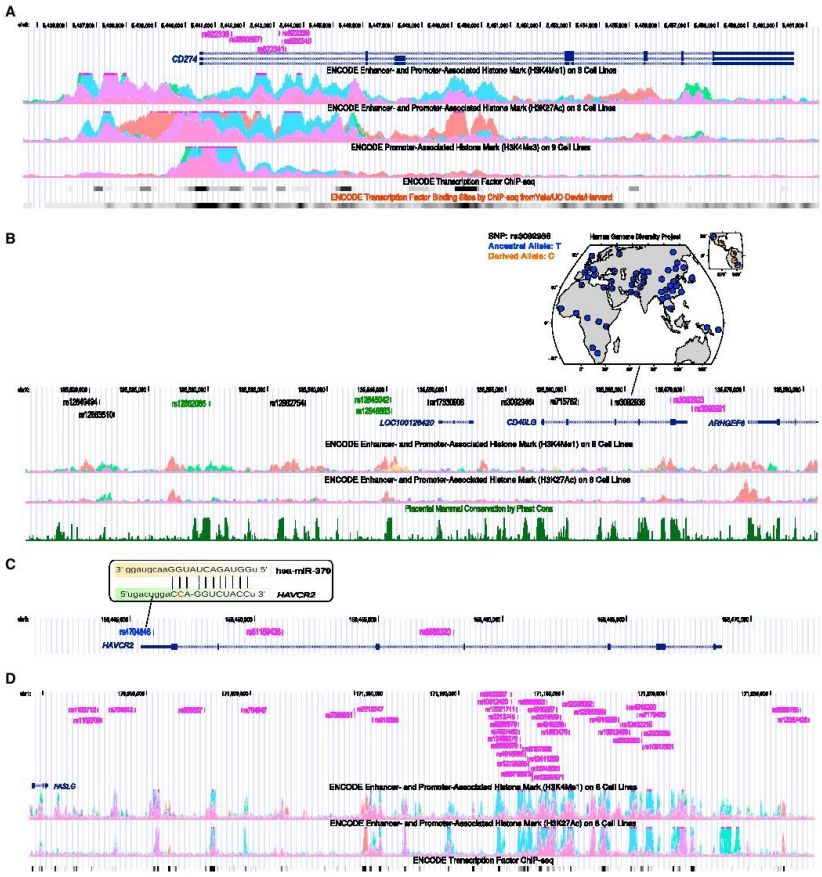


Figure 3. Analysis of Selected Variants

Location of the most likely selection targets in *CD274* (A), *CD40LG* (B), *HAVCR2* (C), and *FASLG* (D) within the UCSC Genome Browser view. Relevant ENCODE tracks are also shown. Variants in blue and magenta represent selection targets in YRI and AS, respectively. Polymorphisms in black represent selected sites in Native Americans (B) and those in green fall within PhastCons elements. See also Figure S3.

modulates *FAS* expression in response to interferon γ (IFN-γ) (Farre et al., 2008). The helminth-selected variant defined a subgroup of haplotypes in clade A (Figure 4B). Overall, these data are consistent with a model of multiallelic balancing selection.

Finally, *PDCD1* was found to display very high nucleotide diversity in AS populations (Figure S2). F_{ST} analysis indicated that several variants with extreme values in the CEU versus AS comparison are located both in the 5′ and 3′ gene regions;

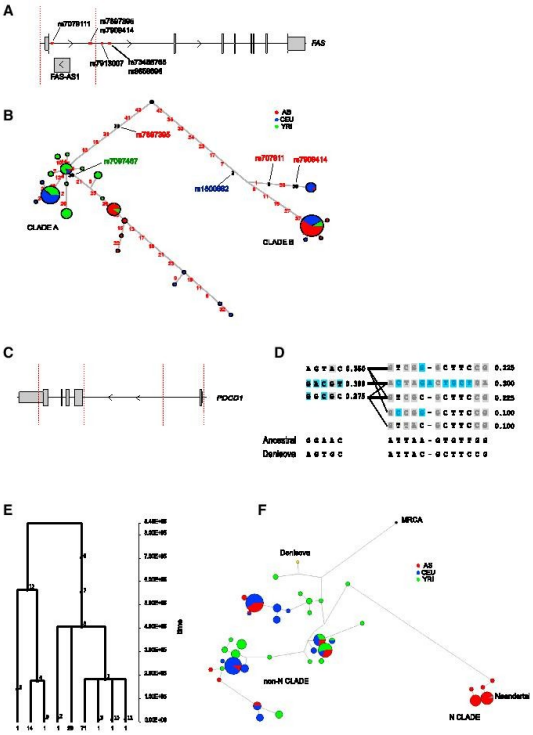


Figure 4. *FAS* and *PDCD1* Analysis

(A) Exon-intron structure of *FAS* with the location of F_{ST} outliers (red dots). The region we resequenced is delimited by the red hatched lines.

(B) Genealogy of haplotypes in the sequenced region reconstructed through a median-joining network. Each node represents a different haplotype, with the size of the circle proportional to frequency. Nucleotide differences between haplotypes are indicated on the branches. Variants with extreme F_{ST} and correlating with helminth diversity are shown in red and green, respectively. The position of rs1800682 (blue), which affects STAT1 binding, is shown.

(C) Exon-intron structure of *PDCD1*; resequenced regions are delimited by hatched lines.

(D) Blocks of LD for the two resequenced regions (AS haplotypes only). Cyan and gray denote position matching the Neandertal sequence and not covered in Neandertal, respectively; the ancestral state and Denisova sequences are also shown.

(E) GENETREE for the LD subregion of *PDCD1*. Variants are represented as black dots; the absolute frequency of each haplotype is reported.

(F) Network analysis of extant haplotypes with the Denisova (yellow) and Neandertal (cyan) sequences. The network was constructed for a 30.5 kb region after discarding positions where either the ancestral state could not be inferred or the Denisova and Neandertal genomes were not covered. Haplotypes occurring only once were also discarded for network construction.

See also Figure S4.

therefore, we resequenced the whole *PDCD1* coding sequence (Figures 4C and S2). Calculation of θ_W and π confirmed high nucleotide diversity in AS and high Tajima's D in this population (Table 1). Conversely, in CEU and YRI diversity was not exceptional and several SFS statistics tended to be low (Table 1). Haplotype analysis revealed two major clades, one of which (N clade) was almost specific to Asians and matching the Neandertal sequence in all positions (seven derived, three ancestral) where the hominin sequence is covered (Figure 4D). We estimated a coalescence time of 602,298 years for the whole haplotype phylogeny and similar TMRCAs of 344,827 and 445,623 years for clade N and non-N haplotypes, respectively. A more robust TMRCA estimate was calculated with GENETREE over a shorter region of LD (Figure S4) and yielded comparable results (Figure 4E). Calculation of π for Asian chromosomes gave estimates of 3.55×10^{-4} and 6.52×10^{-4} for the N and non-N clade, respectively. To gain further insight, we retrieved human haplotypes for a ~30 kb region centered around *PDCD1*. For each polymorphic position, we determined the ancestral state (from at least two primate genomes) and the status in the two Neandertal and Denisova genomes. A median-joining network indicated that the Neandertal and Denisova sequences cluster with N and non-N haplotypes, respectively (Figure 4F). Finally, we calculated the extended haplotype homozygosity (EHHA) over a 200 kb region: a similar EHHA plot was obtained for the two alleles of a variant tagging the N and non-N haplotypes (Figure S4).

Natural Selection Shaped Allele Frequencies at Disease Variants

To study the selective patterns of disease SNPs in T cell regulatory molecule genes, we focused on variants that have been identified through genome-wide association studies (GWASs). A total of 20 GWAS variants were retrieved. Out of these, 13 SNPs were available in the HGDP-CEPH panel and 7 correlated with the diversity of at least one pathogen species (Table 2).

Although some of these seven variants are located relatively close to each other, LD among SNPs pairs is relatively low (r^2 not exceeding 0.8 in either CEU, YRI, or AS), suggesting that

Table 1. Nucleotide Diversity and Neutrality Tests for Sanger-Sequenced Regions

				θ_w[a]		π[b]		Tajima's D		Fu and Li's D*		Fu and Li's F*		MLHKA	
Gene	L[c]	Pop[d]	S[e]	Value	Rank[f]	Value	Rank[f]	Value	p[g]	Value	p[g]	Value	p[g]	K[h]	p[i]
FAS	5.4	YRI	31	13.48	0.89	9.95	0.74	−0.90	0.34	1.08	0.02	0.47	0.11	3.17	00042
		CEU	30	13.05	0.95	21.73	0.98	2.29	0.01	1.29	0.04	1.94	0.01	4.17	0.00027
		AS	31	13.48	0.97	21.21	0.98	1.98	0.03	0.11	0.45	0.89	0.19	5.13	0.00014
CTLA4	2.4	YRI	21	21.19	0.98	26.20	0.99	0.79	0.049	0.74	0.07	0.89	0.041	2.64	0.026
		CEU	15	14.45	0.98	22.85	0.98	1.85	0.028	1.56	0.009	1.95	0.0074	2.31	0.079
		AS	12	11.56	0.95	14.90	0.94	0.89	0.23	−0.05	0.45	0.30	0.42	2.10	0.095
ICOSLG	2.4	YRI	21	20.42	0.98	18.40	0.97	−0.33	0.62	0.68	0.12	0.41	0.16	3.34	0.0045
		CEU	17	16.53	0.98	31.96	>0.99	3.03	<0.001	1.60	0.0018	2.45	<0.001	3.68	0.0061
		AS	19	18.47	0.99	35.09	>0.99	2.96	0.0017	1.29	0.05	2.18	0.0026	4.53	0.00083
PDCD1	4.3	YRI	22	12.04	0.81	8.23	0.61	−1.06	0.24	−2.46	0.025	−2.35	0.035	0.82	0.09
		CEU	20	10.95	0.91	7.98	0.76	−0.90	0.15	−2.82	0.021	−2.57	0.028	1.15	0.18
		AS	21	11.49	0.95	19.12	0.98	2.21	0.02	0.35	0.35	1.15	0.1	1.51	0.34

See also Figure S2.
[a] θ_w estimation per site ($\times 10^{-4}$).
[b] π estimation per site ($\times 10^{-4}$).
[c] Length of analyzed resequenced region (in kb).
[d] Population (YRI, Yoruba; CEU, European; AS, Asian).
[e] Number of segregating sites.
[f] Percentile rank relative to a distribution of 238 5 kb windows from NIEHS genes.
[g] p value obtained from coalescent simulations.
[h] Selection parameter (k > 1 indicates an excess of polymorphism compared to divergence).
[i] p value for the MLHKA test.

they might represent at least partially independent selective events. We evaluated the probability to obtain a similar number of pathogen-selected variants by applying a resampling approach (Supplemental Experimental Procedures). Results indicated that disease variants in T cell regulatory molecule genes are preferential targets of pathogen-driven selection ($p = 0.013$).

Analysis of the seven variants indicated that in six cases (i.e., for all autoimmune diseases), the susceptibility allele was associated with higher pathogen diversity (Table 2); in the case of IgA deficiency, the opposite situation was observed, with the susceptibility allele having higher frequency in regions where pathogen load is lower.

Pathogen-driven variations in allele frequencies can occur under different selective scenarios such as directional or balancing selection. Therefore, we calculated F_{ST} and performed the DIND test for disease variants and estimated SFS-based statistics in 5 kb windows around each disease SNP.

The two variants in *FASLG* that correlate with pathogen diversity also showed a significant DIND test in AS (Table S6), very high DAF in this population, and extreme F_{ST} values (Table S6). These two SNPs lie in the large intergenic spacer separating *FASLG* from *TNFSF18* and analysis of the HDGP data indicated that this region showed significantly high XP-EHH and iHS scores in East Asia and America. Haplotype analysis in AS showed that many variants in a large genomic region covering the intergenic spacer had extreme F_{ST} values (Figure 3D). Several of these SNPs cluster in a region where histone marks associated with transcriptional enhancers have been described and some fall within ChIP-seq-identified transcription factor binding sites (Figure 3D).

The rs231735 variant associated with rheumatoid arthritis (RA) (Table 2) is located upstream the transcription start site of *CTLA4*. Given the high values for both θ_w and π in most populations and the positive SFS-based statistics in YRI and CEU (Table S7), we Sanger-resequenced a ~2.5 kb region centered around the variant. θ_w and π values higher than the 95th were confirmed in YRI and CEU (Table 1); the MLHKA test revealed a significant excess of polymorphism over divergence in YRI (borderline in CEU, Table 1). In non-Asian populations, simulations indicated significantly high results for most SFS-based statistics (Table 1), and the estimated TMRCA for the haplotype phylogeny ranged from 2.84 to 1.9 My (Table S5). Overall, these data support the action of balancing selection on this region in African populations and possibly in CEU. The RA risk variant that correlated with helminth and bacteria diversity (Table 2) separates the two major haplotype clades (Figure 5A). Interestingly, another SNP (rs6715389) on the major branch fell within a sequence that is conserved in mammals and changes an almost invariant position, suggesting that it might represent the balancing selection target (Figure 5A).

Finally, for *ICOSLG*, two risk SNPs (rs762421, which strongly correlates with bacterial diversity, and rs2838519) for inflammatory bowel disease (IBD) are located downstream the transcription end site in close proximity to each other. Resequencing of the interval encompassing the two disease variants confirmed (Table S7) high diversity in all populations and significant MLHKA results were obtained (Table 1). In CEU and AS most SFS-based statistics were significantly higher than expected under neutrality (Table 1). Thus, the region carrying the two IBD risk variants represents a balancing selection target. Haplotype analysis revealed a genealogy (TMRCA 4.37–2.39 My, Table S5) with two

Table 2. Correlations with Pathogen Diversity for SNPs Associated with Different Traits

GWAS SNP	Risk Allele	Trait	Gene or Region	Virus Diversity		Bacteria Diversity		Protozoa Diversity		Helminth Diversity	
				p Value[a]	Rank[b]	p Value[a]	Rank[b]	p Value[a]	Rank[b]	p Value[a]	Rank[b]
rs231735	T*	rheumatoid arthritis	intergenic (*CD28, CTLA4*)	>0.05	0.79	1.8×10^{-4}	0.95	>0.05	0.73	9.3×10^{-9}	1.00
rs1024161	T*	Graves' disease; alopecia areata	intergenic (*CD28, CTLA4*)	>0.05	0.76	2.3×10^{-4}	0.87	5.2×10^{-4}	0.90	8.3×10^{-6}	0.98
rs762421	G*	Crohn's disease	intergenic (*ICOSLG*)	>0.05	0.75	7.7×10^{-7}	0.99	>0.05	0.73	3.7×10^{-3}	0.80
rs9282641	G*	multiple sclerosis	Intronic or UTR (*CD86*)	4.1×10^{-4}	0.98	>0.05	0.42	1.9×10^{-5}	0.96	6.8×10^{-4}	0.87
rs2234978	A	IgA deficiency	exonic, synonymous (*FAS*)	>0.05	0.78	1.2×10^{-6}	0.98	1.0×10^{-5}	0.97	1.4×10^{-4}	0.93
rs859637	A*	celiac disease	intergenic (*FASLG*)	>0.05	0.68	4.3×10^{-6}	0.97	9.4×10^{-4}	0.89	4.1×10^{-5}	0.97
rs9286879	G*	Crohn's disease	intergenic (*FASLG*)	>0.05	0.73	5.3×10^{-6}	0.96	8.9×10^{-4}	0.89	2.0×10^{-5}	0.97

NOTE: the risk allele is denoted with an asterisk if it positively correlates with pathogen diversity (i.e., the frequency of the risk allele increases with pathogen diversity).
[a]p values are Bonferroni corrected for 13 tests.
[b]Percentile rank of tau relative to the distribution of SNP control sets matched for allele frequency.

major clades (A and B) (Figure 5B). Most haplotypes in clade B carry both risk alleles for IBD. We next evaluated the induction of *ICOSLG* mRNA after Staphylococcal enterotoxin B (SEB) treatment in peripheral blood mononuclear cells (PBMCs) from 18 healthy volunteers with different rs762421 genotype (six individuals for each genotype). A significant difference in *ICOSLG* induction was observed among the three genotype groups (one-way ANOVA, $p = 0.043$) with the C allele (which confers increased risk for CD and correlates with bacterial diversity) determining higher expression (Figure 5C). On average, SEB induction led to a 2.5-fold higher *ICOSLG* expression in PBMCs from subjects homozygous for the C allele compare to T homozygotes (t test, two-tailed, $p = 0.033$).

DISCUSSION

Evolutionary analyses can provide information on the location and nature of adaptive changes in genomic regions, highlighting the presence of functional variation. We aimed at providing a comprehensive analysis of T cell regulatory molecule gene evolution, although we stress that those we analyzed herein by no means represent the whole set of molecules involved in the regulation of T cell activity. We show that the majority (9 out of 15) of T cell regulatory molecule genes have been targeted by positive diversifying selection in mammals. Notably, some genes that were not positively selected in mammals represented selection targets during the recent evolutionary history of human populations and carry selected alleles associated with autoimmune diseases (e.g., *ICOSLG* and *FASLG*). This may reflect not only the variability of selective pressures and the different time spans involved but also the availability of different methods to study evolutionary processes at the inter- and intraspecific level. At both levels, infectious agents may have represented a powerful selective force acting on T cell regulatory molecules. Pathogens either may develop strategies to modulate the transcription of these genes (Khan et al., 2012) or they may encode molecules that directly bind T cell regulatory molecules and alter their function. Also, T cell regulatory molecules may be commonly exploited as receptors by viruses (Dermody et al., 2009). Indeed, the IgV domains of several proteins are used as viral receptors; in these cases the viral components invariably engage sites in the CC′FG β sheet of the immunolglobulin fold (Dermody et al., 2009). Residues Val60 and Phe61 in TIM-3 and Val56 in CD80 occur on the CC′ β strand (Ikemizu et al., 2000) of the IgV domain, whereas position 51 in CD274 immediately flanks the C strand (Lázár-Molnár et al., 2008). Thus, these sites might have evolved to avoid recognition by some extant or extinct viral species. The hypothesis that diversifying selection at T cell regulatory molecule genes is at least partially virus driven was tested more directly for CD86. Indeed, the KSHV MIR2 ubiquitinase directly binds CD86 through its JM and TM regions (Kajikawa et al., 2012). Docking analysis indicated that the two positively selected sites in the TM region of CD86 are crucial for the interaction with MIR2. Therefore, the selective pressure exerted by MIR2 might have driven the evolution of the CD86 TM region to decrease binding by viral-encoded ubiquitinases or to displace the ubiquitine ligase domain from its targets in the cytoplasmic domain of CD86. Interestingly, MIR2-like proteins are encoded by other herpesviruses and poxviruses, suggesting that they represent a conserved viral mechanisms to control the host immune response (Mansouri et al., 2003).

Several positively selected sites we identified are very close to or overlap with positions directly involved in interactions with the binding partner or with cellular components. In TIM-3, two of the positively selected sites at the top of the CC′ loop rim the pocket that accommodates PtdSer, a central signal exposed by apoptotic cells and exploited by intracellular pathogens to dampen the host response (Wanderley and Barcinski, 2010). Substitution of the mouse residues at the top of the CC′ loop of TIM-3 (WSQ) with the corresponding human amino acids (VFE) significantly decreases PtdSer binding (DeKruyff et al., 2010), providing direct evidence that positive selection at these sites affected the functional properties of TIM-3. Experiments in mice have recently shown that the MILIBS is also important for the interaction with the alarmin HMGB1 (Chiba et al., 2012). Thus, it will be interesting to evaluate whether the positively selected sites affect the efficiency of TIM-3 binding to HMGB1. We also detected several positively selected sites in the

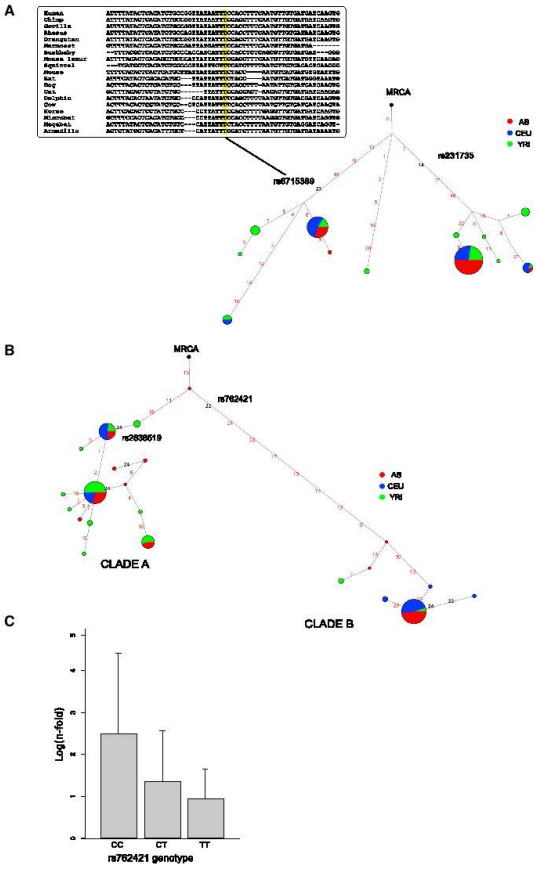


Figure 5. *CTLA4* and *ICOSLG* Analysis

(A) Median-joining network for haplotypes in the region surrounding rs231735 (RA susceptibility variant). The position of rs231735 is shown, as well as the position and multiple alignment for rs6715389.

(B) Genealogy of haplotypes in the region encompassing rs762421 and rs2838519 (downstream *ICOSLG*).

(C) Analysis of *ICOSLG* expression after stimulation of PBMCs with SEB. *ICOSLG* abundance is shown as fold-change expression from the unstimulated sample. Mean values ± standard errors are shown.

extracellular JM regions of PD-L2 and CD274 and one in the JM portion of CD40. The corresponding "stalk" regions of TNF receptors modulate the responsiveness to the soluble but not to the membrane-bound form of TNF (Richter et al., 2012), suggesting that the JM regions of the PD-1 receptors and CD40 might play still unknown relevant functions. Indeed, CD40 belongs to the TNFR family and signals through both a membrane-bound and a soluble form of CD40L. The three positively selected sites in CD28 are located in the TM domain; this region could contain signals that modulate the association of CD28 with lipid rafts, as demonstrated, for example, for the TM domains of CD40 and FcγRIIA (Bock and Gulbins, 2003; García-García et al., 2007).

T cell regulatory molecules have been the subject of extremely intense research. Data herein suggest caution when extrapolating results from specific experiments in model organisms, because a considerable portion of genetic diversity at these genes has accumulated not as a result of neutral processes but in response to adaptive events. As such, several interspecific changes are expected to entail functional effects, as shown in the case of the human and mouse TIM-3 orthologs.

Data herein indicate that adaptive change has also played a relevant role in shaping T cell regulatory molecule gene diversity in human populations. All the selective events we identified in humans target noncoding variants. In addition to the functional *FAS* polymorphism, we determined that the putative *HAVCR2* selection target is located in the 3′ UTR and affects a predicted binding site for miR-379. This microRNA is downregulated after $CD4^+$ T cell exposure to HIV-1 antigens (Bignami et al., 2012), and expression of TIM-3 defines a population of exhausted T cells in chronic HIV-1 infection (Sakuishi et al., 2011). In general, TIM-3 has a crucial role in the development of T cell exhaustion in other chronic viral infections such as HCV (Sakuishi et al., 2011), suggesting that it might be targeted by several pathogens as a strategy to downmodulate host response.

A likely regulatory role on gene expression can also be envisaged for the putative selection target in *CD274*, as well as for the variants defining the *CD40LG* and *FASLG* selected haplotypes. The signatures we detected at these loci were mainly geographically restricted and in the case of *CD274* and *FASLG* we found

the selected variants to correlate with pathogen diversity, suggesting that they might confer resistance to one or more infectious agents. In fact, different pathogens, including mycobacteria, *Helicobacter pylori*, and HIV-1, upregulate the expression of CD274 to dampen or evade the host immune response (Khan et al., 2012). Signatures of local adaptation were also detected at *PDCD1* and haplotype analysis revealed that the detected signatures are accounted for by the presence of a haplotype clade that shares many alleles with the Neandertal sequence. Neandertals and Denisovans are genetically more similar to contemporary European and Asian populations than to Africans (Green et al., 2010; Reich et al., 2010), and gene flow from archaic hominins to humans occurred at other loci, including *STAT2* and *HLAB* (Abi-Rached et al., 2011; Mendez et al., 2012). Analysis of the *PDCD1* region indicates that: (1) the TMRCA of the N and non-N clades are similar and much deeper than the time when introgression is expected to have occurred (37,000 to 86,000 years ago) (Sankararaman et al., 2012); also, the TMRCA of the whole genealogy is relatively recent; (2) nucleotide diversity is lower for N compared to non-N haplotypes but not dramatically so, whereas strong reduction of diversity would be expected for a recently acquired haplotype; and (3) the LD pattern is not unusual and the N and non-N haplotypes have similar haplotype homozygosity. We note that TMRCA inferences should be taken with caution, and no extensive LD might be expected in a subtelomeric region, where *PDCD1* is located. Nonetheless, these observations do not support the Neandertal-to-human introgression hypothesis for *PDCD1*. However, introgression may have occurred from human populations to Neandertals (in line with the observation that the Neandertal and Denisova sequences are quite divergent in the region we analyzed). In this case, humans would contribute genomic regions that are expected to have TMRCAs solely dependent on human history. Thus, balancing selection might have maintained the two *PDCD1* haplotype clades in Asian populations, due to some local selective pressure, and the N haplotype might have introgressed Neandertal populations through hybridization. It is presently impossible to establish how gene flow occurred (e.g., a single episode, multiple episodes, or continuous gene flow) (Sankararaman et al., 2012).

We also addressed the question of whether adaptive events at T cell regulatory molecule genes have affected the spread of human disease alleles. We found variants for human diseases to be preferential targets of pathogen-driven selection. The disease variants we describe herein as selection targets are located in noncoding regions, in line with regulation of gene expression being a major determinant of phenotypic variation and a common target of adaptive evolution (Lappalainen and Dermitzakis, 2010). Specifically, our data show that, whereas all autoimmune risk alleles correlate positively with pathogen diversity, the opposite situation is observed for IgA deficiency, suggesting that (1) risk alleles for autoimmunity confer higher protection against infections, and (2) their spread in human populations results from adaptation. We provided a demonstration of (1) by showing that the IBD risk allele (rs762421), which correlates with bacterial diversity, indeed increased the expression of *ICOSLG* mRNA in response to a bacterial superantigen. Moreover, the observation that the IgA deficiency allele is more common where pathogen load is low supports the validity of our approach, because IgA-deficient subjects are more susceptible to recurrent bacterial infections and to giardiasis (Yel, 2010). As for the second point, three out of seven variants displaying signatures of pathogen-driven selection are located in nonneutrally evolving regions, as assessed by different tests. Therefore, these data expand previous observations (Fumagalli et al., 2009; Zhernakova et al., 2010) indicating adaptation to infection as the underlying explanation for the maintenance of a set of autoimmune risk alleles in human populations.

EXPERIMENTAL PROCEDURES

Evolutionary Analysis in Mammals

Most mammalian sequences were retrieved from the Ensembl website (http://www.ensembl.org/index.html). The chimpanzee sequences for *CD40* and *PDCD1* were reconstructed through BLAST search of Trace Archives and sequencing of a *Pan troglodytes* individual (see below), respectively. DNA alignments were performed with the RevTrans 2.0 utility (Wernersson and Pedersen, 2003). Alignment uncertainties were removed by trimAL (automated1 mode) (Capella-Gutiérrez et al., 2009). We selected models of amino acid substitution and constructed phylogenetic trees with ProtTest3 (Abascal et al., 2005). GARD and MEME analyses were performed through the DataMonkey sever (http://www.datamonkey.org). Further details on evolutionary analyses are given in Supplemental Experimental Procedures.

DNA Samples, Sequencing, and Population Genetic Analyses

Human genomic DNA from HapMap subjects (20 Yoruba [YRI], 20 European [CEU], and 20 Asians [AS]) was obtained from the Coriell Institute for Medical Research. The genomic DNA of a *Pan troglodytes* was obtained from the Gene Bank of Primates, Germany. Details on Sanger sequencing and on the analysis of the Neandertal and Denisova sequences are available as Supplemental Experimental Procedures. Data from the Pilot 1 phase of the 1000 Genomes Project were retrieved from the dedicated website (http://www.1000genomes.org/) (1000 Genomes Project Consortium et al., 2010). All details on population genetic analyses are available in Supplemental Experimental Procedures.

HGDP-CEPH Panel Data and Pathogen-Driven Selection

F_{ST} was calculated for all HGDP-CEPH variants among continental groups; F_{ST} distributions were calculated for MAF (minor allele frequency)-matched SNP classes; outliers were defined as variants with an F_{ST} higher than the 95th percentile in the distribution of SNPs in the same MAF class.

The approach used to identify variants selected by different pathogen species has been described elsewhere (Fumagalli et al., 2009) and is briefly summarized in the Supplemental Experimental Procedures. We considered a SNP to be significantly associated with pathogen diversity if it displayed a significant correlation and a τ rank higher than 0.95.

SEB Stimulation and *ICOSLG* Transcript Quantification

Peripheral blood mononuclear cells (PBMCs) from 18 volunteers (age 22–28 years) were stimulated with SEB; quantification of the *ICOSLG* and *GAPDH* transcripts was performed by real-time PCR (Supplemental Experimental Procedures). The study was reviewed and approved by the institutional review board of the Scientific Institute IRCCS E. Medea.

ACCESSION NUMBERS

The GenBank accession number for the chimpanzee *PDCD1* gene sequence reported in this paper is KC535541.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, and eight tables and can be found with this article online at http://dx.doi.org/10.1016/j.immuni.2013.04.008.

ACKNOWLEDGMENTS

This work was supported by the Broad Medical Research Program of The Broad Foundation (grant IBD-0294). D.F. is supported by a fellowship of the Doctorate School of Molecular Medicine, University of Milan.

Received: October 2, 2012
Accepted: April 23, 2013
Published: May 23, 2013

REFERENCES

Abascal, F., Zardoya, R., and Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. Bioinformatics *21*, 2104–2105.

1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

Abi-Rached, L., Jobin, M.J., Kulkarni, S., McWhinnie, A., Dalva, K., Gragert, L., Babrzadeh, F., Ghazidadeh, 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.

An, H.J., Kim, Y.J., Song, D.H., Park, B.S., Kim, H.M., Lee, J.D., Paik, S.G., Lee, J.O., and Lee, H. (2011). Crystallographic and mutational analysis of the CD40-CD154 complex and its implications for receptor activation. J. Biol. Chem. *286*, 11226–11235.

Anisimova, M., and Liberles, M.A. (2012). Detecting and understanding natural selection. In Codon Evolution: Mechanisms and Models, G. Cannarozzi and A. Schneider, eds. (Oxford: Oxford University Press), pp. 73–96.

Anisimova, M., Bielawski, J.P., and Yang, Z. (2002). Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol. Biol. Evol. *19*, 950–958.

Anisimova, M., Nielsen, R., and Yang, Z. (2003). Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics *164*, 1229–1236.

Bajorath, J. (1999). Identification of the ligand binding site in Fas (CD95) and analysis of Fas-ligand interactions. Proteins *35*, 475–482.

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.

Bignami, F., Pilotti, E., Bertoncelli, L., Ronzi, P., Gulli, M., Marmiroli, N., Magnani, G., Pinti, M., Lopalco, L., Mussini, C., et al. (2012). Stable changes in CD4+ T lymphocyte miRNA expression after exposure to HIV-1. Blood *119*, 6259–6267.

Bock, J., and Gulbins, E. (2003). The transmembranous domain of CD40 determines CD40 partitioning into lipid rafts. FEBS Lett. *534*, 169–174.

Cao, E., Zang, X., Ramagopal, U.A., Mukhopadhaya, A., Fedorov, A., Fedorov, E., Zencheck, W.D., Lary, J.W., Cole, J.L., Deng, H., et al. (2007). T cell immunoglobulin mucin-3 crystal structure reveals a galectin-9-independent ligand-binding surface. Immunity *26*, 311–321.

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973.

Chen, L. (2004). Co-inhibitory molecules of the B7-CD28 family in the control of T-cell immunity. Nat. Rev. Immunol. *4*, 336–347.

Chiba, S., Baghdadi, M., Akiba, H., Yoshiyama, H., Kinoshita, I., Dosaka-Akita, H., Fujioka, Y., Ohba, Y., Gorman, J.V., Colgan, J.D., et al. (2012). Tumor-infiltrating DCs suppress nucleic acid-mediated innate immune responses through interactions between the receptor TIM-3 and the alarmin HMGB1. Nat. Immunol. *13*, 832–842.

DeKruyff, R.H., Bu, X., Ballesteros, A., Santiago, C., Chim, Y.L., Lee, H.H., Karisola, P., Pichavant, M., Kaplan, G.G., Umetsu, D.T., et al. (2010). T cell/transmembrane, Ig, and mucin-3 allelic variants differentially recognize phosphatidylserine and mediate phagocytosis of apoptotic cells. J. Immunol. *184*, 1918–1930.

Dermody, T.S., Kirchner, E., Guglielmi, K.M., and Stehle, T. (2009). Immunoglobulin superfamily virus receptors and the evolution of adaptive immunity. PLoS Pathog. *5*, e1000481.

Elgueta, R., Benson, M.J., de Vries, V.C., Wasiuk, A., Guo, Y., and Noelle, R.J. (2009). Molecular mechanism and function of CD40/CD40L engagement in the immune system. Immunol. Rev. *229*, 152–172.

Farre, L., Bittencourt, A.L., Silva-Santos, G., Almeida, A., Silva, A.C., Decanine, D., Soares, G.M., Alcantara, L.C., Jr., Van Dooren, S., Galvão-Castro, B., et al. (2008). Fas 670 promoter polymorphism is associated to susceptibility, clinical presentation, and survival in adult T cell leukemia. J. Leukoc. Biol. *83*, 220–222.

Fay, J.C., and Wu, C.I. (2000). Hitchhiking under positive Darwinian selection. Genetics *155*, 1405–1413.

Fu, Y.X., and Li, W.H. (1993). Statistical tests of neutrality of mutations. Genetics *133*, 693–709.

Fumagalli, M., Pozzoli, U., Cagliani, R., Comi, G.P., Riva, S., Clerici, M., Bresolin, N., and Sironi, M. (2009). Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions. J. Exp. Med. *206*, 1395–1408.

García-García, E., Brown, E.J., and Rosales, C. (2007). Transmembrane mutations to FcgammaRIIA alter its association with lipid rafts: implications for receptor signaling. J. Immunol. *178*, 3048–3058.

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.

Hudson, R.R., Kreitman, M., and Aguadé, M. (1987). A test of neutral molecular evolution based on nucleotide data. Genetics *116*, 153–159.

Ikemizu, S., Gilbert, R.J., Fennelly, J.A., Collins, A.V., Harlos, K., Jones, E.Y., Stuart, D.I., and Davis, S.J. (2000). Structure and dimerization of a soluble form of B7-1. Immunity *12*, 51–60.

Kajikawa, M., Li, P.C., Goto, E., Miyashita, N., Aoki-Kawasumi, M., Mito-Yoshida, M., Ikegaya, M., Sugita, Y., and Ishido, S. (2012). The intertransmembrane region of Kaposi's sarcoma-associated herpesvirus modulator of immune recognition 2 contributes to B7-2 downregulation. J. Virol. *86*, 5288–5296.

Kanemitsu, S., Ihara, K., Saifddin, A., Otsuka, T., Takeuchi, T., Nagayama, J., Kuwano, M., and Hara, T. (2002). A functional polymorphism in fas (CD95/APO-1) gene promoter associated with systemic lupus erythematosus. J. Rheumatol. *29*, 1183–1188.

Khan, N., Gowthaman, U., Pahari, S., and Agrewala, J.N. (2012). Manipulation of costimulatory molecules by intracellular pathogens: veni, vidi, vicil! PLoS Pathog. *8*, e1002676.

Kosakovsky Pond, S.L., Posada, D., Gravenor, M.B., Woelk, C.H., and Frost, S.D. (2006). Automated phylogenetic detection of recombination using a genetic algorithm. Mol. Biol. Evol. *23*, 1891–1901.

Lappalainen, T., and Dermitzakis, E.T. (2010). Evolutionary history of regulatory variation in human populations. Hum. Mol. Genet. *19* (R2), R197–R203.

Lázár-Molnár, E., Yan, Q., Cao, E., Ramagopal, U., Nathenson, S.G., and Almo, S.C. (2008). Crystal structure of the complex between programmed death-1 (PD-1) and its ligand PD-L2. Proc. Natl. Acad. Sci. USA *105*, 10483–10488.

Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., and Myers, R.M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science *319*, 1100–1104.

Lin, D.Y., Tanaka, Y., Iwasaki, M., Gittis, A.G., Su, H.P., Mikami, B., Okazaki, T., Honjo, T., Minato, N., and Garboczi, D.N. (2008). The PD-1/PD-L1 complex resembles the antigen-binding Fv domains of antibodies and T cell receptors. Proc. Natl. Acad. Sci. USA *105*, 3011–3016.

Lu, L.F., Cook, W.J., Lin, L.L., and Noelle, R.J. (2003). CD40 signaling through a newly identified tumor necrosis factor receptor-associated factor 2 (TRAF2) binding site. J. Biol. Chem. *278*, 45414–45418.

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.

Mansouri, M., Bartee, E., Gouveia, K., Hovey Nerenberg, B.T., Barrett, J., Thomas, L., Thomas, G., McFadden, G., and Früh, K. (2003). The PHD/LAP-domain protein M153R of myxomavirus is a ubiquitin ligase that induces the rapid internalization and lysosomal destruction of CD4. J. Virol. *77*, 1427–1440.

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.

Morio, T., Hanissian, S.H., Bacharier, L.B., Teraoka, H., Monoyama, S., Seki, M., Kondo, J., Nakano, H., Lee, S.K., Geha, R.S., and Yata, J. (1999). Ku in the cytoplasm associates with CD40 in human B cells and translocates into the nucleus following incubation with IL-4 and anti-CD40 mAb. Immunity *11*, 339–348.

Murrell, B., Wertheim, J.O., Moola, S., Weighill, T., Scheffler, K., and Kosakovsky Pond, S.L. (2012). Detecting individual sites subject to episodic diversifying selection. PLoS Genet. *8*, e1002764.

Nagae, M., Nishi, N., Murata, T., Usui, T., Nakamura, T., Wakatsuki, S., and Kato, R. (2006). Crystal structure of the galectin-9 N-terminal carbohydrate recognition domain from *Mus musculus* reveals the basic mechanism of carbohydrate recognition. J. Biol. Chem. *281*, 35884–35893.

Nei, M., and Li, W.H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA *76*, 5269–5273.

Nurieva, R.I., Liu, X., and Dong, C. (2009). Yin-Yang of costimulation: crucial controls of immune tolerance and function. Immunol. Rev. *229*, 88–100.

Pullen, S.S., Dang, T.T., Crute, J.J., and Kehry, M.R. (1999). CD40 signaling through tumor necrosis factor receptor-associated factors (TRAFs). Binding site specificity and activation of downstream pathways by distinct TRAFs. J. Biol. Chem. *274*, 14246–14254.

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.

Richter, C., Messerschmidt, S., Holeiter, G., Tepperink, J., Osswald, S., Zappe, A., Branschädel, M., Boschert, V., Mann, D.A., Scheurich, P., and Krippner-Heidenreich, A. (2012). The tumor necrosis factor receptor stalk regions define responsiveness to soluble versus membrane-bound ligand. Mol. Cell. Biol. *32*, 2515–2529.

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.

Sakuishi, K., Jayaraman, P., Behar, S.M., Anderson, A.C., and Kuchroo, V.K. (2011). Emerging Tim-3 functions in antimicrobial and tumor immunity. Trends Immunol. *32*, 345–349.

Sankararaman, S., Patterson, N., Li, H., Pääbo, S., and Reich, D. (2012). The date of interbreeding between Neandertals and modern humans. PLoS Genet. *8*, e1002947.

Schmidl, C., Klug, M., Boeld, T.J., Andreesen, R., Hoffmann, P., Edinger, M., and Rehli, M. (2009). Lineage-specific DNA methylation in T cells correlates with histone methylation and enhancer activity. Genome Res. *19*, 1165–1174.

Scott, F.L., Stec, B., Pop, C., Dobaczewska, M.K., Lee, J.J., Monosov, E., Robinson, H., Salvesen, G.S., Schwarzenbacher, R., and Riedl, S.J. (2009). The Fas-FADD death domain complex structure unravels signalling by receptor clustering. Nature *457*, 1019–1022.

Stamper, C.C., Zhang, Y., Tobin, J.F., Erbe, D.V., Ikemizu, S., Davis, S.J., Stahl, M.L., Seehra, J., Somers, W.S., and Mosyak, L. (2001). Crystal structure of the B7-1/CTLA-4 complex that inhibits human immune responses. Nature *410*, 608–611.

Starling, G.C., Kiener, P.A., Aruffo, A., and Bajorath, J. (1998). Analysis of the ligand binding site in Fas (CD95) by site-directed mutagenesis and comparison with TNFR and CD40. Biochemistry *37*, 3723–3726.

Strasser, A., Jost, P.J., and Nagata, S. (2009). The many roles of FAS receptor signaling in the immune system. Immunity *30*, 180–192.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585–595.

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.

Wanderley, J.L., and Barcinski, M.A. (2010). Apoptosis and apoptotic mimicry: the *Leishmania* connection. Cell. Mol. Life Sci. *67*, 1653–1659.

Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. *7*, 256–276.

Wernersson, R., and Pedersen, A.G. (2003). RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res. *31*, 3537–3539.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.

Yel, L. (2010). Selective IgA deficiency. J. Clin. Immunol. *30*, 10–16.

Zhernakova, A., Elbers, C.C., Ferwerda, B., Romanos, J., Trynka, G., Dubois, P.C., de Kovel, C.G., Franke, L., Oosting, M., Barisani, D., et al.; Finnish Celiac Disease Study Group. (2010). Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. Am. J. Hum. Genet. *86*, 970–977.

3.2 An evolutionary analysis of antigen processing and presentation across different timescales reveals pervasive selection

OPEN ACCESS Freely available online

PLOS | GENETICS

An Evolutionary Analysis of Antigen Processing and Presentation across Different Timescales Reveals Pervasive Selection

Diego Forni[1☯], Rachele Cagliani[1☯], Claudia Tresoldi[1], Uberto Pozzoli[1], Luca De Gioia[2], Giulia Filippi[2], Stefania Riva[1], Giorgia Menozzi[1], Marta Colleoni[1], Mara Biasin[3], Sergio Lo Caputo[4], Francesco Mazzotta[4], Giacomo P. Comi[5], Nereo Bresolin[1,5], Mario Clerici[6,7], Manuela Sironi[1]*

1 Scientific Institute IRCCS E. MEDEA, Bioinformatics, Bosisio Parini, Italy, **2** Department of Biotechnology and Biosciences, University of Milan-Bicocca, Milan, Italy, **3** Department of Biomedical and Clinical Sciences, University of Milan, Milan, Italy, **4** Infectious Disease Unit, S. Maria Annunziata Hospital, Florence, Italy, **5** Dino Ferrari Centre, Department of Physiopathology and Transplantation, University of Milan, Fondazione Ca' Granda IRCCS Ospedale Maggiore Policlinico, Milan, Italy, **6** Chair of Immunology, Department of Physiopathology and Transplantation, University of Milan, Milan, Italy, **7** Don C. Gnocchi Foundation ONLUS, IRCCS, Milan, Italy

Abstract

The antigenic repertoire presented by MHC molecules is generated by the antigen processing and presentation (APP) pathway. We analyzed the evolutionary history of 45 genes involved in APP at the inter- and intra-species level. Results showed that 11 genes evolved adaptively in mammals. Several positively selected sites involve positions of fundamental importance to the protein function (e.g. the TAP1 peptide-binding domains, the sugar binding interface of langerin, and the CD1D trafficking signal region). In CYBB, all selected sites cluster in two loops protruding into the endosomal lumen; analysis of missense mutations responsible for chronic granulomatous disease (CGD) showed the action of different selective forces on the very same gene region, as most CGD substitutions involve aminoacid positions that are conserved in all mammals. As for ERAP2, different computational methods indicated that positive selection has driven the recurrent appearance of protein-destabilizing variants during mammalian evolution. Application of a population-genetics phylogenetics approach showed that purifying selection represented a major force acting on some APP components (e.g. immunoproteasome subunits and chaperones) and allowed identification of positive selection events in the human lineage. We also investigated the evolutionary history of APP genes in human populations by developing a new approach that uses several different tests to identify the selection target, and that integrates low-coverage whole-genome sequencing data with Sanger sequencing. This analysis revealed that 9 APP genes underwent local adaptation in human populations. Most positive selection targets are located within noncoding regions with regulatory function in myeloid cells or act as expression quantitative trait loci. Conversely, balancing selection targeted nonsynonymous variants in *TAP1* and *CD207* (langerin). Finally, we suggest that selected variants in *PSMB10* and *CD207* contribute to human phenotypes. Thus, we used evolutionary information to generate experimentally-testable hypotheses and to provide a list of sites to prioritize in follow-up analyses.

Citation: Forni D, Cagliani R, Tresoldi C, Pozzoli U, De Gioia L, et al. (2014) An Evolutionary Analysis of Antigen Processing and Presentation across Different Timescales Reveals Pervasive Selection. PLoS Genet 10(3): e1004189. doi:10.1371/journal.pgen.1004189

Editor: David D. Pollock, University of Colorado School of Medicine, United States of America

Received September 5, 2013; **Accepted** January 6, 2014; **Published** March 27, 2014

Funding: This work was supported by the Broad Medical Research Program of The Broad Foundation (grant IBD-0294). DF was supported by a fellowship of the Doctorate School of Molecular Medicine, University of Milan. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: manuela.sironi@bp.lnf.it

☯ These authors equally contributed to this work.

Introduction

Cell mediated immune responses are initiated by the recognition of an MHC/antigen complex on the surface of an APC (antigen presenting cell) by a T cell receptor (TcR). MHC class I and II molecules present peptides to T cells that express the CD8 or CD4 molecules, respectively.

Non-conventional T cell populations also exist that express TcRs with semi-invariant α-chains: MAIT (mucosal-associated invariant T) cells recognize antigens bound to the class Ib MHC molecule MR1, and iNKT (invariant natural killer T) cells respond to lipids and glycolipid antigens bound to CD1D.

Whatever the nature of the presenting molecule, the limited dimension of its cleft makes it impossible for macromolecules to be presented: only fragments deriving from the lysis of such molecules will be nested in the cleft. Most steps leading to the formation of MHC class I- and II-peptide complexes have been defined [1]. Peptides that will be embedded into the cleft of class I molecules are initially processed by the proteasome, a complex structure located in the cytoplasm. Immune cells and other cell types exposed to interferon gamma express a variant of the proteasome referred to as the immunoproteasome and differing in a few subunit components (Figure 1) [1]. The proteasome activity can be complemented in the cytosol by endopeptidases (Figure 1) [1,2].

Author Summary

Antigen-presenting cells digest intracellular and extracellular proteins and display the resulting antigenic repertoire on cell surface molecules for recognition by T cells. This process initiates cell-mediated immune responses and is essential to detect infections. The antigenic repertoire is generated by the antigen processing and presentation pathway. Because several pathogens evade immune recognition by hampering this process, genes involved in antigen processing and presentation may represent common natural selection targets. Thus, we analyzed the evolutionary history of these genes during mammalian evolution and in the more recent history of human populations. Evolutionary analyses in mammals indicated that positive selection targeted a very high proportion of genes (24%), and revealed that many selected sites affect positions of fundamental importance to the protein function. In humans, we found different signatures of natural selection acting both on regions that are expected to regulate gene expression levels or timing and on coding variants; two human selected polymorphisms may modulate the susceptibility to Crohn's disease and to HIV-1 infection. Therefore, we provide a comprehensive evolutionary analysis of antigen processing and we show that evolutionary studies can provide useful information concerning the location and nature of functional variants, ultimately helping to clarify phenotypic differences between and within species.

Channels formed by TAP molecules (TAP1 and TAP2) allow peptides generated in the cytoplasm to be transported into the endoplasmic reticulum (ER), where they may be trimmed at their N-terminal end by ERAP proteins. In the ER, MHC class I are bound to the TAP complex through tapasin (TAPBP), and they are further stabilized by two chaperones, calreticulin (CALR) and ERp57 (PDIA3) [1] (Figure 1). The whole complex is referred to as the peptide-loading complex (PLC). The peptide/MHC class I dimer will then bind a molecule of β2 microglobulin; this results in the stabilization of the complex that will be exported to the cell surface by an exocytic vescicle [1].

MHC class II molecules wait for the proper peptide in endosomes; these will fuse with lysosomes where the exogenous proteins have been processed by resident proteases (Figure 1). The removal of the CD74-derived invariant DM peptide by cathepsin S or L (CTSS, CTSL1) from the cleft of the MHC molecule will render it available to the incoming peptides. The resulting MHC/peptide complexes will then be exported to the cell surface by endosomes [1].

Finally, in cross-presentation phagocytosed antigens are partially degraded, exported to the cytoplasm for further processing, and then loaded onto MHC class I molecules. A central role in this process is played by the superoxide-producing phagocyte NADPH-oxidase, a multiprotein complex (Figure 1) which regulates alkalinization of the phagosomal lumen [3].

Classic MHC molecules are encoded by genes that show extreme levels of polymorphism in most vertebrates and several studies have demonstrated that diversity at the peptide binding region is maintained by natural selection [4]. Thus, their role in adaptive immunity and their pattern of diversity indicate adaptation to a wide range of pathogen species leading to aminoacid diversification of the antigen binding cleft. Nonetheless, the generation and loading of the antigenic repertoire presented by MHC molecules also depend on the action of a number of molecules, as detailed above. Therefore, it is straightforward to imagine that a proportion of these should be targeted by natural selection, as well. The observation whereby several pathogens encode molecules that hijack specific components of the antigen processing and presentation (APP) pathway further supports this possibility [5]. Herein, we investigated the evolutionary history of 45 genes with a central role in APP by analyzing inter-specific divergence in mammals and intra-specific diversity in human populations.

Results

Several APP genes evolved adaptively in mammals

To analyze the evolutionary history of the APP pathway, we compiled a list of 45 genes that play roles of central importance in this process. Specifically, based on Gene Ontology classification, we included genes involved in the processing of both endogenous and exogenous antigens and in the presentation via class I, class II or class Ib MHC molecules (see methods for details of gene selection criteria) (Figure 1, Supplementary Table S1). Because they have already been the topic of extensive investigation, *HLA* genes were not included. Moreover, genes involved in APP, but also in general cellular processes (e.g. components of the constitutive proteasome, genes involved in vesicle trafficking) were not analyzed.

First we analyzed the evolutionary history of these genes in mammals by retrieving coding sequence information for all available species. For *CTSL1* and *CTSL2* only primate sequences were included because the two genes originated from a relatively recent duplication event (which occurred before the split of modern primates) and, due to their high similarity, it is very difficult to establish one-to-one orthology with more distantly related mammals.

Analysis of sequence alignments revealed that all genes evolved under purifying selection, as the average non-synonymous substitution rate (dN) was generally lower than the rate for synonymous substitutions (dS) (Supplementary Table S2). Yet, positive selection can operate on specific residues or domains within coding regions that are otherwise selectively constrained. To test this possibility we applied maximum-likelihood analyses by comparing models of gene evolution that allow (NSsite models M2a and M8) or disallow (NSsite models M1a, and M7) a class of codons to evolve with dN/dS>1 [6]. After accounting for the presence of recombination (that might yield false positive results [7]) and using different models of codon frequency (see Materials and Methods and Supplementary Figure S1), eleven APP genes (*BLMH*, *CD1D*, *CD207*, *CTSL2*, *CTSG*, *CYBB*, *ERAP2*, *LNPEP*S, *TAPBP*, *TAPBPL*, and *TAP1*) were found to evolve adaptively in mammals (Table 1, Figure 1, Supplementary Table S3 and S4). To identify specific sites subject to positive selection, we applied two methods: the Bayes Empirical Bayes (BEB) analysis (with a cut-off of 0.90) from M8 [8], and the Mixed Effects Model of Evolution (MEME) (with the default cutoff of 0.1) [9]. Only sites detected using both methods were considered and these are listed in Table 1.

In order to explore possible variations in selective pressure among different lineages, we used the branch site-random effects likelihood (BS-REL) method [10], which was applied to the 45 APP gene alignments or to sub-regions (alignments were split on the basis of recombination breakpoint location). BS-REL makes no a priori assumption about which lineages are more likely to represent selection targets. We focused our attention on genes showing evidences of episodic positive selection in lineages that include the human species (i.e. the human lineage or branches leading to great apes) or in branches leading to murids (due to the

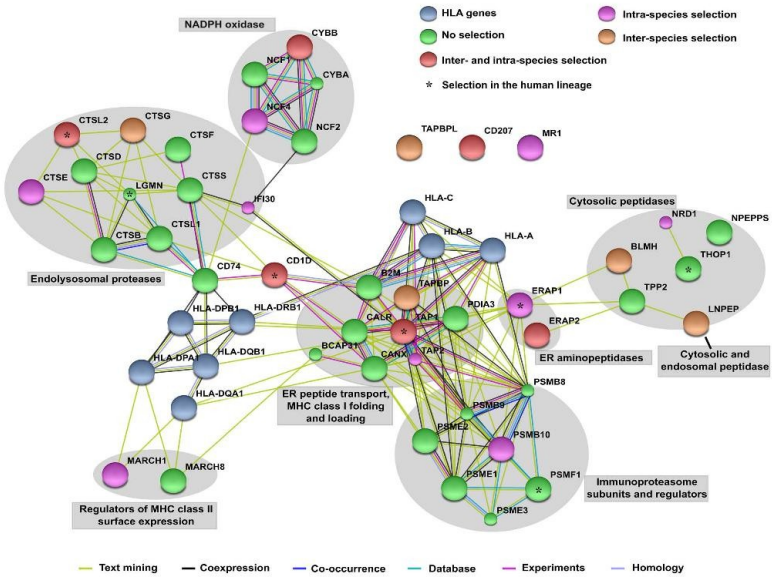


Figure 1. STRING interaction diagram of the analyzed genes. Classic *HLA* class I and class II genes are also shown (although not analyzed). Each filled node denotes a gene; edges between nodes indicate interactions between protein products of the corresponding genes. Different edge colors represent the types of evidence for the association. Annotation of genes and gene clusters refers to their major activity/location in APP. Genes are colored according to the observed selection signatures either described herein or in previous works [33,43,44].
doi:10.1371/journal.pgen.1004189.g001

relevance these species have as model organisms). Thus, three genes were selected for further analysis: *CD207*, *CTSG*, and *CYBB* (Figure 2 and Supplementary Figure S2). For these alignments, the primate/murid branches detected by BS-REL were cross-validated using the branch-site models implemented in PAML [11], which apply a likelihood ratio test to compare a model (MA) that allows positive selection on one or more lineages (foreground lineages) with a model (MA1) that does not allow such positive selection. As suggested [12], a false discovery rate (FDR) correction was applied to these p values, as multiple hypotheses are being tested on the same phylogeny. As shown in figure 2, PAML confirmed episodic positive selection at 1 and 2 branches in *CD207* and *CTSG*, respectively; no *CYBB* branch was validated by PAML (Supplementary Table S5, Supplementary Figure S2). The PAML branch-site models can identify specific sites that evolved under positive selection in the foreground branches; this is achieved through implementation of a BEB analysis, which is accurate but has low statistical power [11]. BEB analysis identified one positively selected site in CTSG (175I) on the lineage leading to simians.

In line with its ability to detect episodic positive selection, the MEME analysis performed on the whole phylogeny also detected the 175I residue in CTSG. Thus, episodic positive selection acted on *CTSG* and *CD207* in simians and murids, respectively.

Positively selected sites involve functional residues

We next analyzed the location of positively selected sites relative to known protein domains or crystal structures.

The extracellular portion of CD1D comprises two domains (α1 and α2) that form the antigen-binding groove and interact with the TcR, plus an α3 domain that interacts with B2M. All positively selected sites we identified in the extracellular portion of the protein are in the α1/α2 domains, and four of them cluster in a spatially defined region in the C' pocket; these positions are not directly involved in the binding of known antigens, and one of them flanks the TcR interaction surface (Figure 3A). One additional positively selected site was located in the short CD1D cytoplasmic tail, which carries signals essential for CD1D cellular trafficking. Specifically, the human 322T residue is essential for transportation to the plasma membrane [13].

CD207 encodes langerin, a C-type lectin that binds glycoconjugates and functions as a trimer [14,15]. The extracellular portion of the protein contains a carbohydrate-recognition domain (CRD) and a neck region that participates in trimer formation. The two

Table 1. Evolutionary analysis of mammalian/primate APP genes.

Gene (length in codons)[a]	N species[b]	N recombination breakpoints[c]	N significant regions[d]	Positively selected sites (human codons)[e]
BLMH (455)	39	2	1	211V, 388A, 390T
CD207 (329)	32	0	1	213P, 289A
CD1D (353)	28	1	2	25L, 108L, 136F, 139K, 157L, 161L, 302M, 322T
CTSG (255)	28	0	1	66W, 69N, 106Q, 122R, 177G, 221S
CTSL2 (334)	11	0	1	262S
CYBB (570)	38	2	1	136P, 148Q, 149N, 233A, 234E, 237A, 240N, 241I, 242T, 243V, 245E, 249S, 250E, 255K
ERAP2 (970)	26	2	1	416Y, 420V, 857A
LNPEP (1025)	38	1	1	872K, 884I, 918N, 1023W
TAP1 (777)	35	1	2	R137, E145, G225, Q516, L557, L562
TAPBP (468)	33	0	1	67S, 225N
TAPBPL (438)	32	0	1	394G, 433T

[a]Only genes subject to positive selection (see text) are shown.
[b]Number of species in the alignment.
[c]Number of recombination breakpoints from GARD.
[d]Number of gene regions showing evidences of positive selection (see text).
[e]Positively selected sites identified by both BEB and MEME.
doi:10.1371/journal.pgen.1004189.t001

positively selected sites are located in the CRD domain; one of them (289A) is directly involved in Ca^{+} mediated carbohydrate binding [14] (Figure 3C); the other site (213P) immediately flanks residues that contribute to the interaction among langerin subunits forming the trimer. The W264R mutation in *CD207* has been associated with Birbeck granule deficiency [16] and 264W is conserved in all mammals (Figure 3C).

The *CYBB* gene encodes an integral membrane protein that functions as the catalytic subunit of the phagocyte NADPH oxidase. Because the crystallographic structure of CYBB has not been solved, we mapped selected sites onto the membrane topology arrangement [17]: results indicated that all sites are located in extracellular/phagosome lumenal loops; specifically several sites cluster in the third loop and one of these (240N) affects a glycosylation site [18]. *CYBB* mutations are responsible for X-linked chronic granulomatous disease (CGD) [19] and for mendelian inheritance to mycobacterial diseases (MSMD) [20]; analysis of MSMD and CGD missense mutations located in the region where the positively selected sites were detected indicated that they all affect extremely conserved positions (Figure 3D).

TAP1 and TAPBP (tapasin) are part of the PLC. TAP1 belongs to the family of ATP binding cassette (ABC) transporters and its membrane topology has been determined [21]. Three of the positively selected sites we identified are located in the transmembrane region or cytoplasmic loops of the TAP1 unique N-terminal domain that is involved in the binding of tapasin (TAPBP) [22]. Interestingly, three additional sites subject to diversifying selection are located within or very close to the pore-forming region of TAP1 - i.e. the region responsible for peptide binding and transportation (Figure 3E) [23]. As for tapasin, one of the two positively selected sites is directly involved in ERp57 binding (225N) [24] and the second one is located at the N-terminus (67S) (Figure 3F). The cystein residue involved in disulfide-bonding with PDIA3 is conserved in all eutheria but not in metatheria (Supplementary Figure S3).

Positively selected sites were also identified in two cathepsin family members whose crystal structure has been solved. In CTSG the six sites subject to pervasive positive selection are located within the serine protease domain and three of them immediately flank (66W and 221S) or overlap (177G) residues that define the substrate binding pockets [25] (Figure 3B). This also applies to the 175I residue, targeted by positive selection in the simian lineage (Figure 3B).

As for CTSL2, one positively selected site was found in the protease domain, outside the substrate binding pockets (Supplementary Figure S3).

LNPEP encodes leucyl/cystinyl aminopeptidase; the four positively selected sites were found to be located in the C-terminal domain 4, which has been shown to possess regulatory activity [26] (Supplementary Figure S3).

Three sites subject to diversifying selection were also detected in *BLMH*, which encodes a cytoplasmic cysteine protease highly conserved from yeast to mammals [27]. One of them (211V) is located on an exposed α-helix (Figure 3G); the other two sites are on an unstructured loop and immediately flank a lysine residue (391K) which undergoes acetylation and ubiquitination [28,29]. The modified lysine and most aminoacids in the region are highly conserved, including a phenylalanine at position −2 relative to 391K that is present in all eutheria (Figure 3G) and represents a highly preferred residue in cytosolic acetylation sites [29].

Finally, in ERAP2 we identified three positively selected sites, which seem not be involved in proteolytic activity. 3D-structure protein analysis indicated that the three residues are located on α helices shaping the internal cavity of the protein where the catalytic Zn ion is coordinated (Figure 4A). Two of these residues are involved in several short-range interactions: 416Y can interact hydrofobically with 362L, 413F, 746W, and 420V (and *vice-versa*); the same kind of interactions can be made by 420V with 417F (not shown); a side-chain side-chain H-bond can be formed by the OH group of 416Y and the NH_2 group of 366R (not shown). Thus, we performed a stability analysis: 416Y and 420V were mutated to all other residues through the use of three different methods. The tyrosine and valine at positions 416 and 420 are the most common aminoacids among the species we analyzed (Figure 4B) and represent the ancestral state residues (see Materials and methods). As shown in Figure 4C, the replacement of the two aminoacids led

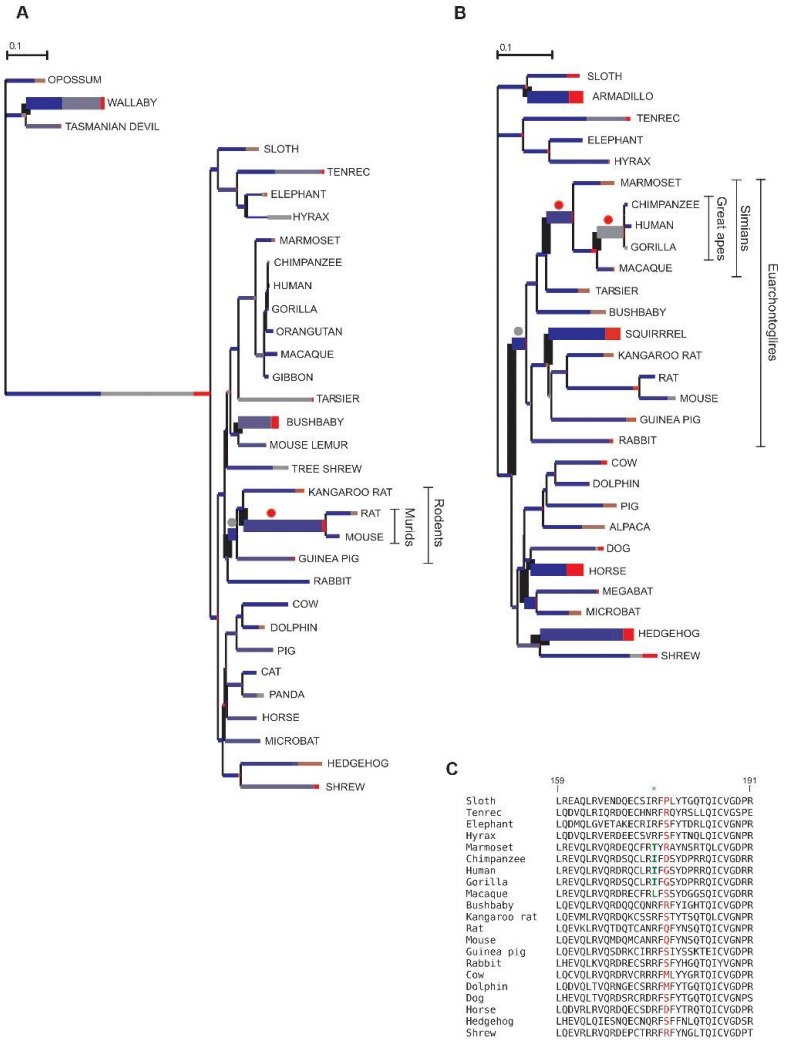

A
0.1
OPOSSUM
WALLABY
TASMANIAN DEVIL
SLOTH
TENREC
ELEPHANT
HYRAX
MARMOSET
CHIMPANZEE
HUMAN
GORILLA
ORANGUTAN
MACAQUE
GIBBON
TARSIER
BUSHBABY
MOUSE LEMUR
TREE SHREW
KANGAROO RAT
RAT
MOUSE
Murids
Rodents
GUINEA PIG
RABBIT
COW
DOLPHIN
PIG
CAT
PANDA
HORSE
MICROBAT
HEDGEHOG
SHREW
B
0.1
SLOTH
ARMADILLO
TENREC
ELEPHANT
HYRAX
MARMOSET
CHIMPANZEE
HUMAN
GORILLA
Great apes
Simians
MACAQUE
TARSIER
BUSHBABY
SQUIRRREL
KANGAROO RAT
RAT
MOUSE
GUINEA PIG
RABBIT
Euarchontoglires
COW
DOLPHIN
PIG
ALPACA
DOG
HORSE
MEGABAT
MICROBAT
HEDGEHOG
SHREW
C
159
191
Sloth LREAQLRVENDQECSIRFPLYTGQTQICVGDPR
Tenrec LQDVQLRIQRDQECHNRFRQYRSLLQICVGSPE
Elephant LQDMQLGVETAKECRIRFSFYTDRLQICVGNPR
Hyrax LQDVQLRVERDEECSVRFSFYTNQLQICVGNPR
Marmoset LREVQLRVQRDEQCFRTYRAYNSRTQLCVGDPR
Chimpanzee LREVQLRVQRDSQCLRIFDSYDPRRQICVGDRR
Human LREVQLRVQRDRQCLRIFGSYDPRRQICVGDRR
Gorilla LREVQLRVQRDSQCLRIFGSYDPRRQICVGDRR
Macaque LREVQLRVQRDRECFRLFSSYDGGSQICVGDRR
Bushbaby LQEVQLRVQRDQQCQNRFRFYIGHTQICVGDPR
Kangaroo rat LQEVMLRVQRDQKCSSRFSTYTSQTQLCVGNPR
Rat LQEVKLRVQTDQTCANRFQFYNSQTQICVGNPR
Mouse LQEVQLRVQMDQMCANRFQFYNSQTQICVGNPR
Guinea pig LQEVQLRVQSDRKCIRRFSIYSSKTEICVGDPR
Rabbit LHEVQLKVQRDRECSRRFSFYHGQTQIYVGNPR
Cow LQCVQLRVQRDRVCRRRFMLYYGRTQICVGDPR
Dolphin LQDVQLTVQRNGECSRRFMFYTGQTQICVGDPR
Dog LHEVQLTVQRDSRCRDRFSFYTGQTQICVGNPS
Horse LQDVRLRVQRDQECSDRFDFYTRQTQICVGDPR
Hedgehog LHEVQLQIESNQECNQRFSFFNLQTQICVGDSR
Shrew LQEVRLRVQRDEPCTRRFRFYNGLTQICVGDPT

Figure 2. Branch-site analysis of positive selection for *CD207* and *CTSG*. BS-REL analysis for *CD207* (A) and *CTSG* (B). Branch lengths are scaled to the expected number of substitutions per nucleotide, and branch colors indicate the strength of selection (dN/dS or ω). Red, positive selection (ω> 5); blue, purifying selection (ω = 0); grey, neutral evolution (ω = 1). The proportion of each color represents the fraction of the sequence undergoing the corresponding class of selection. Thick branches indicate statistical support for evolution under episodic diversifying selection as determined by BS-REL. Dots denote branches that were confirmed (red) or not (gray) to be under positive selection using the PAML branch-site models (after FDR correction for multiple tests). (C) Alignment of a portion of the CTSG peptidase domain for a few representative mammals showing positively selected residues in simians (green) and in the whole phylogeny (red).
doi:10.1371/journal.pgen.1004189.g002

to changes of different magnitude in ΔG. Although the three programs yielded different ΔΔG values for every mutated residues, the trend was maintained (in particular between I-Mutant and PoPMuSiC) and indicated that replacement of 416Y and 420V with any other aminoacid likely results in protein destabilization (i.e. positive ΔΔG values) (Figure 4C). These observations suggest that positive selection might have driven the recurrent appearance of destabilizing variants in ERAP2.

Different evolutionary scenarios for APP genes in the human lineage

We next applied a recently developed population genetics-phylogenetics approach to study the evolution of APP genes in the human species. Specifically, we used the gammaMap program [30], that jointly uses intra-specific variation and inter-specific diversity to estimate the distribution of fitness effects (DFE) (i.e. selection coefficients, γ) along coding regions. To this aim, we exploited data from the 1000 Genomes Pilot project deriving from the low-coverage whole genome sequencing of 179 individuals with different ancestry: Europeans (CEU), Yoruba from Nigeria (YRI), and East Asians (AS; Japanese plus Chinese) [31]. Ancestral sequences were reconstructed by parsimony from the human, chimpanzee, orangutan and macaque sequences. We noted that no human variant mapped to *NCF1* in CEU and AS. Inspection of accessibility by pair-end next generation sequencing approaches (see Materials and Methods) indicated that *NCF1* is poorly covered in the 1000 Genomes Project data, possibly because of the presence of segmental duplications. We thus discarded genes with less than 80% of accessible sequence; this resulted in the removal of *NCF1* and *NPEPPS*, which were excluded from further analyses.

We first analyzed the overall distribution of selection coefficients along the 43 APP genes. We observed a general preponderance of codons evolving under negative selection ($\gamma<0$) in all APP genes, with few exception including *CD1D*, *CD207*, *CTSG*, and *PSMF1* (Figure 5). The strongest level of negative selection was evident for genes encoding chaperones or proteins involved in MHC class I binding and transport, as well as for loci encoding immunoproteasome subunits. Likewise several endolysosomal proteases and peptidases located in the cytosol showed considerable levels of negative selection (Figure 5).

GammaMap also allows to identify specific codons evolving under positive selection. Herein we defined positively selected codons as those having a cumulative probability >0.80 of $\gamma\geq1$. Some of these residues had previously been identified in the positive selection analysis we conducted on the whole mammalian phylogeny (Table 2). For example, the 302M residue in CD1D had been detected by both MEME and BEB. Additional selected sites were identified in human CD1D. Among these, residue 200 is at the end of an α-helix that connects domains α1/α2 with α3; this position is occupied by a negatively charged aminoacid in all analyzed primates and mammals (not shown), but the human protein carries a lysine (Figure 5). Likewise, two of the positively selected sites in LGMN were also detected by MEME (Table 2): they are located in the activation peptide (which needs to be removed to generate catalytically active LGMN); in particular, 288R involves the alpha-cleavage site (^{287}KRK^{289}) [32] (Figure 5). In ERAP1 one of the positively selected sites (R528K, rs30187) has previously been described as a target of balancing selection in human populations [33] (Supplementary Figure S3). Analysis of TAP1 selected sites indicated that they are located in the tapasin binding region, where three sites positively selected in mammals are also observed (Figure 3). As for PSMF1, two positively selected sites map to the N-terminal PI31 proteasome regulator and flank a highly conserved motif important for protein structure [34] (Figure 5). Finally, in THOP1, one of the identified residues is an exposed cystein, which might be involved in multimerization [35] (Supplementary Figure S3).

Natural selection at APP genes is widespread in human populations

To investigate the evolutionary pattern of APP genes during the more recent history of human populations, we again exploited data from the 1000 Genomes Pilot project. A work-flow of the methods we applied is available as Supplementary Figure S4. Briefly, we integrated different neutrality tests that rely on distinct signatures left by natural selection. Thus, over whole gene regions we calculated: 1) θ_W [36] and π [37], which describe genetic diversity; 2) Tajima's D [38], normalized Fay and Wu's H [39], as well as Fu and Li's F* and D* [40], which represent site frequency spectrum (SFS)-based statistics. Also, for all SNPs located within APP genes we calculated F_{ST} [41], a measure of population genetic differentiation in pairwise comparisons (CEU/YRI, YRI/AS, and AS/CEU), and we performed the DIND (Derived Intra-allelic Nucleotide Diversity) test [42], which is based on haplotype homozygosity.

Because the low-coverage 1000 Genomes data suffer from a bias in the SFS [31], and in order to account for the influence of human demographic history, we applied an outlier approach by deriving empirical distributions of the same parameters calculated for a randomly selected set of human genes (see Materials and methods).

Analysis of θ_W and π for APP genes indicated that 8 of them had values higher than the 95th percentiles in at least one population (Supplementary Figure S5); after excluding *ERAP1*, *ERAP2*, and *TAP2*, which have previously been described as selection targets [33,43,44], these genes were considered as balancing selection candidates and were Sanger-resequenced, as detailed below.

For the remaining genes, we investigated whether they have been targets of selective sweeps. To minimize the identification of false positive signals, APP genes were considered targets of directional selection if they represented outliers (in the 5% tails of empirical distributions) in the same population for at least three parameters based on distinct signatures (e.g. F_{ST}, DIND and SFS-statistics) or in at least two parameters based on different features and both in the 1% tails of empirical distributions. Ten genes satisfied these criteria and for all of them analyses were extended to a 100 kb flanking region (50 kb up- and down-stream) to account for the large span of selective sweeps.

As detailed below, we combined multiple tests to identify the most likely selection target (i.e. the advantageous mutation

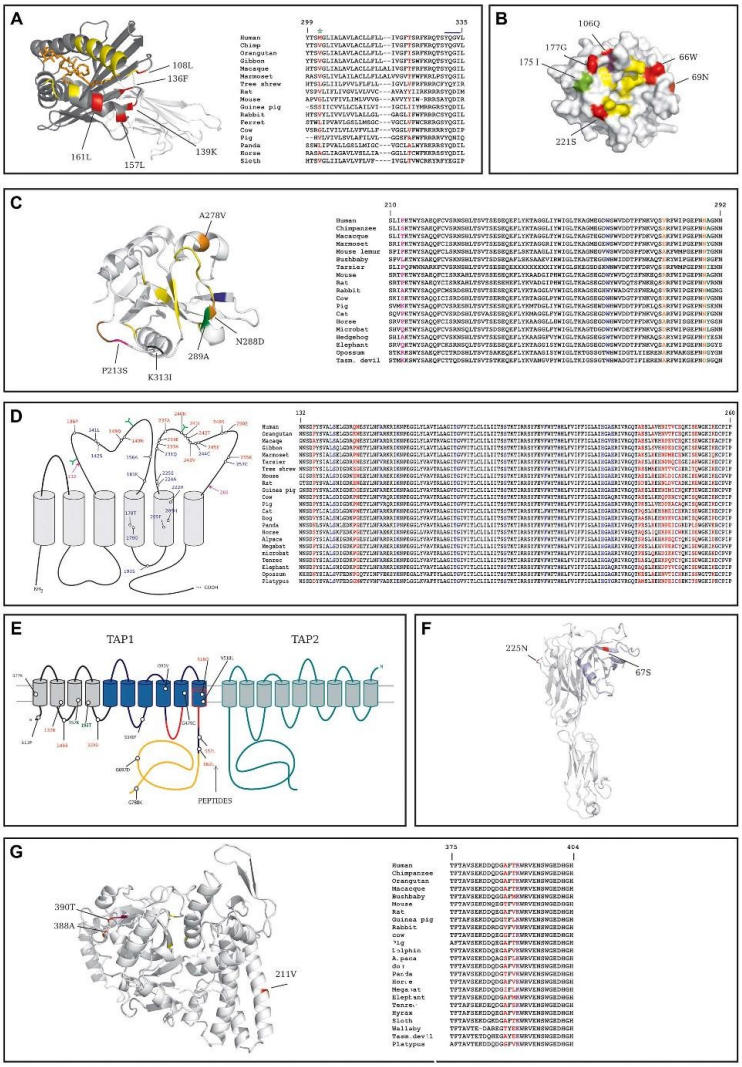

A
108L
136F
161L
157L
139K
B
106Q
177G
175I
66W
69N
221S
C
A278V
N288D
289A
P213S
K313I
D
COOH
E
TAP1
TAP2
PEPTIDES
F
225N
67S
G
390T
388A
211V

Figure 3. Analysis of positively selected sites. In all panels aminoacid numbering refers to the human protein. (A) Left: ribbon diagram of the extracellular domain of human CD1D bound to α-galactosylceramide (orange). Positively selected sites are shown in red, the α1/α2 and α3 domains are depicted in dark and light grey, respectively. Yellow residues form the contact interface with the TcR. Right: alignment of the transmembrane and cytoplasmic domains of CD1D for a few representative mammals; positively selected sites are in red and the YxxZ sequence is marked (blue line); the green asterisk denotes a site positively selected in the human lineage. (B) Surface structure of the protease domain of human CTSG; sites that define substrate binding pockets or form the catalytic triad are shown in yellow; positively selected sites are in red (whole phylogeny) and green (simians). The violet residue confers to CTSG the ability to cleave Shigella virulence factors if mutated. 126R is not visible as it is located on the back surface. (C) Left: ribbon diagram of the human CD207 CRD. Color codes are as follows: yellow, sites directly involved in sugar binding; green, positively selected site at the sugar binding interface; brown, sites involved in trimer formation; orange, nonsynonymous SNPs; magenta, positively selected site that is polymorphic in humans; black, missense SNP at the sugar binding interface; blue, a human mutation responsible for Birbeck granule deficiency. Right: alignment of a portion of the CRD for a few representative mammals; color codes are as in the left panel. (D) Positively selected sites for CYBB are shown relative to the membrane topology (left); sites subject to diversifying selection are in red, mutations responsible for CGD or MSMD are in blue (note that mutations are shown only if falling in the region where positively selected sites are located); glycosylation sites are represented in green; the magenta arrows denote the region which is represented in the multiple species alignment (right, color codes as in the membrane topology diagram). (E) Membrane topology arrangement and positively selected sites for TAP1; TAP2 (green profiled) is shown although no positively selected sites were identified. The TAP1 unique N-terminal domain is shown as grey cylinders, the ABC transporter domain is in blue; the nucleotide binding domain is in orange and the protein portions that bind peptides are profiled in red. Sites subject to diversifying selection are in red, human missense polymorphisms in black, positively selected sites in the human lineage are in green. (F) Ribbon diagram of human tapasin; positively selected sites are shown in red; the 87 N-terminal aminoacids that facilitate the folding of MHC I-peptide complexes are in light blue. (G) Left: ribbon diagram of human BLMH (one subunit of the hexameric complex is shown); positively selected sites are in red, the acetylated/ubiquitinated lysine (391K) is in violet, the catalytic triad in yellow. Right: alignment of the region surrounding 391K and two positively selected sites for a few representative mammals; color codes as in the left panel.
doi:10.1371/journal.pgen.1004189.g003

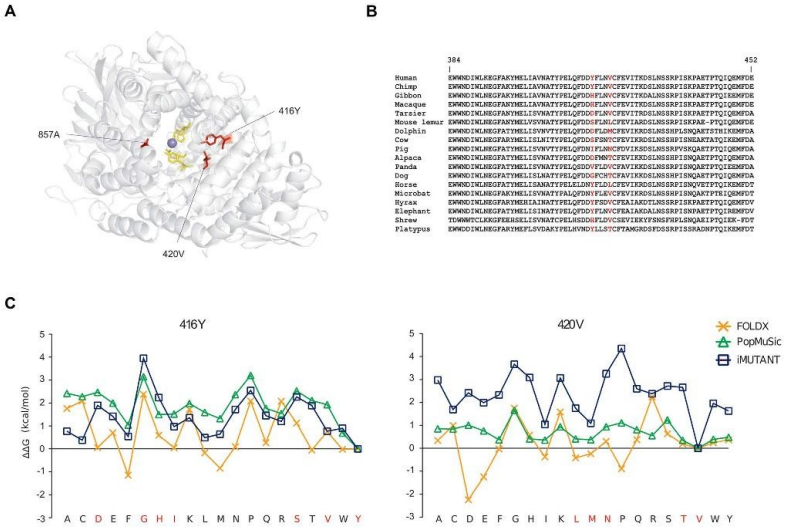


Figure 4. Analysis of positively selected sites in *ERAP2*. (A) Ribbon diagram of ERAP2: positively selected sites are shown in red and those that coordinate the Zn ion (violet) in yellow. (B) Alignment of the region surrounding 416Y and 420V (in red) for a few representative mammals. (C) ΔΔG in kcal/mol for 416Y (left), 420V (right) mutations to all other 19 residues of the ERAP2 structure or sequence; results are shown for FoldX, PopMuSiC, and I-Mutant.
doi:10.1371/journal.pgen.1004189.g004

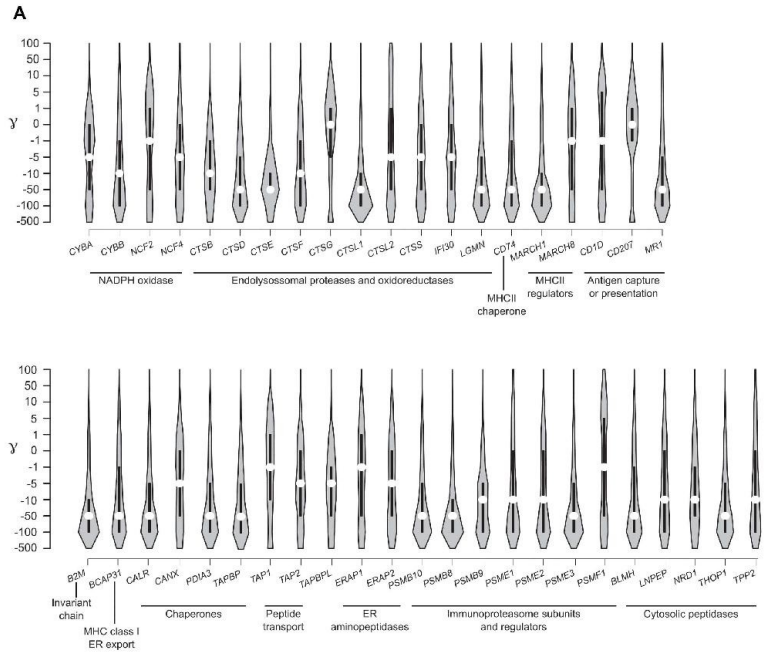


Figure 5. Analysis of selective pressure in the human lineage for APP genes. (A) Violin plot of selection coefficients for APP genes (median, white dot; interquartile range, black bar). Selection coefficients (γ) are classified as strongly beneficial (100, 50), moderately beneficial (10, 5), weakly beneficial (1), neutral (0), weakly deleterious (−1), moderately deleterious (−5, −10), strongly deleterious (−50, −100), and inviable (−500). (B) Ribbon diagram of CD1D; the α1/α2 and α3 domains are depicted in dark and light grey, respectively. Positively selected sites in humans are in green; in red sites selected in the whole phylogeny. (C) Ribbon diagram of LGMN with the activation peptide in dark grey. Human positively selected sites are in green.
doi:10.1371/journal.pgen.1004189.g005

underlying the sweep). Finally, we verified whether these signals could also be detected using other tests based on extended haplotype homozygosity [45,46].

Selective sweeps drove the frequency increase of regulatory polymorphisms in APP genes

Among genes coding for immunoproteasome-specific subunits, *PSMB10* and *PSME3* showed evidences of selection; nonetheless, variants in *PSME3* might have hitchhiked with a selected allele in a nearby gene, highlighting the need to analyze flanking regions to avoid incorrect inference of selection at a given gene. In fact, *PSME3* showed low diversity and SFS statistics in all populations (Supplementary Figure S5, Supplementary Table S6); in YRI one variant in the gene (rs3785545) had a significant DIND test (Supplementary Figure S6) and represented an outlier in the YRI/CEU F_{ST} distributions (Supplementary Figure S6). Yet, analysis of 5′ and 3′ flanking regions revealed that a SNP (rs61995868) in full linkage disequilibrium (LD) with rs3785545 ($r^2 = 1$ in YRI) was an F_{ST} outlier and had a DIND higher than rs3785545. This variant is a nonsynonymous substitution in the nearby *CNTD1* gene and is likely to represent the selection target (Supplementary Figure S7). Conversely, *PSMB10* was subject to directional selection; indeed the gene showed low diversity in CEU and AS (Supplementary Figure S5) and negative Fay and Wu's H in CEU (Supplementary Table S6). One synonymous variant (rs14178) was an outlier in the distribution of YRI/CEU F_{ST} values and in the distribution of DIND-DAF values (Supplementary Figure S6); analysis of 100 kb surrounding the gene revealed no SNP with higher F_{ST} and DIND ranks than rs14178. In CEU the SNP falls in a region of local reduction in Fay and Wu's H, and it is located in the fifth exon of the small *PSMB10* gene (Figure 6A). In this region DNaseI hypersensitive sites and transcription factor binding sites have been mapped by CHIP-seq in several cell lines (Figure 6A). In CEU rs14178 is in full LD ($r^2 = 1$) with rs11574514, which is located 1850 bp apart and has been associated with Crohn's disease (CD) in genome-wide association studies [47].

The activity of the proteasome is complemented by cytoplasmic peptidases [2]. One of these, *NRD1*, was found to represent a selection target in Asian populations. The gene showed low diversity (Supplementary Figure S5) and negative SFS-based statistics (Supplementary Table S6); several SNPs were outliers in the YRI/AS F_{ST} distribution and in AS three of these also showed a very high DIND test (Supplementary Figure S6). The three variants had similar DAF (0.94) in AS, and rs1538881 had the highest DIND rank; in an extended region no other variant showed outlier values for DIND and F_{ST}. A sliding-window analysis along the region indicated that rs1538881 falls in a valley of Fay and Wu's H calculated on AS chromosomes (Figure 6B). The variant is located at the beginning of the long first intron of the gene, a region where open chromatin signals and H3K4Me1 histone marks have been described in K562 and lymphoblastoid cells (Figure 6B).

Among genes involved in MHC class II presentation, *IFI30* (also known as *GILT*), *CTSE*, and *CTSL2* were found to represent selection targets. Analysis of *IFI30* indicated negative Fay and Wu's H values in AS (Supplementary Table S6) and one outlier SNP (rs7125) in the DIND-DAF distribution for the same population (Supplementary Figure S6). Analysis of the extended region revealed no SNP with higher rank in the DIND test. The variant is synonymous and falls within a nuclease accessible site in CD34- maturing myeloid cells [48] (Figure 6C).

CTSL2 encodes a cysteine protease also referred to as *CTSV*; analysis of the gene showed a significant negative Fay and Wu's H in CEU (Supplementary Table S6); F_{ST} analysis indicated rs7037968 as an outlier in the CEU/AS distribution (Supplementary Figure S6); analysis of an extended region revealed one single variant with F_{ST} (rs4361859) similar to rs7037968. Sliding window analysis of Fay and Wu's H in CEU indicated that rs7037968 (but not rs4361859) is in a local valley, suggesting that it represents the selection target (Figure 6D). No functional annotation has been described for rs7037968.

As for *CTSE*, encoding cathepsin E, the gene region showed reduced diversity in AS (Supplementary Figure S5) and low Tajima's D and Fay and Wu's D* and F* in this same population (Supplementary Table S6). F_{ST} analysis was performed for all variants in the gene and for genomic flanks, although the region immediately telomeric to *CTSE* is not covered in the human reference sequence, therefore only variants centromeric to the gene were included. Several SNPs were found to be outliers in the YRI/AS F_{ST} distribution (Supplementary Figure S6) and closer inspection revealed that in a number of cases this is due to derived alleles that are fixed or almost fixed in AS, while remain at intermediate frequency in African populations. Most variants cluster in a region upstream *CTSE* or within the transcription unit (Figure 6E), suggesting that a complete/almost complete selective sweep has occurred in AS and targeted *CTSE*; mapping of these variants indicated that many of them fall within potential regulatory regions carrying H3K4Me1 histone marks in different cell types (Figure 6E).

MARCH1 has also been involved in APP, as it regulates the surface expression of MHC class II molecules [1]. Two variants in the gene (rs2036905 and rs13125648) had an extremely high DIND test in CEU (Supplementary Figure S6) and represented outliers in the YRI/CEU F_{ST} distribution (Supplementary Figure S6). The two variants are located ~9 kb apart and have similar DAF in CEU (0.61 and 0.66, respectively); interestingly, rs2036905 falls within a sequence that is highly conserved in mammals and affects a position invariant in most species (Figure 6F). In AS, 9 variants with a similar DAF (0.12 to 0.16) had very high DIND test values and represented outliers in the YRI/AS or CEU/AS F_{ST} comparisons or in both (Supplementary Figure S6). Several of these variants are located in a ~4 kb region in intron 1, and one of them (rs12509765) is within a nuclease accessible site in maturing myeloid cells (CD34- cells) [48] (Figure 6F), suggesting a role in the regulation of *MARCH1* transcription.

Antigen presentation to T cell populations distinct from CD4 and CD8 occurs through specialized molecules encoded by genes that are not located in the MHC. *MR1* showed two variants (rs4048650 and rs6686208) with very high DIND test in CEU and a similar DAF of 0.48 (Supplementary Figure S6); both SNPs are located in the long 3′UTR. rs4048650 also represented an outlier in the YRI/CEU F_{ST} distribution; analysis of an extended region revealed no additional variants showing similarly high DIND and F_{ST} values. rs4048650 is located in the 3′UTR and affects no known microRNA binding site, but it lies in a region showing H3K4Me1 histone marks in lymphoblastoid cell lines (Figure 6G). Consistently, this SNP represents an expression QTL for *MR1* [49]. As for *CD1D*, the gene showed low SFS-based statistics in YRI (Supplementary Table S6). Several variants in the gene and in flanking regions displayed extreme DIND test values in YRI and represented outliers in the YRI/CEU or YRI/AS or in both FST distributions (Supplementary Figure S6). Specifically, one of these variants (rs73012242) is located upstream the transcription start site of *CD1D* and has a DAF of 0.95 in YRI; the remaining variants are positioned downstream the transcription end site and have a DAF ranging from 0.27 to 0.41 (Figure 6H). Sliding window analysis indicated that the 5′ portion of *CD1D* and the

Table 2. Positively selected sites in the human lineage.

Gene	Codon	Ancestral AA	Human AA	dbSNP	Frequency[a] (YRI; CEU; AS)	Pr[b]	Other methods	Domain/region
CD1D	270	Ala	Val	-		0.909	-	Alpha3
	247	Asp	Gly	-		0.893	-	Alpha3
	302	Val	Met	-		0.859	MEME e BEB	Cytoplasmic tail
	200	Glu	Lys	-		0.816	-	Alpha 2
CTSL2	207	Met	Val	-		0.942	-	peptidase
ERAP1	515	Leu	Val	-		0.912	MEME	Domain II
	528	Lys	Arg	rs30187	0.585; 0.688; 0.514	0.858	-	Domain III
LGMN	294	Ile	Val	-		0.981	MEME	activation peptide
	295	Ser	Pro	-		0.977	-	activation peptide
	288	His	Arg	-		0.955	MEME	activation peptide
PSMF1	36	Tyr	Cys	rs1803415	0.79; 0.818; 0.448	0.949	-	PI31 proteasome regulator
	18	Thr	Arg	-		0.938	-	PI31 proteasome regulator
	203	His	Pro	-		0.916	MEME	Proline-rich
	192	Ala	Val	rs79465651	0.943; 1; 1	0.914		Proline-rich
TAP1	198	Ser	Thr	-		0.867	-	Transmembrane IV
	157	Glu	Lys	-		0.822	-	Transmembrane III
THOP1	350	Arg	Cys	rs148139735	0.994; 1[c]	0.971	-	Peptidase
	333	His	Arg	-		0.964	-	Peptidase

[a]Frequencies derive fron the 1000 Genomes Phase I data.
[b]Posterior probability of $\gamma > 0$ as detected by gammaMap.
[c]These frequencies derive from the NHLBI Exome Sequencing Project (ESP) in African American and European Americans, respectively.
doi:10.1371/journal.pgen.1004189.t002

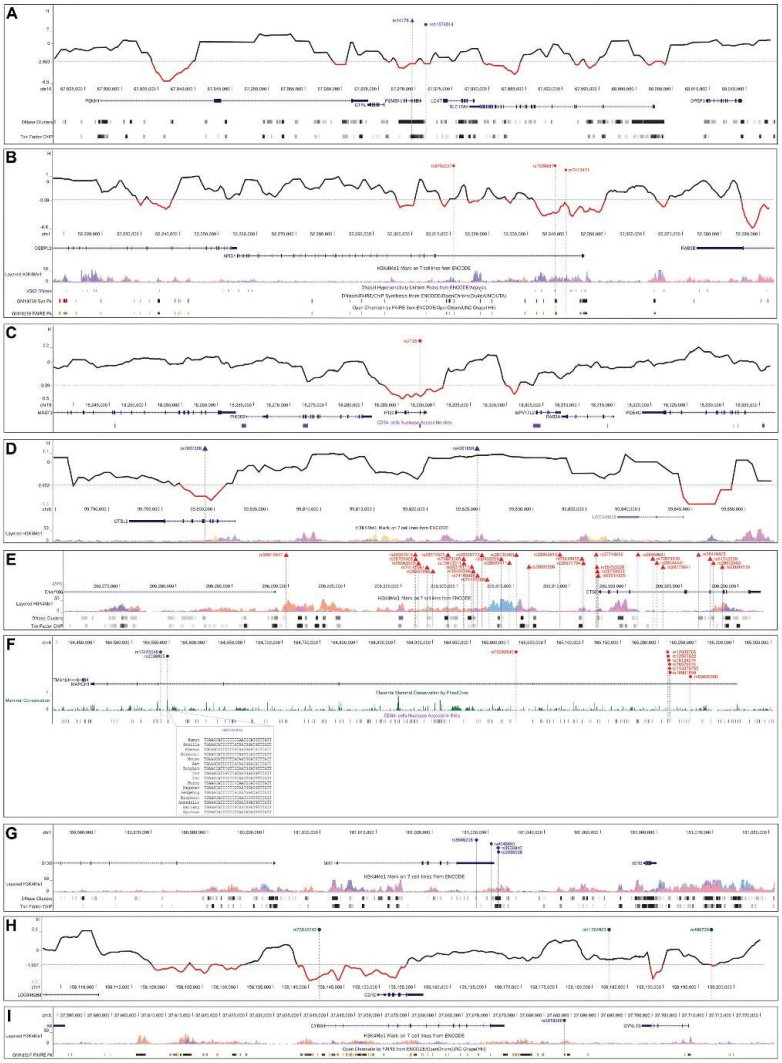

A
B
C
D
E
F
G
H
I

Figure 6. Analysis of selected variants. Location of the most likely selection targets in *PSMB10* (A), *NRD1* (B), *IFI30* (C), *CTSL2* (D), *CTSE* (E), *MARCH1* (F), *MR1* (G), *CD1D* (H), and *CYBB* (I) within the UCSC Genome Browser view. Relevant annotation tracks are shown. For *MARCH1* a short alignment of the highly conserved sequence encompassing rs2036905 is reported. For *PSMB10*, *NRD1*, *IFI30*, *CTSL2*, and *CD1D* a sliding-window analysis of Fay and Wu's H is also shown, as mentioned in the text. The hatched horizontal line represents the 5th percentile (see methods) and significantly negative values are in red. Variants in blue, red and green represent selection targets in CEU, AS, and YRI, respectively. SNP notation is as follows: triangle, F_{ST} outlier; diamond, DIND outlier; dot, both F_{ST} and DIND outlier; square, SNP associated with a disease.
doi:10.1371/journal.pgen.1004189.g006

upstream region encompassing rs73012242 correspond to a valley of Fay and Wu's H (Figure 6H), suggesting that this SNP represents the selection target at the *CD1D* locus and that the downstream polymorphisms might result from a distinct selective event possibly involving telomeric genes. The derived allele of rs73012242 is fixed in CEU and AS, suggesting that the sweep is complete in these populations. No functional annotation is reported for this variant.

Finally, *CYBB* showed low diversity (Supplementary Figure S5) and negative SFS-based statistics in AS (Supplementary Table S6); in CEU θ_W was reduced (Supplementary Figure S5). Analysis of an extended region indicated that one variant (rs5918386) had extremely high DIND test in both CEU and AS and represented an outlier in the YRI/CEU and YRI/AS F_{ST} distributions (Supplementary Figure S6). This variant is located downstream the transcription end site of *CYBB*, in a region where open chromatin and H3K4Me1 histone marks have been described in lymphoblastoid cell lines (Figure 6I). Sliding window analysis was not performed due to the low number of variants segregating in the region.

Finally, we assessed whether the selection signatures we identified above could also be detected using other tests based on extended haplotype homozygosity, namely lnRsb [45] and iHS [46], and if they overlapped with previous positive selection scans. The lnRsb test contrasts extended haplotype homozygosity between two populations and has good power for selective events at high frequency [45], whereas iHS compares the homozygosity decay for haplotypes carrying the ancestral and derived alleles for a given variant in the same population. The test has maximum power for intermediate frequency selective events [46]. As above, an empirical distribution was obtained for lnRsb (CEU/YRI, CEU/AS, and CEU/AS) and iHS values. Six of the selection targets we identified in the analyses above showed very high lnRsb values: rs1538881 in *NRD1* ($lnRsb_{AS/YRI}$: 1.63, rank: 0.951), rs7037968 in *CTSL2* ($lnRsb_{CEU/AS}$: 2.56, rank: 0.988; $lnRsb_{CEU/YRI}$: 2.38, rank: 0.990), most SNPs in *CTSE* and flanking regions (strogest SNP: rs57713692, $lnRsb_{AS/YRI}$: 3.60, rank>0.999), rs2036905 in *MARCH1* ($lnRsb_{CEU/YRI}$: 1.68, rank: 0.950), rs4048650 in *MR1* ($lnRsb_{CEU/YRI}$: 2.30, rank = 0.987), and rs5918386 downstream *CYBB* ($lnRsb_{CEU/YRI}$: 2.62, rank = 0.994) (Supplementary Figure S8). In the case of rs14178, lnRsb was high but not exceptionally so ($lnRsb_{CEU/YRI}$: 1.22, rank = 0.888). As for the iHS test, no variant showed outlier results, the best value being iHS = −1.80 (rank = 0.93) for rs7125 in AS. Nonetheless, it should be noted that most variants we identified have high DAF, thus being difficult to detect through the iHS. Also, the selective event at rs73012242 (upstream *CD1D*) is almost impossible to detect using either lnRsb or iHS as the sweep is at very hight frequency in YRI and likely complete in AS and CEU.

To evaluate the overlap between the signal we detected and those identified in previous scans of positive selection, we retrieved data from 9 genome-wide studies [45,46,50–56] that applied different approaches. This analysis indicated that large genomic regions covering portions of *MARCH1* had been previously identified in both CEU and AS by Williamson and co-workers [50], who applied a composite likelihood ratio (CLR) model (the *MARCH1* regions have CLR p values <0.01), and by Tang et al. [45], by application of the lnRsb test (Supplementary Figure S9). These latter authors also described a genomic region encompassing *NRD1* as a selection target in AS (Supplementary Figure S9). No overlaps were detected for the remaining genes.

Balancing selection targeted coding variants in APP genes

Balancing selection is more difficult to detect than positive selection, mainly because its signal (an excess of polymorphism) is often confined to narrow genomic regions [57]. Because the low-coverage 1000 Genomes Pilot Project data are skewed against singletons and low-frequency variants, and because this bias is not homogeneous along the genome, local minor differences might have a comparatively high weight when the selection signal is restricted to relatively small regions. Thus, to obtain unbiased estimates of nucleotide diversity and of the SFS, we Sanger resequenced the putative balancing selection targets in 60 HapMap subjects (20 YRI, 20 CEU and 20 AS).

In particular, resequencing was performed for the entire coding sequences of *CD207*, *PSMB9* and *TAP1*. Given the large size of the genes, two sub-regions of 4.6 and 3 kb, respectively were resequenced for *CTSB* and *NCF4* (Figure 7A); these genomic portions were selected because they contain outlier SNPs in the distribution of F_{ST} values (Supplementary Figure S6).

For each analyzed region/gene, nucleotide diversity was assessed by calculating θ_W and π; as a control for demographic effects, both indexes were calculated for 5 kb windows deriving from 238 genes resequenced by the NIEHS (National Institute of Environmental Health Sciences) SNP Program. Because under neutral evolution the amount of within-species diversity is predicted to correlate with levels of between-species divergence, we also applied a Maximum-Likelihood-ratio HKA (MLHKA) test [58] to assess whether an excess of polymorphism was observed relative to divergence.

Estimates of nucleotide diversity higher than the 95th percentile were obtained for all genes/regions in at least one population (Table 3, Supplementary Table S7). Nonetheless, a significant excess of nucleotide diversity versus inter-species divergence (as detected by the MLHKA test) was observed only for *CD207* and *TAP1* in YRI, and for *NCF4* in AS (Table 3, Supplementary Table S7). High levels of diversity in human populations that are paralleled by high inter-species diversity (i.e. non-significant MLHKA test) are difficult to interpret and raise the possibility that polymorphisms are not being maintained by selection but result from a high local mutation rate or from relaxation of functional constraints. Thus, we considered candidates of balancing selection only genes/regions that rejected neutrality based on the MLHKA results (in at least one population). For *TAP1*, *CD207*, and *NCF4* we verified whether the neutral model could be rejected by SFS-based statistics through coalescent simulations. Positive values of Tajima's D and of Fu and Li's D* and F* indicate an excess of intermediate frequency variants and are a hallmark of balancing selection, although non-significant SFS statistics may be observed when balancing selection is multiallelic or when balanced haplotypes/alleles are not at intermediate

frequency. Significantly high SFS tests were observed for at least one statistic for *TAP1* and *CD207* in YRI, as well as for *NCF4* in AS (Table 3). The values of Tajima's D and of Fu and Li's D* and F* were also compared to the distributions obtained from 5 kb windows deriving from Sanger resequenced NIEHS genes; also these statistics were calculated using the 1000 Genomes Pilot Project data (Supplementary Table S8). Overall, high concordance was observed between coalescent simulation p values and percentile ranks obtained from Sanger sequencing, whereas the 1000 Genomes Project data yielded few values higher than the 95th percentile (Supplementary Table S8), suggesting that Sanger sequencing or high-coverage data may be better suited for the detection of balancing selection.

To further extend these analyses, haplotype phylogenies were reconstructed for *NCF4*, *TAP1*, and *CD207*. The haplotype phylogeny for the resequenced NCF4 region showed 3 main haplotype groups (hapI–III, Figure 7B) with an estimated time to the most recent common ancestor (TMRCA) ranging from 840,000 to 1,790,000 years (Supplementary Table S9, Supplementary Figure S10). One of them (hap I) has low frequency in all populations and carries putative regulatory variants (Figure 7A–B). Hap II carries the derived allele of rs3788524, which is an outlier in the YRI/AS F_{ST} distribution (Supplementary Figure S6); in AS and CEU this variant is in strong LD with L272P (rs2075939), which also defines HapII. The derived allele of a putative regulatory variant (rs738148) defines hapIII (Figure 7A–B). Overall, these data support a scenario of multiallelic balancing selection at the *NCF4* gene, with both missense and regulatory variants being maintained in human populations.

In the case of *TAP1*, the haplotype network showed a complex scenario and revealed a few recurrent mutations, possibly originating from recombination or gene conversion. One major cluster of haplotypes is evident, and all these chromosomes carry the derived alleles at aminoacid residues 393 and 697 (393I and 697D). Two distantly related haplotypes are observed in YRI (YRI-hapI and YRI-hapII, Figure 7C) and both carry at least one distinctive nonsynonymous variant (V518L and G77R plus Q788K, respectively). The presence of highly differentiated haplotypes with restricted geographic distribution might be suggestive of ancient population structure [59]; nonetheless, calculation of the TMRCA for the haplotype phylogeny yielded estimates ranging from 1,670,000 to 660,000 years (Supplementary Table S9, Supplementary Figure S10),which are not consistent with population structure in Africa. Although some variants that affect putative gene transcription regulatory elements are also located on the branches of the haplotype genealogy, the balancing (or diversifying) selection targets are likely to be accounted for by aminoacid substitutions.

Finally, the haplotype network of *CD207* was reconstructed using variants located in a sub-region of relatively tight linkage disequilibrium (covering the whole transcription unit with the exclusion of exon 1 and intron 1); nonetheless, some recurrent mutations were evident (Figure 7D). The two major haplogroups carry different alleles at two polymorphisms that affect residues in the CRD: N288D, which was shown to affect binding to mannose [60], and K313I, where the lysine residue forms the sulfated glycan recognition interface [61]. Within the more common haplotype cluster, other missense variants are observed, including A278V, which does not influence sugar binding or protein stability [60]. Few CEU chromosomes are differentiated at the S213P variant (Figure 7D); reconstruction of ancestral state at this site is difficult as different primates carry distinct residues, in line with the fact that this position was found to be positively selected in mammals (Figure 3). Overall, these data suggest that in humans balancing selection targeted two nonsynonymous variants -K313I and N288D- resulting in two major langerin forms (288N-313K and 288D-313I), that segregate in human populations and are likely different in their sugar binding specificity.

CD207/langerin can internalize HIV-1 to Birbeck granules where it is degraded [62]. Thus, we explored the possibility that the selected functional variants in the CRD domain affect the susceptibility to sexually-transmitted HIV-1 infection. To this aim, we genotyped rs13383830 (N288D) in a cohort of 87 Italian heterosexual HIV-exposed seronegative (HESN) individuals who have a history of unprotected sex with their seropositive partners [63] and in 436 randomly selected Italian subjects (controls). The variant significantly deviated from Hardy-Weinberg equilibrium (HWE) with an excess of homozygotes in HESN alone (Table 4). This observation may be explained by the underlying genetic model (i.e. protection from HIV-1) or by spurious effects; application of a goodness-of-fit test [64] indicated that a recessive model with only genetic effects adequately explains HWE deviation in HESN. Comparison of allele frequencies in the HESN and control samples indicated no significant difference for rs13383830. Conversely, the genotype distribution of the SNP was significantly different in the two cohorts, with 288D/288D homozygotes being much more common in HESN than in controls (permutation $p = 0.015$ and 0.023 for a genotypic and a recessive model, respectively, Table 4). Thus, homozygosity for the 288D allele may be a factor in determining protection from sexually-transmitted HIV-1 infection.

Discussion

Adaptive evolution acts at the level of genetic variants that determine advantageous phenotypic traits. Selection signatures can therefore be exploited to detect genomic regions/positions underlying phenotypic diversity and adaptation. This has recently been demonstrated within a host-pathogen arms race scenario whereby an evolutionary-guided approach was used to identify a protein loop in MX1 (myxovirus resistance 1) that determines antiviral activity [65]. Similarly, it has been known for years that natural selection has specifically acted on the peptide-binding cleft of antigen presenting molecules [4]. Because the repertoire of peptides that is available for presentation is generated by APP gene products, we performed an evolutionary analysis of these loci.

Evolutionary analysis at the inter-specific level indicated that 11 genes have been targeted by diversifying positive selection; this represents a substantial fraction (24%) of analyzed genes, despite our application of a conservative approach. Moreover, an analysis of positive selection in the human species identified positively selected codons at four additional genes. Although large-scale analyses had previously identified immune response loci as preferential targets of positive selection in mammals [66,67], those studies had limited power due to the inclusion of a small number of species. Thus, in Kosiol et al. [66] the percentage of positively selected genes among those involved in APP only amounted to 14%.

Likewise, we identified several genes targeted by natural selection during the more recent history of human populations. Integration of different tests for selection was recently shown to be a powerful tool to identify and finely map positive selection targets [68]; the approach we applied herein differs in a number of ways from that proposed by Grossman and co-workers [51,68]. We did not apply the integrated haplotype score (iHS) or its derivatives, but rather relied on the DIND test, which was proven to be more powerful than iHS in most ranges of selected allele frequency [42]. We used the normalized H statistic (as it has higher power than the

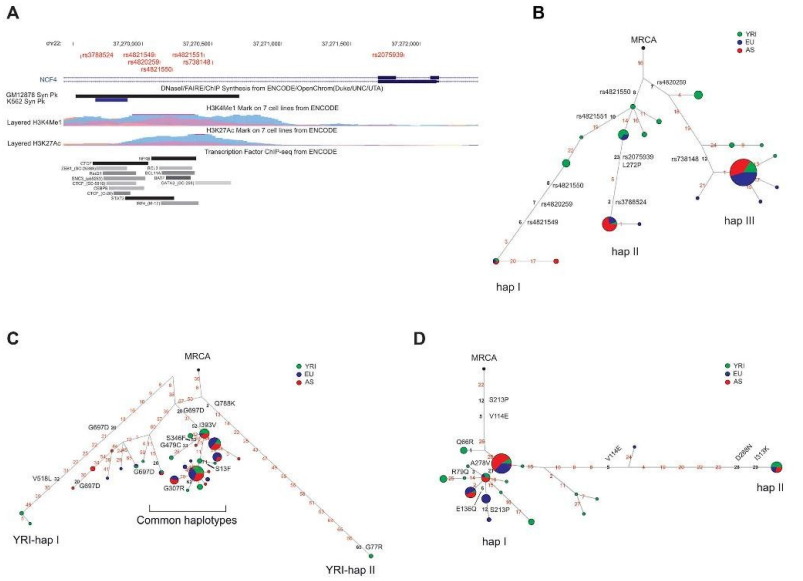


Figure 7. Analysis of *NCF4*, *CD207*, and *TAP1*. (A) Exon-intron structure of the resequenced *NCF4* region with inclusion of a few annotation tracks. The position of SNPs shown in the haplotype network (B) is reported. In the network each node represents a different haplotype, with the size of the circle proportional to frequency. Nucleotide differences among haplotypes are indicated on the branches. The most recent common ancestor (MRCA) is also shown (black circle). The relative position of mutations along a branch is arbitrary. Haplotype phylogenies for *TAP1* (C) and *CD207* (D) were reconstructed through median-joining networks. Nonsynonymous variants are shown, as well as SNPs that fall within potential regulatory elements.
doi:10.1371/journal.pgen.1004189.g007

original non-normalized Fay and Wu's H [39]) rather than the ΔDAF test [51,68], and we included SFS-based statistics. Thus, due to the different power of distinct tests, none of the variants described herein was identified in previous scans for positive selection. At the gene level, analysis of genome-wide scans of positive selection indicated that regions encompassing *MARCH1* and *NRD1* had previously been described as positive selection targets [45,50], whereas no overlap was detected for the remaining genes. Low concordance of positive selection signals among studies has been previously noted (for a recent review [69]). However, most previous positive selection scans have been performed using SNP genotype data (however dense in some cases) whereas we used resequencing data (although low-coverage), which are expected to increase the power to detect selection [70]. Indeed, even tests based on extended haplotype homozygosity, that are less sensitive to the ascertainment bias, have increased power when the actual selection target is included in the analysis [46]. One extreme example of this is accounted for by *CTSE*, where no SNP mapped in HapMap releases predating 2008 and which is still poorly covered by HapMap data.

Several reports have indicated that genes involved in immune response may be preferential targets of both positive [45,46,51,56] and balancing [71] selection in humans, with some immune-response pathways possibly being particularly enriched in selection signals. Tang et al. [45] found an over-representation of genes coding for cytokines (IL-1 receptor agonists in particular) among their top signals; likewise, other authors indicated an enrichment for complement-mediated and class I MHC-related immune response genes [46,51]. Beside genome-wide scans, studies that focused on specific families of immune response loci often revealed a high proportion of selected genes; these include, for example, type III interferon genes [72], genes coding for T-cell regulatory molecules [73], and NOD-like receptors [74]. These observations clearly reflect the extremely important role of immune response for survival in the face of infections. Nonetheless, analyses herein also indicated that for some components of the APP pathway (e.g. immunoproteasome subunits, chaperones, several lysosomal proteases) negative selection likely represented the major evolutionary force. Conversely, genes that code for APP components that, at different levels, directly interact with the antigens to be presented (e.g. *CD1D*, *CD207*, *TAP1*, *ERAP2*, and *CYBB*) have been constantly targeted by positive or balancing selection, as determined by both inter- and intra-species analyses.

Besides providing a general picture of the evolutionary forces acting on the APP pathway, our aim was to describe in detail the

Table 3. Nucleotide diversity and neutrality tests.

Gene	L[a]	Pop[b]	S[c]	θ_w[d]		π[f]		Tajima's D		Fu and Li's D*		Fu and Li's F*		MLHKA	
				value	rank[e]	value	rank[e]	value	p[g]	value	p[g]	value	p[g]	K[h]	p'
CD207	4.7	YRI	40	19.90	**0.98**	20.13	**0.98**	0.04	0.10	0.51	**0.045**	0.41	**0.042**	2.53	**0.014**
		CEU	27	13.43	**0.96**	12.06	0.87	−0.35	0.25	0.68	0.13	0.39	0.26	2.22	0.066
		AS	20	9.95	0.90	9.13	0.82	−0.27	0.25	1.32	**0.013**	0.93	0.099	1.77	0.39
NCF4	3.0	YRI	20	15.60	0.94	15.33	0.94	−0.056	0.24	−0.75	0.38	−0.61	0.49	1.97	0.13
		CEU	17	13.26	**0.95**	11.31	0.88	−0.48	0.31	−1.14	0.18	−1.08	0.19	2.31	0.071
		AS	16	12.48	**0.96**	17.88	**0.97**	1.39	0.098	1.58	**0.007**	1.79	**0.016**	2.27	**0.039**
TAP1	7.2	YRI	53	17.17	**0.97**	10.98	0.79	−1.29	0.098	1.32	**0.0017**	0.49	0.057	2.92	**0.0065**
		CEU	24	7.77	0.72	6.14	0.60	−0.71	0.19	−0.66	0.29	−0.80	0.23	1.66	0.73
		AS	23	7.45	0.77	7.41	0.40	−0.018	0.41	−0.45	0.32	−0.36	0.33	1.75	0.77

[a]length of analyzed resequenced region (in kb);
[b]population;
[c]number of segregating sites;
[d]θ_w estimation per site ($\times 10^{-4}$);
[e]percentile rank relative to a distribution of 238 5 kb windows from NIEHS genes;
[f]π estimation per site ($\times 10^{-4}$);
[g]p value calculated by coalescent simulations;
[h]selection parameter (k>1 indicates an excess of polymorphism compared to divergence; k<1 indicates the opposite situation);
[i]p values obtained by applying a calibrated population genetics model, as described in the text.
doi:10.1371/journal.pgen.1004189.t003

Table 4. Genotype counts, HWE proportions and association analysis for rs13383830.

SNP ID	Phenotype	Genotype counts	Genotype counts (recessive model)	p^a (HWE)	p^b (genotype)	p^b (recessive)
rs13383830 (N288D)	**HESN**	3/12/72	3/84	0.04	0.015	0.023
	CTR	2/59/375	2/434	>0.99		

[a]HWE deviation p value.
[b]*p* value obtained from 10,000 phenotype-label swapping permutations.
doi:10.1371/journal.pgen.1004189.t004

specific sites and variants targeted by natural selection so that this information can be exploited to prioritize functional characterization in follow-up analyses. We defined positively selected sites in mammals by the combined use of two methods, BEB and MEME; this choice was taken to limit the number of false positive results, although we most likely underestimated the number of selected sites. In fact, MEME was developed to detect both episodic and pervasive positive selection [9], whereas sites evolving under episodic selection are likely to be missed by BEB. Thus, the combination of the two methods is expected to result in the confident identification of sites evolving under pervasive diversifying selection only.

Nonetheless, several sites evolving adaptively were identified and they are expected to define positions and protein regions that affect functional properties. As an example, our data indicate that a threonine residue (322T) that functions as a trafficking signal in the cytoplasmic region of CD1D [13] is present in primates only and represents a selected site, suggesting that different motifs evolved in distinct mammalian species to modulate CD1D expression at the plasma membrane. Indeed, differences in intracellular trafficking between mouse and human CD1D molecules have been reported [75]. Interestingly, it has been proposed that the 322T signal is exploited by HSV-1 to downmodulate the surface expression of CD1D molecules as an evasion strategy [13]. Thus, the cytoplasmic tail and the transmembrane region of CD1D might have evolved under virus-driven selective pressure. Indeed, different pathogens, including HSV-1, HPV, HIV-1, VSV, and KSHV, interfere with CD1D expression and recycling [75], although the specific contact interfaces between viral products and CD1D molecules are unknown. Adaptive evolution was also evident in the extracellular domains of CD1D; sites positively selected in mammals are spatially clustered and flank the TcR interaction surface and the lipid binding pocket, suggesting that they may exert indirect effects on binding specificity, especially in light of the broad array of lipid molecules presented by CD1D [75]. Similarly, a human-specific positively selected site at the $\alpha 2/\alpha 3$ domain interface might modulate CD1D activity by altering the flexibility or relative positioning of the extracellular domains.

Different viral species are known to encode products that counteract specific components of the APP pathway other than CD1D. This represents a strategy to evade the host immune system by hampering the presentation of immunogenic epitopes. Specifically, several viral proteins target the PLC by binding TAP or tapasin [5]. Viral inhibition of the PLC is suggested to be of pivotal importance for efficient infection; for example different herpesviruses encode distinct TAP inhibitors, which are unrelated in genome location, structure, and mechanism of action, suggesting convergent evolution [76]. This indicates that some of the positively selected sites we identified in TAP1 and tapasin (TAPBP) might have evolved to avoid targeting by viral products. One of these is the US3 immunomodulator encoded by HCMV; this protein directly binds the tapasin ER luminal domain, but has no effect on the formation of the TAP-tapasin complex [5]. US3 might interfere with recruitment of ERp57 by tapasin [5], suggesting that the tapasin 225N residue -located at the ERp57 binding interface - might be involved in this process.

Three of the positively selected sites in TAP1 are located in the channel forming region and one of them (516Q) maps to a transmembrane domain that directly interacts with peptides. Because TAP is known to select peptides for transportation in a species-specific manner [77], it would be interesting to evaluate the effect of the identified residues on TAP binding affinity and transportation preference, as well as on the sensitivity to viral inhibitors. TAP contributes to the shaping of the overall repertoire available for MHC presentation. On the one hand this property *per se* represents a possible target for host-pathogen arms races, as decreasing transport of specific peptides would translate in reduced presentation. On the other hand, it has been noticed that in human, mouse, and rat, the specificity of TAP transportation correlates with the predominant peptide binding profiles of the corresponding MHC class I molecules, suggesting co-evolution [77].

Co-evolution with MHC class I molecules might also be driving aminoacid replacements at BLMH and tapasin. Indeed, the N-terminal domain of tapasin, where one of the selected residues (67S) is located, was shown to facilitate MHC-peptide complex folding depending on the identity of both the peptide and of the HLA I heavy chain. As for BLMH, experiments in human cells indicated that its depletion affects peptide loading and MHC class I surface expression in a *HLA* class I allele-dependent manner [78]. *BLMH* is highly conserved from yeast to mammals, suggesting strong constraints [27]. As a consequence, selection might have acted at the level of aminoacid residues that modulate protein abundance at the post-translational level, as suggested by their location. Likewise, natural selection might have acted at the *ERAP2* locus to modulate protein stability and, consequently, abundance. Although the observation that protein-destabilizing variants have been favored during evolution might seem counter-intuitive, it should be noted that an *ERAP2* haplotype that results in a truncated (and degraded) protein product is maintained by balancing selection in human populations [33,44]. Also, some rodent species, including mice and rats, lack a functional *ERAP2* gene, suggesting that loss or decreased abundance of ERAP2 protein might confer some advantage, possibly related to selective antigen trimming. We also detected human-specific selective events at *ERAP1*. One of the two variants we identified had previously been shown to represent a balancing selection target in human populations [33]. The variant affects enzymatic properties [79] and associates with the susceptibility to different autoimmune diseases, often in interaction with *HLA* allelic status [80].

Analysis of *CYBB* and *CD207* also provides remarkable examples of the action of different selective forces on the very same gene region, as both highly variable and strongly constrained

positions are observed in close proximity at these loci. Indeed, most missense substitutions that cause mendelian immunologic defects involve aminoacid positions that are conserved in all mammals, indicating that negative selection at these sites prevents aminoacid replacements affecting host resistance to pathogens. The pattern of positive selection at *CTBB* indicates that the two long loops protruding in the extracellular space or in the phagosome lumen are strongly targeted by diversifying selection. These protein regions are expected to be mostly exposed to a direct interaction with pathogen components, suggesting that they have evolved to avoid inhibition by bacterial/fungal products, a possibility that awaits experimental validation. In addition to its role in cross-presentation, the NADPH oxidase complex directly participates in the killing of pathogenic microbes through the production of superoxide and other oxidants in neutrophils. This activity is also required to activate cathepsin G and other proteases that, in turn, kill and digest engulfed pathogens [81]. Most positively selected sites we identified in *CTSG* are likely to modulate substrate specificity as they rim the binding pockets. Likewise, the site targeted by positive selection in simians is located at the edge of the substrate binding pocket on an exposed loop that also carries 177G (positively selected in the whole phylogeny); this loop has previously been shown to confer substrate specificity to other serine proteases [25,82]. Interestingly, a site subject to diversifying selection (106Q) is adjacent to a position (104T) that, if replaced with the equivalent aminoacid in elastase (T104N), confers to CTSG the ability to cleave *Shigella* virulence factors [82] (Figure 3B). Thus, the selective pressure acting on both *CTBB* and *CTSG* might be related to their direct antimicrobial role in addition to participation in APP. Finally, analysis of human-specific positively selected sites in *LGMN*, which also encodes a lysosomal protease, indicated that one of them maps to the α cleavage site of the activation peptide. Although the identity of the protease(s) responsible for cleavage is presently unknown, the multistep activation of LGMN is thought to have a regulatory significance and is modulated by the maturation status of dendritic cells, possibly via acidification of the endosome/lysosome compartments [83].

Results herein also indicate a continuum in selective pressure acting on different timescales and targeting the coding sequences of *TAP1* and *CD207*, as aminoacid-replacement variants are likely to represent the selection targets in human populations. In both cases balancing selection signatures were detected in African populations only. Because we accounted for demography events both in coalescent simulations and by the empirical comparison with genes resequenced in the same populations, the signatures we detect are unlikely to represent demographic effects, but instead indicate stronger selective pressure in Africa. Interestingly, one of the putative balancing selection targets in *TAP1*, the V518L variant, is located in the peptide binding domain, close to a positively selected site (516Q), and defines a minor haplotype in YRI; this variant might affect the affinity of TAP1 for one or more antigenic peptides. Likewise, in the case of *CD207* one positively selected site (289A) immediately flanks a human polymorphic position representing a balancing selection target (N288D) with known effect on sugar binding [60]. The second site subject to diversifying selection (213P) is polymorphic in humans (P213S), although its positioning on the haplotype network does not suggest that it is a major target of balancing selection in humans. Indeed, the two major haplotype clades of *CD207* carry, in addition to N288D, a second variant, K313I, that also affects langerin binding to glycan substrates [61]. This indicates that balancing selection has maintained two alternative langerin forms that differ in binding specificity and may recognize distinct microbial glycan structures, ultimately affecting the susceptibility to specific infections. We show that homozygosity for the 288D-313I langerin haplotype may be associated with protection against sexually transmitted HIV-1 infection. The HIV-1 gp120 protein, which is bound by CD207, is heavily glycosylated with both oligomannose and complex N-glycans [84]; the 288D allele displays reduced binding to mannose-containing structures [60], but may confer increased affinity for more complex sugars, as suggested by the broad specificity of langerin. Overall, although the recessive effect of the rare haplotype is consistent with the trimeric nature of langerin, and its frequency differs in HESN and controls (3.45% and 0.46%, respectively), the association results should be regarded as preliminary and treated with caution due to the small sample size and the low frequency of the putative protective haplotype. Thus, replication in an independent cohort and functional analyses on the role of the 288D allele in HIV-1 recognition and internalization will be needed.

One nonsynonymous polymorphism (L272P) in *NCF4*, encoding a cytosolic regulatory component of the NADPH oxidase complex, was also identified as a possible balancing selection target in human populations. This SNP is located in an intron of the gene that may be retained in the transcript as a result of alternative splicing. Nonetheless, the selection target might also be accounted for by variants with a regulatory function on *NCF4* expression. Indeed SNPs located on the branches of the haplotype genealogy fall within Chip-seq mapped binding sites for transcription factors including STAT3, which is regulated by RAC1 [85], a modulators of NAPH oxidase activity [86], and NFKB, a central transcriptional regulator in myeloid cells. Similarly, we found all adaptive variants subject to directional selection to represent likely modulators of gene expression levels.

As recently suggested [68], the use of large-scale low coverage data, while posing challenges due to the biased SFS, may allow identification of the causal variant underlying the selective event. This represents a valuable advantage by providing a list of targets that may be directly tested in functional analyses. Moreover, integration of selection signals with extensive functional annotations generated by the ENCODE project and by eQTL studies further increases the possibility to underscore adaptive alleles. Our analysis indicated that two of the selected variants (in *IFI30* and *MARCH1*) are located within nuclease accessible sites in maturing myeloid cells, suggesting they affect transcription regulatory elements activated during cell differentiation [48] and the selected variant in *MR1* represents an eQTL. Likewise, selected variants in or close to *NRD1* and *CTBB* fall within open chromatin regions in lymphoblastoid cell lines, and the synonymous variant in *PSMB10* maps to DNAse I sensitive sites in different cell types and to transcription factor binding sites. Interestingly, this variant is in full LD in CEU with a risk SNP for Crohn's disease [47], again supporting the view that adaptive events underlie phenotypic variability. In general, most of the positive selection events we described occur at positions with a likely role controlling gene expression. Grossman and co-workers [68] finely mapped causal variants in 412 candidate selected regions and determined the large majority of these may modulate transcription levels. Likewise, Vernot et al. [87] performed a genome-wide analysis of DNase I hypersensitive regions and indicated that these harbor a number of variants targeted by positive selection in human populations. Thus, our data are in agreement with previous findings and help substantiate the view that regulatory variation represents a major target for adaptive evolution in humans.

Materials and Methods

Gene selection

The initial list of genes to be included in the study was obtained from Gene Ontology (GO). Specifically, we queried GO for all the all human genes (n = 180) associated with the following GO terms (and children): GO:0019884 (antigen processing and presentation of exogenous antigen), GO:0019883 (antigen processing and presentation of endogenous antigen), GO:0002474 (antigen processing and presentation of peptide antigen via MHC class I), GO:0002495 (antigen processing and presentation of peptide antigen via MHC class II), GO:0002428 (antigen processing and presentation of peptide antigen via MHC class Ib). From this initial list we removed *HLA* class I (n = 7) and class II genes (n = 15), as they have been the topic of intense investigation, as well as immunoglobulin receptors (n = 3) and integrins (n = 2), as they are not directly involved in the process that leads to antigen processing and presentation (APP). We also pruned genes that, although participating in APP, play non-specific roles including components of the constitutive proteasome (n = 34), general ubiquitination factors (n = 4), molecules involved in the formation and transport of clathrin-coated vescicles (n = 19), proteins involved in vesicle trafficking across different cellular compartments (n = 14), dynamins and dyneins (n = 11), dynactins (n = 6), and kinesins (n = 19). Two ubiquitine-ribosomal protein gene fusions were discarded as well, as their function is poorly understood. Finally, *HFE*, encoding a nonclassical MHC class Ib molecule, was discarded because this gene is believed to have no antigen-presentation function [88]. Thus, we concentrated our efforts on a list of 43 genes, which are considered to be central components of the APP pathway. Notably, *THOP1* and *NRD1* were also included in the final group of genes given their recently established role in antigen processing [2]; this lead to a final list of 45 genes (Supplementary Table S1).

Evolutionary analysis in mammals

Mammalian sequences for APP genes were retrieved from the Ensembl database. Mammalian orthologs of human APP genes were included only if they represented 1-to-1 orthologs as reported in the EnsemblCompara GeneTrees [89]. As mentioned in the text only primate sequences were included for *CTSL1* and *CTSL2* (Supplementary Table S2).

DNA alignments were performed using the RevTrans 2.0 utility [90], which uses the protein sequence alignment as a scaffold for constructing the corresponding DNA multiple alignment. This latter was checked and edited by hand to remove alignment uncertainties. Trees were generated by maximum-likelihood using the program DnaML (PHYLIP Package). To detect selection, NSsite models that allow (M2a, M8,) or disallow (M1a, M7) sites to evolve with dN/dS>1 were fitted to the data two models of equilibrium codon frequencies: the F3x4 model (codon frequencies estimated from the nucleotide frequencies in the data at each codon site) and the F61 model (frequencies of each of the 61 non-stop codons estimated from the data). Results for the two codon frequency models are reported in Supplementary Tables S3 and S4. Whenever maximum-likelihood trees showed differences (always minor) from the accepted mammalian phylogeny, analyses were repeated using the accepted tree, and the same results were obtained in all cases. Sites under selection with the M8 model were identified using Bayes Empirical Bayes (BEB) analysis with a significance cutoff of 0.90 [8,91].

In order to identify specific branches with a proportion of sites evolving with $\omega>1$, we used BS-REL [10]. Branches identified using this approach were cross-validated with the branch-site likelihood ratio tests from PAML (the so-called modified model A and model MA1, "test 2") [11]. A false discovery rate correction was applied to account for multiple hypothesis testing (i.e. we corrected for the number of tested lineages), as suggested [12]. BEB analysis from MA (with a cut-off of 0.90) was used to identify sites that evolve under positive selection on specific lineages. Ancestral site reconstruction for positions 416 and 420 in ERAP2 was obtained through the DataMonkey sever by ASR utility, which implements three different methods. GARD [92], MEME [9], SLAC [93], and BS-REL [10] analyses were performed through the DataMonkey server [94] (http://www.datamonkey.org).

In silico analysis of protein stability

Intra-protein interaction calculations were performed using PIC (Protein Interactions Calculator) [95]. Stability analysis was carried out using three different methods. FoldX 3.0 [96] and PoPMuSiC (web-server version) [97], were used on the chain A of the X-ray structure of ERAP2 (PDB code: 3SE6). I-Mutant 2.0 [98] was used on the corresponding protein sequence retrieved from UniprotKB (Q6P179). In FoldX and I-Mutant the ΔΔG values are calculated as follows: $\Delta\Delta G = \Delta G_{mutant} - \Delta G_{wild\text{-}type}$. In FoldX and I-Mutant ΔΔG values >0 kcal/mol indicate mutations that decrease protein stability, whereas in PoPMuSiC ΔΔG values >0 kcal/mol are mark of mutation increasing protein stability. Therefore, PoPMuSiC ΔΔG values were multiplied by −1 to obtain homogeneous results.

In the analysis carried out with FoldX 3D, the three-dimensional structure of the protein was repaired using the < RepairPDB> command. Mutations were introduced using the < BuildModel> command with <numberOfRuns> set to 5 and < VdWdesign> set to 0. Temperature (298K), ionic strength (0.05 M) and pH (7) were set to default values and the force-field predicted the water molecules on the protein surface. Residues His370, His374, Glu393 and Tyr455, which coordinates the zinc ion, were kept fixed during reparation and mutation procedures.

HapMap DNA samples and sequencing

Human genomic DNA from HapMap subjects (20 Yoruba, YRI, 20 European, CEU, and 20 Asians, AS) was obtained from the Coriell Institute for Medical Research. All analysed regions were PCR amplified and directly sequenced. PCR products were treated with ExoSAP-IT (USB Corporation Cleveland Ohio, USA), directly sequenced on both strands with a Big Dye Terminator sequencing Kit (v3.1 Applied Biosystems) and run on an Applied Biosystems ABI 3130 XL Genetic Analyzer (Applied Biosystems). Sequences were assembled using AutoAssembler version 1.4.0 (Applied Biosystems), and inspected manually by two distinct operators. All primers sequences are available in Supplementary Table S10.

Population genetics-phylogenetics analysis

Data from the Pilot 1 phase of the 1000 Genomes Project were retrieved from the dedicated website [31]. SNP genotypes were organized in a MySQL database. Coding sequence information was obtained for the 45 APP genes. Accessibility of gene region by paired-end next-generation sequencing was evaluated using the "1000 Genomes Project Phase 1 Paired-end Accessible Regions - Pilot Criteria" UCSC track.

To analyze the DFE for APP genes we used gammaMap [30]. We assumed θ (neutral mutation rate per site), k (transitions/transversions ratio), and T (branch length) to vary among genes following log-normal distributions. For each gene we set the neutral frequencies of non-STOP codons (1/61) and the

probability that adjacent codons share the same selection coefficient ($p = 0.02$). For selection coefficients we considered a uniform Dirichlet distribution with the same prior weight (0.1) for each selection class. For each gene we run 100,000 iterations with thinning interval of 10 iterations.

Population genetic analyses

A set of programs was developed to retrieve genotypes from the 100 Genomes Pilot Project MySQL database and to analyse them according to selected regions/populations. These programs were developed in C++ using the GeCo++ [99] and the libsequence [100] libraries. Genotype information was obtained for the 45 APP genes. In order to obtain a control set of ~1,000 genes to use as a reference set, we initially selected 1,200 genes by random sampling of those included in the RefSeq list. For these genes we retrieved orthologous regions in the chimpanzee, orangutan or macaque genomes (outgroups) using the LiftOver tool; genes showing less than 80% human-outgroup aligning bases were discarded. This originated a final set of 987 genes, hereafter referred to as control set. These data were used to calculate θ_W [36], π [37], as well as Tajima's D [38], Fu and Li's D* and F* [40], and normalized Fay and Wu's H [39,101] over each entire gene region.

Data from the control gene set were used to calculate empirical distributions of these parameters, as specified in the text.

Normalized Fay and Wu's H was also calculated in 5 kb sliding windows moving with a step of 500 bp. Sliding window analyses have an inherent multiple testing problem that is difficult to correct because of the non-independence of windows. In order to partially account for this limitation, we applied the same procedure to the control gene set, and the distribution of normalized Fay and Wu's H was obtained for the corresponding windows. This allowed calculation of the 5th percentile and visualization of regions below this threshold.

F_{ST} [41] and the DIND test [42] were calculated for all SNPs mapping to the control and APP gene sets. Because F_{ST} values are not independent from allele frequencies, we binned variants based on their MAF (50 classes) and calculated the 95th and 99th percentiles for each MAF class. As for the DIND test, it was originally developed for application to Sanger or high coverage sequencing data [42], so that statistical significance can be inferred through coalescent simulations. This is not the case for the 1000 Genomes Project data; thus, we calculated statistical significance by obtaining an empirical distribution of DIND-DAF value pairs for variants located within control genes. Specifically, DIND values were calculated for all SNPs using a constant number of 40 flanking variants (20 up- and down-stream). The distributions of DIND-DAF pairs for YRI, CEU and AS was binned in DAF intervals (100 classes) and for each class the 95th and 99th percentiles were calculated. As suggested previously [42], for values of $i\pi_D = 0$ we set the DIND value to the maximum obtained over the whole dataset plus 20. Due to the nature of low-coverage data, for low DAF values most $i\pi_D$ resulted equal to 0 (i.e. the 95th percentile could not be calculated); thus, we did not calculated DIND in these ranges and we consequently cannot detect selection acting on low frequency derived alleles.

The lnRsb and iHS tests were calculated as previously described [45,46] using the rehh R package [102]. Specifically, lnRsb and iHS were calculated for all tested SNPs using information from 200 kb flanking regions (100 kb 5′ and 3′). To obtain empirical distributions, we randomly selected 100 genic SNPs and calculated lnRsb and iHS values for all SNPs in their 200 kb flanks. Data obtained from these randomly selected variants were alos used to calculate the median and standard deviation for lnRsb' normalization [45].

As mentioned in the text, an approach based on coalescent simulations was applied with Sanger sequencing data. In particular, calibrated coalescent simulations were performed using the cosi package [103] and its best-fit parameters for YRI, CEU, and AS populations with 10,000 iterations. Demographic parameters for YRI, CEU and AS implemented in cosi are described in [103]. Simulations were conditioned on mutation and recombination rates. Estimates of the population recombination rate parameter ρ were obtained from resequencing data with the use of the Web application MAXDIP [104] and converted to cM/Mb.

For Sanger-resequenced regions the percentile ranks of θ_W and π were obtained from the distribution of the same parameters calculated for 5 Kb windows deriving from 238 human genes resequenced by NIEHS (National Institute of Environmental Health Sciences) SNPs Program, as previously described [105]. The maximum-likelihood-ratio HKA test was performed using the MLHKA software [58], as previously proposed [105].

Haplotype analysis and TMRCA calculation

Haplotypes were inferred from Sanger resequencing data using PHASE version 2.1 [106,107]. Median-joining networks to infer haplotype genealogy were constructed using NETWORK 4.5 [108]. Estimates of the time to the most common ancestor (TMRCA) was obtained using different methods: i) a phylogeny based approach implemented in NETWORK 4.5 using a mutation rate based on the number of fixed differences between chimpanzee and humans [108]; ii) GENETREE, which is based on a maximum-likelihood coalescent method [109,110] assuming an infinite-site model without recombination; haplotypes and sites that violate these assumptions were removed; iii) a previously described method [111] that calculates the average pairwise difference between all chromosomes and the MRCA: this value was converted into years on the basis of mutation rate retrieved as above. The SD for this estimate was calculated as previously described [112].

We based calculations on the assumption that the divergence between human and chimpanzee occurred 6 MY ago [113] and that the generation time is 25 years.

Human subjects, genotyping and association analysis

Inclusion criteria for HESN were a history of multiple unprotected sexual episodes for more than 4 years at the time of the enrolment, with at least 3 episodes of at-risk intercourse within 4 months prior to study entry and an average of 30 (range, 18 to > 100) reported unprotected sexual contacts per year. These HESN subjects are part of a well characterized cohort of serodiscordant heterosexual couples that has been followed since 1997 (reviewed in [63]).

No HESN was homozygous for the *CCR5Δ32* variant, which confers resistance to R5 HIV-1 strains [114]. As for controls, 436 Italian donors were also included in the study, irrespective of their HIV infection status. The study was reviewed and approved by the institutional review board of the S. M. Annunziata Hospital, Florence. Written informed consent was obtained from all subjects.

HWE deviation was analysed as suggested by Wittke-Thompson and co-workers [64]. The equations are parametrized in q (susceptibility allele frequency), α (risk in non-susceptible homozygotes), β (heterozygote relative risk), γ (homozygote relative risk) and K_p (trait prevalence in the general population). We obtained ML estimates for these parameters minimizing the goodness-of-fit test statistic (as reported in [64]) using the BFGS method. Using an

estimate of K_p the procedure was repeated with a general model estimating q, β and γ, and for constrained specific models, estimating q and gamma (dominant: β = γ; recessive: β = 1, γ > 1; additive: β = (γ+1)/2, γ > 1; multiplicative: β = sqrt(γ), γ > 1). Given the different number of parameters in the general model, the Akaike Information Criteria (AIC) was used for the best fit model selection. A p value was then calculated for the minimal value of the test statistic using a χ_2 distribution with 1 or 2 df for the general and constrained models respectively. Using a K_p (prevalence of HESN phenotype in the general population) of 0.20 [115,116], the best model fitting the genotypic proportions in HESN and controls was a recessive model with q (susceptibility allele frequency) = 0.079, α (risk in non-susceptible homozygotes) = 0.20, β (heterozygote relative risk) = 1, and γ (homozygote relative risk) = 3.23. For this model, the goodness-of-fit test was not significant ($\chi_2 = 1.81$, p = 0.40, df = 2), indicating that a recessive model with only genetic effects adequately explains HWE deviation. We performed the same analysis using a range of K_p (from 0.10 to 0.30) and similar results were obtained (not shown).

Association p values for the genotypic and recessive models were calculated using PLINK [117] by performing 10,000 phenotype-label swapping permutations.

Supporting Information

Figure S1 Work-flow and main results for the inter-species analysis. Genes that were defined as targets of positive selection are shown in red.
(PDF)

Figure S2 Branch-site analysis of positive selection for *CYBB*. Branch lengths are scaled to the expected number of substitutions per nucleotide, and branch colors indicate the strength of selection (dN/dS or ω). Red, positive selection (ω > 5); blue, purifying selection (ω = 0); grey, neutral evolution (ω = 1). The proportion of each color represents the fraction of the sequence undergoing the corresponding class of selection. Thick branches indicate statistical support for evolution under episodic diversifying selection as determined by BS-REL. Grey dots denote branches that were tested but not confirmed to be under positive selection using the PAML branch-site models.
(PDF)

Figure S3 Alignment of a TAPBP region and positively selected sites in CTSL2, LNPEP, ERAP1, THOP1, and PSMF1. (A) Multiple alignment of a TAPBP region for a few representative mammalian species. A positively selected site (67S) is colored in red, the cystein residue involved in disulfide-bonding is colored in blue. (B) Ribbon diagram of human CTSL2; sites that define substrate binding are shown in yellow; positively selected sites are in red (whole phylogeny) or green (humans). (C) Schematic representation of LNPEP domains; positively selected sites are indicated in red. (D) Ribbon diagram of ERAP1 with positively selected sites in orange (polymorphic) or green (fixed in humans); the active site is represented in yellow. (E) Ribbon diagram of THOP1 sites subject to positive selection in the human lineage highlighted in green. The active site is shown in yellow. (F) ribbon diagram of PSMF1; the dark grey helix indicates a motif important for protein stability. Positively selected sites are in orange or green depending on their being polymorphic or not, respectively, in humans.
(PDF)

Figure S4 Work-flow and main results for the intra-species analysis. Genes that were defined as targets of positive or balancing selection are shown in red.
(PDF)

Figure S5 Nucleotide diversity estimates for APP genes. π is plotted against θ_W. The dashed lines represent the 5th and 95th percentiles of a distribution of ~1000 randomly selected human genes, represented by grey dots.
(PDF)

Figure S6 DIND test and F_{ST} results. (A) The ratio between the ancestral and derived nucleotide diversity, $i\pi_A/i\pi_D$, is plotted against the derived allele frequency (DAF). The dashed line represents the 95th percentile of a distribution of ~1000 randomly selected human genes. The grey shaded areas represent frequency ranges where the ratio could not be calculated. (B) F_{ST} values are plotted against the minor allele frequency (MAF). The dashed lines represent the 95th and 99th percentiles of a distribution of SNPs deriving from ~1000 randomly selected human genes. Black crosses mark SNPs mentioned in the text which display F_{ST} values higher than the 95th percentile.
(PDF)

Figure S7 Analysis of positively selected sites in the *PSME3/CNTD1* region. Location of the most likely selection targets in *PSME3/CNTD1* region within the UCSC Genome Browser view. Relevant annotation tracks are shown. Variants in green represent both F_{ST} and DIND outliers in AS population.
(PDF)

Figure S8 Extended haplotype homozygosity (EHH) decay plots for variants showing a high lnRsb test.
(PDF)

Figure S9 Overlap between the signals we detected and those identified in previous scans of positive selection. Previously identified regions are represented as black bars and are tagged by author name and population showing selection signatures. The best candidate variants we identified in *NRD1* (upper panel) and *MARCH1* (lower panel) are also shown. Figure S9. Overlap between the signals we detected and those identified in previous scans of positive selection. Previously identified regions are represented as black bars and are tagged by author name and population showing selection signatures. The best candidate variants we identified in *NRD1* (upper panel) and *MARCH1* (lower panel) are also shown.
(PDF)

Figure S10 GENETREE analyses. Estimated haplotype trees for the LD sub-region of *CD207* (A), and for the sequenced regions of *NCF4* (B) and *TAP1* (C). Mutations are represented as black dots and named for their physical position along the region. The absolute frequency of each haplotype is also reported at the bottom of each lineage.
(PDF)

Table S1 List of analysed genes.
(PDF)

Table S2 Average non-synonymous/synonymous substitution rate ratio (dN/dS).
(PDF)

Table S3 Likelihood ratio test statistics for models of variable selective pressure among sites (F3x4 model of codon frequency).
(PDF)

Table S4 Likelihood ratio test statistics for models of variable selective pressure among sites (F61 model of codon frequency).
(PDF)

Table S5 Likelihood ratio test statistics for branch-site models (*CD207*, *CTSG*, and *CYBB*).
(PDF)

Table S6 SFS-based statistics calculated over whole gene regions using data from the 1000 Genomes Project.
(PDF)

Table S7 Nucleotide diversity and neutrality tests for *CTSB* and *PSMB9* gene regions.
(PDF)

Table S8 Nucleotide diversity and neutrality tests using low coverage 1000 Genomes Project data for the Sanger-resequenced regions.
(PDF)

Table S9 TMRCA estimates.
(PDF)

Table S10 Primer sequences.
(PDF)

Author Contributions

Conceived and designed the experiments: MS MCl. Performed the experiments: DF RC CT UP GF SR MCo MB. Analyzed the data: DF RC UP LDG GF MB MS NB GPC GM. Contributed reagents/materials/analysis tools: SLC FM GM. Wrote the paper: MS MCl DF RC LDG.

References

1. Neefjes J, Jongsma ML, Paul P, Bakke O. (2011) Towards a systems understanding of MHC class I and MHC class II antigen presentation. Nat Rev Immunol 11(12): 823–836.
2. Kessler JH, Khan S, Seifert U, Le Gall S, Chow KM, et al. (2011) Antigen processing by nardilysin and thimet oligopeptidase generates cytotoxic T cell epitopes. Nat Immunol 12(1): 45–53.
3. Savina A, Jancic C, Hugues S, Guermonprez P, Vargas P, et al. (2006) NOX2 controls phagosomal pH to regulate antigen processing during crosspresentation by dendritic cells. Cell 126(1): 205–218.
4. Hughes AL, Yeager M. (1998) Natural selection at major histocompatibility complex loci of vertebrates. Annu Rev Genet 32: 415–435.
5. Hansen TH, Bouvier M. (2009) MHC class I antigen presentation: Learning from viral evasion strategies. Nat Rev Immunol 9(7): 503–513.
6. Yang Z. (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol 24(8): 1586–1591.
7. Anisimova M, Nielsen R, Yang Z. (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164(3): 1229–1236.
8. Yang Z, Wong WS, Nielsen R. (2005) Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol 22(4): 1107–1118.
9. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, et al. (2012) Detecting individual sites subject to episodic diversifying selection. PLoS Genet 8(7): e1002764.
10. Kosakovsky Pond SL, Murrell B, Fourment M, Frost SD, Delport W, et al. (2011) A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol 28(11): 3033–3043.
11. Zhang J, Nielsen R, Yang Z. (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22(12): 2472–2479.
12. Anisimova M, Yang Z. (2007) Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol 24(5): 1219–1228.
13. Liu J, Shaji D, Cho S, Du W, Gervay-Hague J, et al. (2010) A threonine-based targeting signal in the human CD1d cytoplasmic tail controls its functional expression. J Immunol 184(9): 4973–4981.
14. Feinberg H, Taylor ME, Razi N, McBride R, Knirel YA, et al. (2011) Structural basis for langerin recognition of diverse pathogen and mammalian glycans through a single binding site. J Mol Biol 405(4): 1027–1039.
15. Feinberg H, Powlesland AS, Taylor ME, Weis WI. (2010) Trimeric structure of langerin. J Biol Chem 285(17): 13285–13293.
16. Verdijk P, Dijkman R, Plasmeijer EI, Mulder AA, Zoutman WH, et al. (2005) A lack of birbeck granules in langerhans cells is associated with a naturally occurring point mutation in the human langerin gene. J Invest Dermatol 124(4): 714–717.
17. Paclet MH, Henderson LM, Campion Y, Morel F, Dagher MC. (2004) Localization of Nox2 N-terminus using polyclonal antipeptide antibodies. Biochem J 382(Pt 3): 981–986.
18. Wallach TM, Segal AW. (1997) Analysis of glycosylation sites on gp91phox, the flavocytochrome of the NADPH oxidase, by site-directed mutagenesis and translation in vitro. Biochem J 321 (Pt 3): 583–585.
19. Royer-Pokora B, Kunkel LM, Monaco AP, Goff SC, Newburger PE, et al. (1986) Cloning the gene for an inherited human disorder–chronic granulomatous disease–on the basis of its chromosomal location. Nature 322(6074): 32–38.
20. Bustamante J, Arias AA, Vogt G, Picard C, Galicia LB, et al. (2011) Germline CYBB mutations that selectively affect macrophages in kindreds with X-linked predisposition to tuberculous mycobacterial disease. Nat Immunol 12(3): 213–221.
21. Schrodt S, Koch J, Tampe R. (2006) Membrane topology of the transporter associated with antigen processing (TAP1) within an assembled functional peptide-loading complex. J Biol Chem 281(10): 6455–6462.
22. Koch J, Guntrum R, Heintke S, Kyritsis C, Tampe R. (2004) Functional dissection of the transmembrane domains of the transporter associated with antigen processing (TAP). J Biol Chem 279(11): 10142–10147.
23. Nijenhuis M, Hammerling GJ. (1996) Multiple regions of the transporter associated with antigen processing (TAP) contribute to its peptide binding site. J Immunol 157(12): 5467–5477.
24. Dong G, Wearsch PA, Peaper DR, Cresswell P, Reinisch KM. (2009) Insights into MHC class I peptide loading from the structure of the tapasin-ERp57 thiol oxidoreductase heterodimer. Immunity 30(1): 21–32.
25. de Garavilla L, Greco MN, Sukumar N, Chen ZW, Pineda AO, et al. (2005) A novel, potent dual inhibitor of the leukocyte proteases cathepsin G and chymase: Molecular mechanisms and anti-inflammatory activity in vivo. J Biol Chem 280(18): 18001–18007.
26. Ascher DB, Cromer BA, Morton CJ, Volitakis I, Cherny RA, et al. (2011) Regulation of insulin-regulated membrane aminopeptidase activity by its C-terminal domain. Biochemistry 50(13): 2611–2622.
27. Ewenkel C, Wolf DH. (1993) BLH1 codes for a yeast thiol aminopeptidase, the equivalent of mammalian bleomycin hydrolase. J Biol Chem 268(10): 7036–7043.
28. Kim W, Bennett EJ, Huttlin EL, Guo A, Li J, et al. (2011) Systematic and quantitative assessment of the ubiquitin-modified proteome. Mol Cell 44(2): 325–340.
29. Choudhary C, Kumar C, Gnad F, Nielsen ML, Rehman M, et al. (2009) Lysine acetylation targets protein complexes and co-regulates major cellular functions. Science 325(5942): 834–840.
30. Wilson DJ, Hernandez RD, Andolfatto P, Przeworski M. (2011) A population genetics-phylogenetics approach to inferring natural selection in coding sequences. PLoS Genet 7(12): e1002395.
31. Genomes Project Consortium, Durbin RM, Abecasis GR, Altshuler DL, Auton A, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467(7319): 1061–1073.
32. Dall E, Brandstetter H. (2013) Mechanistic and structural studies on legumain explain its zymogenicity, distinct activation pathways, and regulation. Proc Natl Acad Sci U S A 110(27): 10940–10945.
33. Cagliani R, Riva S, Biasin M, Fumagalli M, Pozzoli U, et al. (2010) Genetic diversity at endoplasmic reticulum aminopeptidases is maintained by balancing selection and is associated with natural resistance to HIV-1 infection. Hum Mol Genet 19: 4705–14. doi: 10.1093/hmg/ddq401.
34. Kirk R, Laman H, Knowles PP, Murray-Rust J, Lomonosov M, et al. (2008) Structure of a conserved dimerization domain within the F-box protein Fbxo7 and the PI31 proteasome inhibitor. J Biol Chem 283(32): 22325–22335.
35. Ray K, Hines CS, Coll-Rodriguez J, Rodgers DW. (2004) Crystal structure of human thimet oligopeptidase provides insight into substrate recognition, regulation, and localization. J Biol Chem 279(19): 20480–20489.
36. Watterson GA. (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7(2): 256–276.
37. Nei M, Li WH. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A 76(10): 5269–5273.
38. Tajima F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3): 585–595.
39. Zeng K, Fu YX, Shi S, Wu CI. (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 174(3): 1431–1439.
40. Fu YX, Li WH. (1993) Statistical tests of neutrality of mutations. Genetics 133(3): 693–709.
41. Wright S. (1950) Genetical structure of populations. Nature 166(4215): 247–249.
42. Barreiro LB, Ben-Ali M, Quach H, Laval G, Patin E, et al. (2009) Evolutionary dynamics of human toll-like receptors and their different contributions to host defense. PLoS Genet 5(7): e1000562.
43. Cagliani R, Riva S, Pozzoli U, Fumagalli M, Comi GP, et al. (2011) Balancing selection is common in the extended MHC region but most alleles with opposite risk profile for autoimmune diseases are neutrally evolving. BMC Evol Biol 11: 171-2148-11-171.
44. Andres AM, Dennis MY, Kretzschmar WW, Cannons JL, Lee-Lin SQ, et al. (2010) Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation. PLoS Genet 6(10): e1001157.

45. Tang K, Thornton KR, Stoneking M. (2007) A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol 5(7): e171.
46. Voight BF, Kudaravalli S, Wen X, Pritchard JK. (2006) A map of recent positive selection in the human genome. PLoS Biol 4(3): e72.
47. Kenny EE, Pe'er I, Karban A, Ozelius L, Mitchell AA, et al. (2012) A genome-wide scan of ashkenazi jewish crohn's disease suggests novel susceptibility loci. PLoS Genet 8(3): e1002559.
48. Gargiulo G, Levy S, Bucci G, Romanenghi M, Fornasari L, et al. (2009) NA-seq: A discovery tool for the analysis of chromatin structure and dynamics during differentiation. Dev Cell 16(3): 466–481.
49. Liang L, Morar N, Dixon AL, Lathrop GM, Abecasis GR, et al. (2013) A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. Genome Res 23(4): 716–726.
50. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, et al. (2007) Localizing recent adaptive evolution in the human genome. PLoS Genet 3(6): e90.
51. Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, et al. (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327(5967): 883–886.
52. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449(7164): 913–918.
53. Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM. (2006) Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res 16(8): 980–989.
54. Kimura R, Fujimoto A, Tokunaga K, Ohashi J. (2007) A practical genome scan for population-specific strong selective sweeps that have reached fixation. PLoS One 2(3): e286.
55. Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, et al. (2005) Genomic regions exhibiting positive selection identified from dense genotype data. Genome Res 15(11): 1553–1565.
56. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. (2008) Natural selection has driven population differentiation in modern humans. Nat Genet 40(3): 340–345.
57. Charlesworth D. (2006) Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet 2(4): e64.
58. Wright SI, Charlesworth B. (2004) The HKA test revisited: A maximum-likelihood-ratio test of the standard neutral model. Genetics 168(2): 1071–1076.
59. Garrigan D, Hammer MF. (2006) Reconstructing human origins in the genomic era. Nat Rev Genet 7(9): 669–680.
60. Ward EM, Stambach NS, Drickamer K, Taylor ME. (2006) Polymorphisms in human langerin affect stability and sugar binding activity. J Biol Chem 281(22): 15450–15456.
61. Tateno H, Ohnishi K, Yabe R, Hayatsu N, Sato T, et al. (2010) Dual specificity of langerin to sulfated and mannosylated glycans via a single C-type carbohydrate recognition domain. J Biol Chem 285(9): 6390–6400.
62. de Witte L, Nabatov A, Pion M, Fluitsma D, de Jong MA, et al. (2007) Langerin is a natural barrier to HIV-1 transmission by langerhans cells. Nat Med 13(3): 367–371.
63. Miyazawa M, Lopalco L, Mazzotta F, Lo Caputo S, Veas F, et al. (2009) The 'immunologic advantage' of HIV-exposed seronegative individuals. Aids 23(2): 161–175.
64. Wittke-Thompson JK, Pluzhnikov A, Cox NJ. (2005) Rational inferences about departures from hardy-weinberg equilibrium. Am J Hum Genet 76(6): 967–986.
65. Mitchell PS, Patzina C, Emerman M, Haller O, Malik HS, et al. (2012) Evolution-guided identification of antiviral specificity determinants in the broadly acting interferon-induced innate immunity factor MxA. Cell Host Microbe 12(4): 598–604.
66. Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, et al. (2008) Patterns of positive selection in six mammalian genomes. PLoS Genet 4(8): e1000144.
67. Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, et al. (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3(6): e170.
68. Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, et al. (2013) Identifying recent adaptations in large-scale genomic data. Cell 152(4): 703–713.
69. Fu W, Akey JM. (2013) Selection and adaptation in the human genome. Annu Rev Genomics Hum Genet 14: 467–489.
70. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, et al. (2006) Positive natural selection in the human lineage. Science 312(5780): 1614–1620.
71. Andres AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, et al. (2009) Targets of balancing selection in the human genome. Mol Biol Evol 26(12): 2755–2764.
72. Manry J, Laval G, Patin E, Fornarino S, Itan Y, et al. (2011) Evolutionary genetic dissection of human interferons. J Exp Med 208(13): 2747–2759.
73. Forni D, Cagliani R, Pozzoli U, Colleoni M, Riva S, et al. (2013) A 175 million year history of T cell regulatory molecules reveals widespread selection, with adaptive evolution of disease alleles. Immunity 38(6): 1129–1141.
74. Vasseur E, Boniotto M, Patin E, Laval G, Quach H, et al. (2012) The evolutionary landscape of cytosolic microbial sensors in humans. Am J Hum Genet 91(1): 27–37.
75. Horst D, Geerdink RJ, Gram AM, Stoppelenburg AJ, Ressing ME. (2012) Hiding lipid presentation: Viral interference with CD1d-restricted invariant natural killer T (iNKT) cell activation. Viruses 4(10): 2379–2399.
76. Ressing ME, Luteijn RD, Horst D, Wiertz EJ. (2012) Viral interference with antigen presentation: Trapping TAP. Mol Immunol 55: 139–42. doi: 10.1016/j.molimm.2012.10.009.
77. Momburg F, Roelse J, Howard JC, Butcher GW, Hammerling GJ, et al. (1994) Selectivity of MHC-encoded peptide transporters from human, mouse and rat. Nature 367(6464): 648–651.
78. Kim E, Kwak H, Ahn K. (2009) Cytosolic aminopeptidases influence MHC class I-mediated antigen presentation in an allele-dependent manner. J Immunol 183(11): 7379–7387.
79. Evnouchidou I, Kamal RP, Seregin SS, Goto Y, Tsujimoto M, et al. (2011) Coding single nucleotide polymorphisms of endoplasmic reticulum aminopeptidase 1 can affect antigenic peptide generation in vitro by influencing basic enzymatic properties of the enzyme. J Immunol 186(4): 1909–1913.
80. Fierabracci A, Milillo A, Locatelli F, Fruci D. (2012) The putative role of endoplasmic reticulum aminopeptidases in autoimmunity: Insights from genomic-wide association studies. Autoimmun Rev 12(2): 281–288.
81. Reeves EP, Lu H, Jacobs HL, Messina CG, Bolsover S, et al. (2002) Killing activity of neutrophils is mediated through activation of proteases by K+ flux. Nature 416(6878): 291–297.
82. Averhoff P, Kolbe M, Zychlinsky A, Weinrauch Y. (2008) Single residue determines the specificity of neutrophil elastase for shigella virulence factors. J Mol Biol 377(4): 1053–1066.
83. Li DN, Matthews SP, Antoniou AN, Mazzeo D, Watts C. (2003) Multistep autoactivation of asparaginyl endopeptidase in vitro and in vivo. J Biol Chem 278(40): 38980–38990.
84. Sanders RW, Venturi M, Schiffner L, Kalyanaraman R, Katinger H, et al. (2002) The mannose-dependent epitope for neutralizing antibody 2G12 on human immunodeficiency virus type 1 glycoprotein gp120. J Virol 76(14): 7293–7305.
85. Kawashima T, Bao YC, Nomura Y, Moon Y, Tonozuka Y, et al. (2006) Rac1 and a GTPase-activating protein, MgcRacGAP, are required for nuclear translocation of STAT transcription factors. J Cell Biol 175(6): 937–946.
86. Lambeth JD. (2004) NOX enzymes and the biology of reactive oxygen. Nat Rev Immunol 4(3): 181–189.
87. Vernot B, Stergachis AB, Maurano MT, Vierstra J, Neph S, et al. (2012) Personal and population genomics of human regulatory variation. Genome Res 22(9): 1689–1697.
88. Rohrlich PS, Fazilleau N, Ginhoux F, Firat H, Michel F, et al. (2005) Direct recognition by alphabeta cytolytic T cells of hfe, a MHC class ib molecule without antigen-presenting function. Proc Natl Acad Sci U S A 102(36): 12855–12860.
89. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, et al. (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19(2): 327–335.
90. Wernersson R, Pedersen AG. (2003) RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res 31(13): 3537–3539.
91. Anisimova M, Bielawski JP, Yang Z. (2002) Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol 19(6): 950–958.
92. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. (2006) Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol 23(10): 1891–1901.
93. Kosakovsky Pond SL, Frost SD. (2005) Not so different after all: A comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22(5): 1208–1222.
94. Delport W, Poon AF, Frost SD, Kosakovsky Pond SL. (2010) Datamonkey 2010: A suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 26(19): 2455–2457.
95. Tina KG, Bhadra R, Srinivasan N. (2007) PIC: Protein interactions calculator. Nucleic Acids Res 35(Web Server issue): W473–6.
96. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, et al. (2005) The FoldX web server: An online force field. Nucleic Acids Res 33(Web Server issue): W382–8.
97. Dehouck Y, Kwasigroch JM, Gilis D, Rooman M. (2011) PoPMuSiC 2.1: A web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinformatics 12: 151-2105-12-151.
98. Capriotti E, Fariselli P, Casadio R. (2005) I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 33(Web Server issue): W306–10.
99. Cereda M, Sironi M, Cavalleri M, Pozzoli U. (2011) GeCo++: A C++ library for genomic features computation and annotation in the presence of variants. Bioinformatics 27(9): 1313–1315.
100. Thornton K. (2003) Libsequence: A C++ class library for evolutionary genetic analysis. Bioinformatics 19(17): 2325–2327.
101. Fay JC, Wu CI. (2000) Hitchhiking under positive darwinian selection. Genetics 155(3): 1405–1413.
102. Gautier M, Vitalis R. (2012) Rehh: An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics 28(8): 1176–1177.

103. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, et al. (2005) Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15(11): 1576–1583.
104. Hudson RR. (2001) Two-locus sampling distributions and their application. Genetics 159(4): 1805–1817.
105. Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi GP, et al. (2009) Widespread balancing selection and pathogen-driven selection at blood group antigen genes. Genome Res 19(2): 199–212.
106. Stephens M, Smith NJ, Donnelly P. (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68(4): 978–989.
107. Stephens M, Scheet P. (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76(3): 449–462.
108. Bandelt HJ, Forster P, Rohl A. (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16(1): 37–48.
109. Griffiths RC, Tavare S. (1995) Unrooted genealogical tree probabilities in the infinitely-many-sites model. Math Biosci 127(1): 77–98.
110. Griffiths RC, Tavare S. (1994) Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc Lond B Biol Sci 344(1310): 403–410.
111. Evans PD, Gilbert SL, Mekel-Bobrov N, Vallender EJ, Anderson JR, et al. (2005) Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. Science 309(5741): 1717–1720.
112. Thomson R, Pritchard JK, Shen P, Oefner PJ, Feldman MW. (2000) Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. Proc Natl Acad Sci U S A 97(13): 7360–7365.
113. Glazko GV, Nei M. (2003) Estimation of divergence times for major lineages of primate species. Mol Biol Evol 20(3): 424–434.
114. Samson M, Libert F, Doranz BJ, Rucker J, Liesnard C, et al. (1996) Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. Nature 382(6593): 722–725.
115. Plummer FA, Ball TB, Kimani J, Fowke KR. (1999) Resistance to HIV-1 infection among highly exposed sex workers in nairobi: What mediates protection and why does it develop? Immunol Lett 66(1–3): 27–34.
116. Fowke KR, Nagelkerke NJ, Kimani J, Simonsen JN, Anzala AO, et al. (1996) Resistance to HIV-1 infection among persistently seronegative prostitutes in nairobi, kenya. Lancet 348(9038): 1347–1351.
117. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3): 559–575.

3.3 Evolutionary analysis of the contact system indicates that kininogen evolved adaptively in mammals and in human populations

Evolutionary Analysis of the Contact System Indicates that Kininogen Evolved Adaptively in Mammals and in Human Populations

Rachele Cagliani,*[1] Diego Forni,[1] Stefania Riva,[1] Uberto Pozzoli,[1] Marta Colleoni,[1] Nereo Bresolin,[1,2] Mario Clerici,[3,4] and Manuela Sironi[1]

[1]Bioinformatics, Scientific Institute IRCCS E. Medea, Bosisio Parini (LC), Italy

[2]Dino Ferrari Centre, Department of Physiopathology and Transplantation, University of Milan, Fondazione Ca' Granda IRCCS Ospedale Maggiore Policlinico, Milan, Italy

[3]Chair of Immunology, Department of Physiopathology and Transplantation, University of Milan, Milano, Italy

[4]Don C. Gnocchi Foundation ONLUS, IRCCS, Milan, Italy

***Corresponding author:** E-mail: rachele.cagliani@bp.lnf.it.

Associate editor: Naoko Takezaki

Abstract

Activation of the contact system leads to the cleavage of kininogen by plasma kallikrein resulting in kinin release and in the initiation of the intrinsic pathway of coagulation. Proteolysis of kininogen also generates antimicrobial peptides (AMPs) and can be induced by diverse pathogens. Thus, the contact system is regarded as a branch of innate immunity. We performed an evolutionary analysis of contact system genes by analyzing both inter- and intraspecies diversity. Results indicated that mammalian kininogen genes evolved adaptively. Positively selected sites are located in all protein domains with the exclusion of the bradykinin region and also involve AMP sequences (including the highly effective NAT26 peptide); positively selected sites also occur at alternative cleavage sites for neutrophil-released kinins. Population genetic analysis in humans indicated that a region of the kininogen gene (*KNG1*) has been a target of long-standing multiallelic balancing selection and that the coalescence time of the haplotype phylogeny dates back to the split between the humans and chimpanzees. No selection signature was detected in the *Pan troglodytes KNG1* gene or in human genes encoding other components of the contact system. The selection targets in human *KNG1* might be accounted for by variants with transcriptional regulatory activity. Results herein indicate a continuum in selective pressure acting on different timescales and targeting *KNG1*. This is in line with evidences suggesting a central role for kininogen in modulating of immune response and with its being a target of an extremely diverse array of pathogen species.

Key words: **contact system genes, *KNG1*, positive selection, balancing selection, innate immunity.**

Introduction

In vertebrates, a complex system of plasma enzymes has evolved to limit blood loss from damaged blood vessels through formation of a fibrin clot. In all vertebrates, the coagulation cascade includes a series of proteolytic reactions that culminate in the generation of thrombin and in the cleavage of fibrinogen to form fibrin (fig. 1) (Jiang and Doolittle 2003). Initiation of fibrin formation through the "extrinsic pathway" occurs when plasma factor VIIa forms a complex with the integral membrane protein tissue factor, which is exposed upon endothelial injury (fig. 1). Alternatively, coagulation may be initiated through the "intrinsic pathway," also referred to as the contact system. Four proteins represent the core components of the contact system; three of them, factor XI (FXI), factor XII (FXII), and plasma kallikrein (PK), are encoded in humans by the *F11*, *F12*, and *KLKB1* genes, respectively, and act as proteases. The fourth component, high-molecular-weight kininogen, is a nonenzymatic glycoprotein (fig. 1). Kininogen is encoded by the *KNG1* gene, which gives rise to two alternatively spliced products differing in their terminal exons; the two transcripts originate from high- and low-molecular-weight kininogen. High-molecular-weight kininogen circulates in human plasma forming complexes with FXI and PK (Muller-Esterl 1989). Interaction of these complexes with negatively charged surfaces activates the system: FXII and prekallikrein are proteolytically cleaved to the active forms, FXIIa and kallikrein, respectively. As a result, FXI is activated, leading to the initiation of a series of enzymatic reactions that lead to fibrin formation. In the process, high-molecular-weight kininogen is cleaved by kallikrein releasing the nonapeptide bradykinin (BK) (Colman and Schmaier 1997) (fig. 1). BK, acts as a vasoactive and proinflammatory peptide, increases the production of nitric oxide and prostaglandins, and causes increased vascular permeability, hypotension, smooth-muscle contraction, and fever. Kininogen can also be cleaved by proteases other than PK, including tissue kallikrein and neutrophil-derived proteinase 3 to release kinins related to BK, namely Lys-BK and Met-Lys-BK-Ser-Ser (Kahn et al. 2009) (fig. 1).

Article

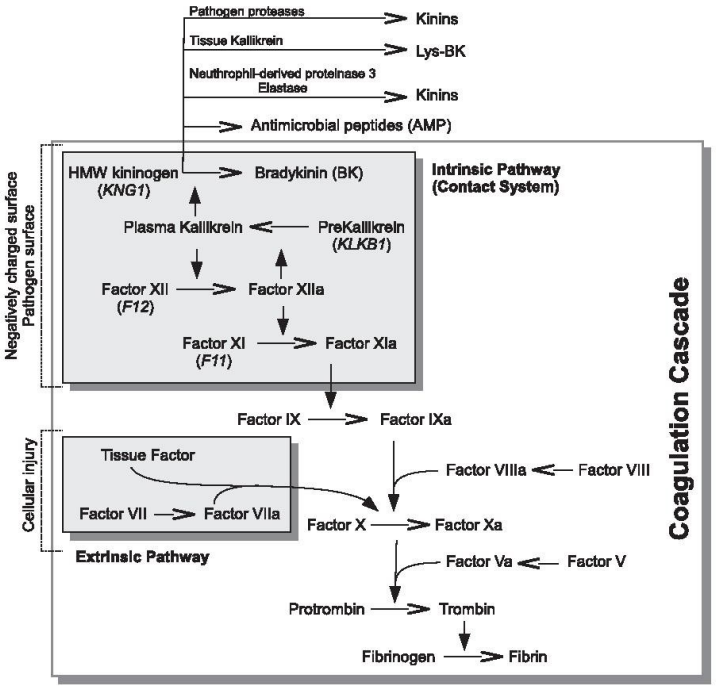


FIG. 1. Schematic representation of the contact system within the coagulation cascade. The intrinsic (contact system) and extrinsic pathways that initiate the coagulation cascade are shown within gray boxes. The official human gene symbol is reported in brackets below each contact system component. Additional endogenous and exogenous components that impinge on kininogen are shown external to the coagulation pathway.

Therefore, the contact system represents a link between the coagulation and inflammatory responses, two systems central to host survival in the face of tissue damage and infection. Nonetheless, the intrinsic coagulation pathway seems to play a minor role in the control of hemostasis, suggesting that its major physiologic functions are exerted by kinin release and include induction of inflammatory responses and regulation of blood pressure (Campbell 2001).

Several evidences have also indicated that the contact system may be regarded as an important branch of innate immunity. In fact, contact system components can bind to the surface of pathogenic Gram-positive and Gram-negative bacteria and become activated, whereby releasing BK and antimicrobial peptides (AMPs) (fig. 1). AMPs are derived from the proteolysis of kininogen domains D3 and $D5_H$ and are active against a wide range of bacterial species (Nordahl et al. 2005; Frick et al. 2006). Moreover, kinins can activate immune responses by different strategies, including recruitment of neutrophils (Paegelow et al. 2002), stimulation of alveolar macrophages (Sato et al. 1996), and induction of dendritic cell maturation via the BK receptor 2 (Scharfstein et al. 2007). As a consequence of these properties, activation of the contact system potentiates the host response against invading pathogens. Nevertheless, BK may also exert adverse effects during infection by inducing hypotension and vascular leaks, eventually contributing to the pathogenesis of sepsis (reviewed in Nickel and Renne [2012]). Moreover, increased vascular permeability potentially facilitates the systemic spread of the infectious agent. Indeed, several pathogens encode proteins that bind contact system components and/or proteases that mediate kininogen proteolysis. One of the first identified microbial kininogenases is cruzipain, a *Trypanosoma cruzi* virulence factor (Del Nery et al. 1997). Cruzipain is encoded by multiple polymorphic genes, and its cleavage of high-molecular-weight kininogen releases Lys-BK, which, in turn, facilitates host cell invasion (Scharfstein et al. 2000). After *T. cruzi*, several pathogens were found to encode kininogenases: These include

Plasmodium falciparum (Bagnaresi et al. 2012), the causative agent of malaria, parasitic worms such as *Schistosoma mansoni* and *Fasciola hepatica* (Carvalho et al. 1998; Cordova et al. 2001), fungi of the *Candida* genus (Rapala-Kozik et al. 2010), and several bacteria species such as *Streptococcus pyogenes* (Herwald et al. 1996), *Staphylococcus aureus* (Imamura et al. 2005), and *Porphyromonas gingivalis* (Imamura et al. 1995).

These observations, and the role of kininogen as a source of AMPs, suggest that contact system genes, and *KNG1* in particular, might have been engaged in host–pathogen genetic conflict.

Results

Evolutionary Analysis of Contact System Genes in Mammals

To investigate the evolutionary history of contact system genes in mammals, we retrieved coding sequence information for all available species from the Ensembl database (http://www.ensembl.org/index.html, last accessed November 3, 2012) for *KNG1*, *KLKB1*, *F11*, and *F12*. Because, due to alternative splicing, *KNG1* originates two alternative products that share a common proximal region, the alignment was split into three portions covering the common region (domains D1–D4, $KNG1_{D1-D4}$), the low-molecular-weight specific domain ($D5_L$ and $KNG1_{D5L}$), and the high-molecular-weight unique portion ($D5_H$-D6 and $KNG1_{D5H-D6}$) (fig. 2). A comparison with the genome-wide distributions of pairwise comparisons (human–macaque, human–mouse, and human–dog) revealed the $KNG1_{D1-D4}$ and $KNG1_{D5H-D6}$ ($KNG1_{D5L}$ was not analyzed due to its short size) tend to have very high dN/dS values, whereas *F11*, *F12*, and *KLKB1* have dN/dS ratios comparable to most genes (supplementary fig. S1, Supplementary Material online).

We next screened the multiple sequence alignments (including all available mammalian species) for evidences of recombination using Genetic Algorithm Recombination Detection (GARD) (Kosakovsky Pond et al. 2006); this analysis uncovered the presence of recombination breakpoints in $KNG1_{D5H-D6}$, *KLKB1*, and *F11* (one breakpoint/alignment) (fig. 2). In the case of $KNG1_{D5H-D6}$, we excluded a region surrounding the breakpoint as it showed limited homology across species. Taking GARD results into account, we calculated the average nonsynonymous (dN) to synonymous (dS) substitution rate ratio (dN/dS) for all alignments. In all cases, dN/dS was lower than 1, with the exception of the very short (29 aligned amino acids) $KNG1_{D5L}$ region (dN/dS = 1.10382, 95% confidence interval [CI] = 0.88–1.36) (supplementary table S1, Supplementary Material online). To formally test whether contact system genes have evolved adaptively in mammals, we used the codeml program to compare models of gene evolution that allow (NSsite models M2a and M8) or disallow (NSsite models M1a and M8a) a class of codons to evolve with $dN/dS > 1$ (Yang 2007). Specifically, these models were applied after dividing the alignments showing evidence of recombination into halves based on the location of the recombination breakpoints. Results indicated that for all alignments covering *KNG1*, the two null models were rejected in favor of the positive selection models (supplementary table S2, Supplementary Material online). For the remaining genes, the M1a/M2a and M8a/M8 comparisons did not support positive selection (not shown). Thus, only *KNG1* can be reliably considered as a target of positive selection, and all gene domains showed evidences of adaptive evolution in mammals (supplementary table S2, Supplementary Material online).

To identify specific sites subject to positive selection in *KNG1*, we applied the Bayes empirical Bayes (BEB) method (with a cutoff of 0.90) from M8 (Anisimova et al. 2002; Yang et al. 2005). Because this approach may yield some false positives when a relatively large number of sequences is analyzed (Kosakovsky Pond and Frost 2005), we used the mixed effects model of evolution (MEME) (with the default cutoff of 0.1) (Murrell et al. 2012) as a second criterion. Thus, only sites detected by both methods were considered to be positively selected, although this may result in an underestimation of the actual number of selected sites. Using these criteria, several positively selected sites were identified, and these are scattered across the whole sequences of *KNG1* with the exclusion of the conserved BK sequence (fig. 2). Cleavage of kininogen domain D3 has been shown to originate AMPs, and several positively selected sites were found to be located within these peptides. Domain $D5_H$ also gives rise to antimicrobial molecules deriving from the proteolysis of the histidine-rich region, which could be only partially aligned in the species we analyzed (fig. 2); one of these antibacterial peptides (GKH17) encompasses one positively selected residue. Three additional selected sites flank the BK region and rim an alternative cleavage site that originates from Met-Lys-BK-Ser-Ser, produced by neutrophils (Kahn et al. 2009) (fig. 2). The binding sites for F11 and PK in domain $D6_H$ also display sites targeted by selection (fig. 2). Finally, residue 577 (Ser in humans) represents a predicted O-glycosylation site (http://www.uniprot.org, last accessed November 3, 2012); interestingly, minor changes in the glycosylation state of high-molecular-weight kininogen have been shown to result in faster cleavage by PK and higher BK production in rats susceptible to chronic intestinal and systemic inflammation (Isordia-Salas et al. 2003).

Population Genetic Analysis in Humans

KNG1, *KLKB1*, *F11*, and *F12* have been almost fully resequenced within the SeattleSNPs Variation Discovery Resource Program (http://pga.gs.washington.edu/, last accessed November 3, 2012). In particular, for each gene, a region encompassing the whole transcription unit plus flanking genomic sequences has been sequenced in 24 African American (AA) and 23 European (EU) subjects (SeattleSNPs DNA Panel 1), with only small sequencing gaps scattered along the introns.

We exploited the availability of these data to calculate θ_W, an estimate of the expected per site heterozygosity (Watterson 1975), and π, the average number of pairwise sequence nucleotide differences between haplotypes (Nei and Li 1979). To compare the values we obtained for contact

FIG. 2. Analysis of positively selected sites in KNG1. The domain structures of low- and high-molecular-weight kininogen are shown (not to scale). Domain color codes are as follows: cyan, cystatin (domains D1–D3); green, BK (D4); yellow, unique light chain of low-molecular-weight kininogen ($D5_L$); and orange, unique light chain of high-molecular-weight kininogen ($D5_H$ and D6). The shaded gray region in $D5_H$ denotes a portion of the histidine-rich region that was excluded from analysis due to poor alignment. The positions (relative to the human sequence) of positively selected sites are shown on the kininogen structures; the location of the recombination breakpoint (black diamond) is reported on the domain structure. The residue in red is a predicted O-glycosylation site in humans. Portions of the multiple species alignments are shown. In all panels, positively selected sites are in red, and cysteines involved in the formation of disulfide bonds are in blue. The upper alignment covers most of D3 and D4; the BK region is shaded in green and the position of AMPs (GKD25, GCP28, NAT26, RET27, and LDC27), is shown. Alignments in the lower portion cover a region in $D5_H$ (left) involved in cell binding, a portion of the histidine-rich region encoding AMPs (GKH17 and HKH20), and the C-terminal portion that carries binding sites for F11 and PK. AMP designation is based on the three initial residues and on the peptide length, as proposed previously (Nordahl et al. 2005; Frick et al. 2006).

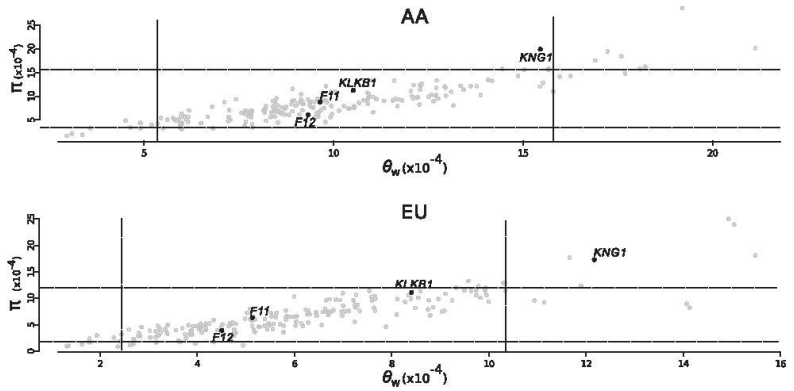


FIG. 3. Nucleotide diversity estimates for contact system genes. π is plotted against θ_W. The dashed vertical and horizontal lines represent the 5th and 95th percentiles of a distribution of 201 genes (gray dots) resequenced within the SeattleSNPs Variation Discovery Resource Program. Average values ($\times 10^{-4}$) for θ_W and π: AA, 9.78 and 8.44; EU, 6.08 and 6.62, respectively.

system genes, we calculated θ_W and π for 201 genes resequenced in the same individuals (Panel 1) by the SeattleSNPs Program. These θ_W and π estimates were used to obtain an empirical distribution, and values higher than the 95th percentile were considered significant. No significant values of θ_W and π were observed for *KLKB1*, *F11*, and *F12* (fig. 3); conversely, for *KNG1* both indexes were high in AA and EU, although the rank of θ_W in AA did not reach the 95th percentile (fig. 3).

High nucleotide diversity might suggest the action of balancing selection—that is, a process whereby genetic variability is maintained in populations due to some selective pressure. As a result of mutation and recombination, balancing selection signatures tend to extend over relatively short genomic intervals (Wiuf et al. 2004; Charlesworth 2006). Thus, to further explore the possible role of balancing selection in shaping diversity at *KNG1*, we divided the gene region into three continuous subregions of approximately 8 kb (*KNG1*-r1, -r2, and -r3) (fig. 4*a*). For each subregion, we calculated nucleotide diversity and compared the values with a distribution of θ_W and π calculated for 5-kb windows (hereafter referred to as reference windows) deriving from 238 genes resequenced by the National Institute of Environmental Health Sciences (NIEHS) Program in the same populations. The percentile ranks corresponding to *KNG1* regions in the distribution of NIEHS gene values are reported in table 1 and indicate that *KNG1-r1* displays extremely high nucleotide diversity in AA and CEU; conversely, no significant values are observed for *KNG1*-r2 and *KNG1*-r3.

To gain further insight into the evolutionary pattern of *KNG1* in human populations, we fully resequenced the region encompassing exons 1–4 in two additional HapMap populations, namely Yoruba (YRI) and East Asians (AS) (fig. 4*a*). Both θ_W and π displayed values higher than the 98th percentile in the distribution of 5-kb reference windows in these populations, as well (table 1).

An effect of balancing selection is a distortion of the site frequency spectrum (SFS) toward intermediate frequency alleles. Common neutrality tests based on the SFS include Tajima's D (D_T) (Tajima 1989) and Fu and Li's D^* and F^* (Fu and Li 1993). Because, population history, in addition to selective processes, is known to affect the SFS, the significance of neutrality tests was evaluated by performing coalescent simulations with population genetics models that incorporate demographic scenarios (see Materials and Methods). As above, we also applied an empirical comparison by calculating the percentile rank of D_T, F^*, and D^* in the *KNG1*-r1 relative to 5-kb reference windows. Neutrality tests indicated departure from neutrality with significantly positive values for most statistics in all populations (table 1).

As mentioned earlier, our data (table 1) indicate that nucleotide diversity indexes are extremely high for the analyzed *KNG1* gene region in all populations. To confirm this observation, we applied a multilocus maximum-likelihood Hudson–Kreitman–Aguadé (MLHKA) test (Wright and Charlesworth 2004) by comparing polymorphism and divergence levels at the *KNG1*-r1 with 16 NIEHS genes resequenced in the same populations (AA, YRI, EU, and AS). Results, summarized in table 2, indicate that a significant excess of nucleotide diversity versus interspecies divergence is detectable in all populations for the *KNG1* study region.

Further insight into the evolutionary history of a gene region can be gained by inferring haplotype genealogies. The presence of recombination may yield unreliable genealogies and affect inference of coalescent times. Thus, we selected a subregion in *KNG1*-r1 based on linkage disequilibrium (LD); in particular, we used data from a 2-kb region (National Center for Biotechnology Information [NCBI]/hg18

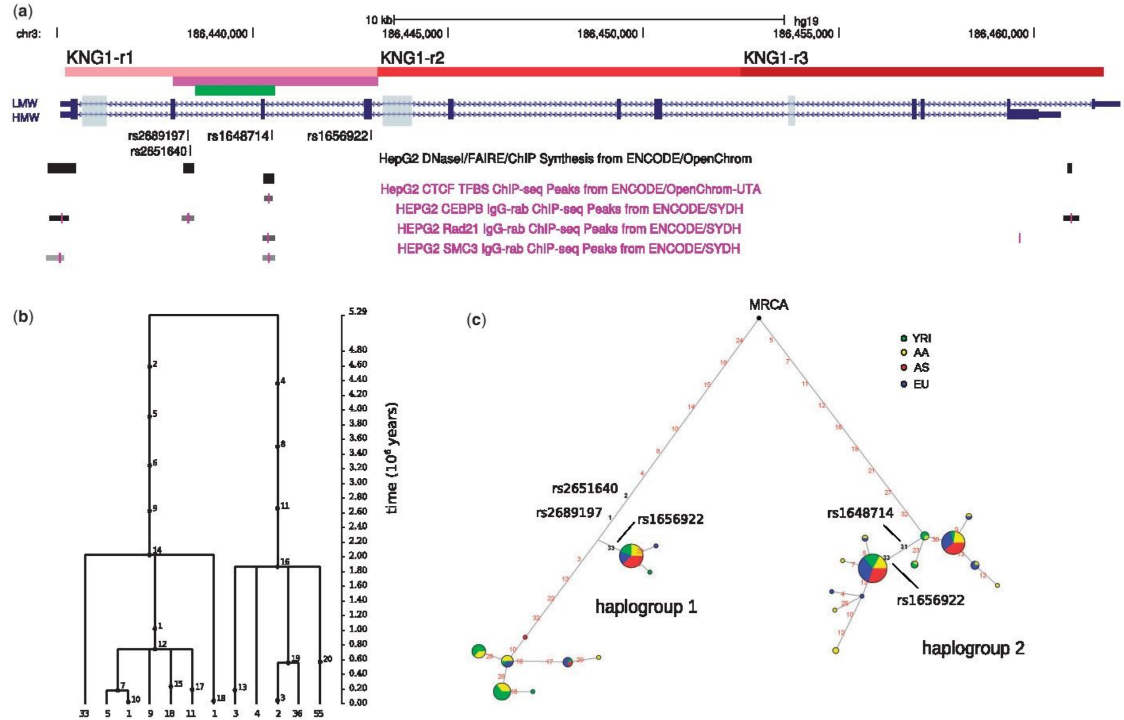


Fig. 4. *KNG1* haplotype analysis. (*a*) Schematic representation of the *KNG1* gene. Both splicing isoforms (originating low-and high-molecular-weight kininogen) are represented. The three regions the *KNG1* genomic region was divided into for population genetic analysis are indicated with different shades of red. The gray regions indicate resequencing gaps in the SeattleSNPs Program data. The subregions within *KNG1-r1* used to perform haplotype analysis (green line) and resequenced in chimpanzees (magenta line) are also shown. Polymorphisms mentioned in the text are shown. ENCODE tracks in HepG2 cells are visualized (as derived from the University of California–Santa Cruz Genome Browser). (*b*) GENETREE analysis. Mutations are represented as black dots and named for their physical position along the region. The absolute frequency of each haplotype is also reported at the bottom of each lineage. (*c*) Median-joining network: The genealogy was reconstructed for the 2-kb *KNG1* region through a median-joining network; each node represents a different haplotype, with the size of the circle proportional to the haplotype frequency. Nucleotide differences between haplotypes are indicated on the branches of the network. Circles are color-coded according to population. The most recent common ancestor (MRCA) is also shown (black circle). The relative position of mutations along a branch is arbitrary.

Table 1. Nucleotide Diversity and Neutrality Tests for *KNG1* Regions.

Population	N^a	S^b	Π ($\times 10^{-4}$)		θ_W ($\times 10^{-4}$)		Tajima's *D*		Fu and Li's D^*		Fu and Li's F^*	
			Value	Rank[c]	Value	Rank[c]	Value (*P*)[d]	Rank[c]	Value (*P*)[d]	Rank[c]	Value (*P*)[d]	Rank[c]
KNG1-r1												
AA	48	73	37.49	≥0.99	22.29	0.99	2.41 (≤0.001)	≥0.99	0.83 (0.091)	0.93	1.68 (≤0.001)	≥0.99
YRI	44	68	36.84	≥0.99	21.18	0.98	2.63 (≤0.001)	≥0.99	1.5 (0.005)	≥0.99	2.27 (0.0045)	≥0.99
EU	46	62	30.72	0.99	19.12	0.98	2.14 (0.010)	0.97	1.44 (0.048)	0.96	2.02 (0.008)	0.98
AS	40	62	33.01	≥0.99	19.75	0.99	2.42 (0.010)	0.99	1.17 (0.066)	0.94	1.91 (0.010)	0.99
KNG1-r2												
AA	48	55	13.41	0.89	12.52	0.81	—	—	—	—	—	—
EU	46	43	12.55	0.90	9.88	0.89	—	—	—	—	—	—
KNG1-r3												
AA	48	50	11.87	0.85	12.24	0.81	—	—	—	—	—	—
EU	46	36	11.28	0.87	8.90	0.85	—	—	—	—	—	—

NOTE.—Significant percentile rank and *P* values are underlined.
[a]Sample size (chromosomes).
[b]Number of segregating sites.
[c]Percentile rank relative to a distribution of 238 5-kb windows from NIEHS genes.
[d]*P* value obtained by coalescent simulations.

Table 2. MLHKA Test for *KNG1-r1*.

Population	MLHKA	
	k^a	*P*
AA	5.70	4.56×10^{-6}
YRI	5.18	9.84×10^{-6}
EU	6.57	2.88×10^{-7}
AS	7.19	1.19×10^{-7}

[a]Selection parameter ($k > 1$ indicates an excess of polymorphisms compared with divergence; $k < 1$ indicates the opposite situation).

chr3:187921248–187923253) with relatively high LD in all analyzed populations (supplementary fig. S2, Supplementary Material online).

As it is evident from both the GENETREE analysis (Griffiths and Tavare 1995) (fig. 4*b*), the haplotype genealogy is split into two major clades (haplogroups 1 and 2) separated by long branches; both clades are further split into relatively deep subclades each containing common haplotypes. The time to the most recent common ancestor for the *KNG1* haplotype phylogeny was obtained using GENETREE and amounted to 5.29 My (standard deviation: 1.24 My), assuming that the human–chimpanzee divergence was 6 Ma. Such deep genealogies are highly unlikely under neutrality as estimates for neutrally evolving autosomal loci range between 0.8 and 1.5 My (Tishkoff and Verrelli 2003; Garrigan and Hammer 2006).

Overall, these analyses indicate that *KNG1*-r1 has been a target of long-standing multiallelic balancing selection in all populations.

To identify the possible selection targets, we searched for putative functional variants within the analyzed region and included them in a haplotype network built on the basis of variants in the LD region (fig. 4*c*). Analysis of intermediate frequency single-nucleotide polymorphisms (SNPs) indicated the presence of one single amino acid-replacing polymorphism in exon 4 (rs1656922, Met178Thr). This position does not overlap with any of the sites targeted by diversifying selection in mammals, and the derived 178Met allele defines major haplotypes within both haplogroups (fig. 4*c*); this observation and the fact that the SNP involves a CpG dinucleotide suggest that it is a recurrent mutation, making it difficult to infer its evolutionary history. Data from the ENCODE project (ENCODE Project Consortium et al. 2012) obtained in HepG2 cells (a hepatocellular carcinoma cell line, as liver is the major expression site of *KNG1*) indicated that the balancing selection region carries three DNase I hypersensitive sites, as well as binding sites for different transcriptional regulators (fig. 4*a*). These include CTCF and two components of the cohesin complex (Rad21 and SMC3); these latter have been shown to be recruited to a subset of DNase I hypersensitive sites by CTCF where they function as transcriptional insulators (Parelho et al. 2008). Two signals for CEBPB (also known as liver activator protein) are observed in the region, as well. Two polymorphic variants (rs2689197 and rs2651640) fall within the intronic CEBPB binding site overlapping one DNase I hypersensitive site (fig. 4*a*); their inclusion in the network indicated that they separate the two major haplotype clades (fig. 4*c*). Finally, rs1648714 is located within the CTCF/SMC3/Rad21 binding sites (fig. 4*a*) and defines a major haplotype within haplogroup 2 (fig. 4*c*).

Population Genetic Analysis of *KNG1* in Chimpanzees

Given the deep coalescence time of the *KNG1* haplotype phylogeny, we analyzed the evolutionary pattern of the gene in chimpanzees. Specifically, we resequenced a 5-kb region (magenta bar in fig.4*a*) in nine unrelated *Pan troglodytes*; the aim was to analyze nucleotide diversity and evaluate the presence of trans-specific polymorphisms (i.e., variants shared between humans and chimpanzees). The total number of segregating sites was 10, and no trans-specific variant was observed. To assess whether the *KNG1* region shows unusual levels of nucleotide variability, we compared θ_W and π to the distribution of these same parameters calculated over 16 genomic regions resequenced in the same individuals (see Materials and Methods); results indicated that nucleotide variability in *KNG1* is not exceptional and,

therefore, that this region is likely to be neutrally evolving in chimpanzees (supplementary fig. S3, Supplementary Material online).

Discussion

Evolutionary studies rely on the signatures left by natural selection to describe regions or sites that evolved adaptively and, as such, entail functional significance and represent determinants of phenotypic variation. Evolutionary analysis along the mammalian phylogeny indicated no consistent evidence of adaptive evolution for *F11*, *F12*, and *KLKB1*. Conversely, strong signatures of diversifying positive selection were detected for *KNG1*. In this analysis, we aimed at minimizing false-positive results by requiring all neutral models to be rejected in favor of the positive selection models and by screening for recombination. Positively selected sites were defined by the use of two methods, BEB and MEME; on the one hand, this choice was motivated by the desire to limit the number of false-positive results. On the other hand, we most likely underestimated the number of positively selected sites. Indeed, MEME allows the distribution of *dN*/*dS* to vary from branch to branch at an individual site, resulting in the ability to detect both episodic and pervasive positive selection (Murrell et al. 2012); conversely, sites evolving under episodic selection are likely to be missed by BEB. Thus, the combination of the two methods results in the confident identification of sites evolving under pervasive diversifying selection only (i.e., evidence of episodic selection are lost). Despite this conservative approach, several residues were found to have evolved adaptively in KNG1. Positively selected residues are located in all domains, with the exception of the highly conserved BK sequence. The scattering of selection targets along the entire coding sequence may at least in part be a consequence of the complex and multifaceted interaction between kininogen and infectious agents. In fact, a number of pathogens such as *Candida albicans*, *S. pyogenes*, *Escherichia coli*, and group G streptococci bind kininogen and activate the contact system (Ben Nasr et al. 1996, 1997; Karkowska-Kuleta et al. 2011; Wollein Waldetoft et al. 2012) (fig. 1); the precise binding sites on the kininogen are unknown and may differ across pathogens, although some interactions have been mapped to domains D3, $D5_H$, and D6. Also, different infectious agents (including bacteria, fungi, helminths, and protozoa) encode proteases that can release kinins (from domain D4) and exploit this ability as a virulence strategy (Del Nery et al. 1997; Carvalho et al. 1998; Cordova et al. 2001; Rapala-Kozik et al. 2010 Bagnaresi et al. 2012) (fig. 1). Interestingly, ScpA and SspB from *Sta. aureus* cleave kininogen at each terminal side of the kinin domain releasing Leu-Met-Lys-BK (Imamura et al. 2005). Therefore, the positively selected 377Ser (fig. 2) represents the N-terminal cleavage site of ScpA/SspB, suggesting that sites flanking the BK sequence are evolving to avoid recognition and cleavage by proteases encoded by infectious agents. Finally, contact system activation at the surface of pathogenic bacteria results in the generation of AMPs from domains D3 and $D5_H$ (fig. 1) (Nordahl et al. 2005; Frick et al. 2006). In the case of domain D3, the pattern of generated AMPs depends on the infecting bacteria (Frick et al. 2006). AMPs are considered pivotal components of innate immunity as they represent a first-line response against invading pathogens; in line with this view, other genes encoding AMPs, such as defensins and cathelicidins, were targeted by positive selection (Zelezetsky et al. 2006; Hollox and Armour 2008). Among AMPs deriving from kininogen D3, NAT26 showed the strongest action against a wide range of bacteria (Frick et al. 2006); the NAT26 region carries two positively selected sites, and additional residues targeted by selection occur within the sequence of other less effective AMPs from D3 and within GKH17 from $D5_H$ (fig. 2). It will be extremely interesting to evaluate whether adaptive changes affect the antimicrobial activity of these peptides, as they have been proposed as possible therapeutic molecules against bacterial and fungal infections (Pasupuleti et al. 2009; Schmidtchen et al. 2009; Sonesson et al. 2011). Overall, these observations suggest that multiple kininogen domains establish diverse interactions with infectious agents, and this may result in widespread selection. Nonetheless, no interaction has been described between domains D1/D2 (where several positively selected sites are located) and pathogens, although domain 2 has protease inhibitory activity (Colman and Schmaier 1997) and may inhibit proteases from infectious agents.

An alternative and not mutually exclusive possibility is that the selective pressure exerted by infectious agents is indirect and derives from the modulation of the inflammatory response. As an example, kinins can be released from high-molecular-weight kininogen by cellular proteases distinct from PK. Neurtrophil-derived elastase and proteinase 3 (fig. 1) produce kinins with terminal extensions: Ser-Leu-Met-Lys-BK-Ser-Ser-Arg-Ile (where the N-terminal Ser and the C-terminal Arg and Ile residues are the positively selected sites, fig. 2), BK-Ser-Ser, and Ser-Leu-Met-Lys-BK (Imamura et al. 2002; Kahn et al. 2009). These kinins may be subsequently processed at both termini, have physiological effects similar to those of BK, and are thought to act as important mediators at sites of inflammation (Imamura et al. 2002; Kahn et al. 2009). Whether the positively selected sites alter the physiologic function of these kinins or the efficiency of their production by neutrophil enzymes remains to be evaluated.

The contact system has been a target of intense investigation for years, as it represents a central link among the coagulation cascade, inflammation, and innate immunity. High-molecular-weight kininogen possibly represents the very molecule at the crossroad of these pathways, as it participates in contact system activation and its cleavage originates both kinins and AMPs; in line with this observation, results herein indicate a continuum in selective pressure acting on different timescales and targeting the *KNG1* gene but not other contact system components. Indeed, analysis of nucleotide diversity in human populations revealed no usual feature for *F11*, *F12*, and *KLKB1*; conversely, we found strong evidences that genetic diversity at *KNG1* has been maintained by balancing selection during the evolutionary history of human populations. In line with this finding, the analysis of *KNG1* haplotypes revealed the presence of two clades separated by long branches approximately dating back to the time

when the human and chimpanzee lineages diverged (Glazko and Nei 2003). Altogether these features represent strong molecular signatures of long-term balancing selection, and the split of the major branches in subclades with relatively deep times of the most recent common ancestor (TMRCA) suggest that the selection is multiallelic (i.e., that more than one selection target are located in the region). Despite the deep coalescence time of the *KNG1* haplotype phylogeny, no selection signature was detected in chimpanzees. Nonetheless, this conclusion should be interpreted with caution as the number of resequenced *P. troglodytes* individuals is relatively small and analysis of additional samples might unveil the presence of low-frequency haplotypes, which might still result from selection (e.g., frequency-dependent selection).

Analysis of variants located along the major branches of the haplotype genealogy indicated that long-standing balancing selection may be acting to maintain regulatory variant in *KNG1*, as previously shown for other genes involved in innate immunity (Cagliani et al. 2008, 2010; Ferrer-Admetlla et al. 2008; Hollox and Armour 2008). Interestingly, two polymorphisms (fig. 4*a*) fall within a CEBPB binding site. CEBPB is a major regulator of the expression of acute phase proteins and is induced upon stimulation with lipopolysaccharide and interleukin-6 (Akira et al. 1990), suggesting that it may link *KNG1* expression to inflammatory and infectious states. Increase in the plasma concentration of acute phase proteins is a hallmark of sepsis, and the contact system was shown to play a central role in this potentially fatal condition (Nickel and Renne 2012). It remains to be evaluated whether variants in the balancing selection region modulate *KNG1* expression during infection, and, possibly, during the systemic inflammatory response syndrome (SIRS) and sepsis, this latter still considered a major cause of death in infants (Watson and Carcillo 2005) and a driver of molecular evolution in humans (Wang et al. 2006; Xue et al. 2006).

In summary, data herein indicate that, unique among contact system genes, *KNG1* has been a target of long-lasting and strong selective pressures. These data reinforce the idea that kininogen plays a central role in the modulation of immune response and is a target of an extremely diverse array of pathogen species.

Materials and Methods

Evolutionary Analysis in Mammals

All mammalian sequences for *KNG1*, *F11*, *F12*, and *KLKB1* were retrieved from the Ensembl website (http://www.ensembl.org/index.html, last accessed November 3, 2012). For *KNG1*, the following species were aligned: *Homo sapiens* (human), *P. troglodytes* (chimpanzee), *Gorilla gorilla* (gorilla), *Nomascus leucogenys* (gibbon), *Macaca mulatta* (macaque), *Callithrix jacchus* (marmoset), *Microcebus murinus* (mouse lemur), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Cavia porcellus* (guinea pig), *Oryctolagus cuniculus* (rabbit), *Bos taurus* (cow), *Tursiops truncatus* (dolphin), *Canis lupus familiaris* (dog), *Ailuropoda melanoleuca* (panda), *Sus scrofa* (pig), *Equus caballus* (horse), *Myotis Lucifugus* (microbat), *Pteropus vampyrus* (megabat), *Erinaceus europaeus* (hedgehog), *Loxodonta africana* (elephant), and *Sorex araneus* (shrew), *Dasypus novemcinctus* (armadillo). For *KLKB1*, *F11*, and *F12*, species list is reported in supplementary table S3, Supplementary Material online. Sequences for all species (excluding human) derive from genome sequencing program predictions. The available *KNG1* transcript sequences for mouse (NM_023125.3 and NM_001102411.1), rat (NM_012696.2), and cow (NM_001113277.1 and NM_175774.3) were checked against the predicted sequences of the respective species and showed 100% identity. The predicted sequences for mouse lemur, guinea pig, and pig showed sequencing gaps in the $KNG1_{D5H\text{-}D6}$ region (fig. 2); missing data do not affect inference of positive selection, and omission of these species from PAML analysis yielded comparable results to those obtained with their inclusion (not shown). DNA alignments were performed using the RevTrans 2.0 utility (Wernersson and Pedersen 2003), which uses the protein sequence alignment as a scaffold for constructing the corresponding DNA multiple alignment. This latter was checked and edited by hand to remove alignment uncertainties. Average dN/dS and its CIs were calculated using SLAC (Kosakovsky Pond and Frost 2005) taking GARD results into account (i.e., using GARD inferred trees). SLAC was run through the DataMonkey server (Delport et al. 2010) (http://www.datamonkey.org, last accessed November 3, 2012). For PAML analyses (Yang 2007), we used trees generated by maximum likelihood using the program PhyML (Guindon et al. 2009).

To detect selection, NSsite models that allow (M8 and M2a) or disallow (M1a and M8a) sites to evolve with $dN/dS > 1$ were fitted to the data using the F3x4 codon frequency model. Whenever maximum-likelihood trees showed differences (always minor) from the accepted mammalian phylogeny, analyses were repeated using the accepted tree, and the same results were obtained in all cases. Sites under selection with the M8 model were identified using BEB analysis using a significance cutoff of 0.90 (Anisimova et al. 2002; Yang et al. 2005). GARD and MEME analyses were performed through the DataMonkey server (Delport et al. 2010) (http://www.datamonkey.org, last accessed November 3, 2012).

Pairwise dN/dS values for human–macaque, human–mouse, and human–dog orthologs were derived from the Ensembl BioMart database (http://www.ensembl.org/biomart/, last accessed November 3, 2012); only 1-to-1 orthologs were included, and genes with $dS > 1$ were discarded. The number of ortholog pairs were 15,435 (human–macaque), 13,411 (human–mouse), and 12,987 (human–dog).

HapMap Samples and Sequencing

Human genomic DNA from YRI and AS individuals was obtained from the Coriell Institute for Medical Research. The analyzed region (NCBI36/hg18: chr3:187920669–187925896) was polymerase chain reaction (PCR) amplified and directly sequenced (primer sequences are available upon request). PCR products were treated with ExoSAP-IT (USB Corporation Cleveland, OH), directly sequenced on both

strands with a Big Dye Terminator sequencing Kit (v3.1, Applied Biosystems) and run on an Applied Biosystems ABI 3130 XL Genetic Analyzer (Applied Biosystems). Sequences were assembled using AutoAssembler version 1.4.0 (Applied Biosystems), inspected manually by two distinct operators. Genotype data for AA and EU were retrieved from the SeattleSNPs website (http://pga.mbt.washington.edu, last accessed November 3, 2012). Information on SNP positions and haplotypes for all subjects is available in supplementary table S4, Supplementary Material online. The genomic DNA of nine *P. troglodytes* was obtained from the Gene Bank of Primates, Primate Genetics, Germany (http://dpz.eu/index.php, last accessed November 3, 2012). These samples have been shown to belong to the *P. troglodytes verus* subspecies (Cagliani et al. 2012).

Population Genetic Analyses

Tajima's *D* (Tajima 1989), Fu and Li's *D** and *F** (Fu and Li 1993) statistics, and diversity parameters θ_W (Watterson 1975) and π (Nei and Li 1979) were calculated using "libsequence" (Thornton 2003). Calibrated coalescent simulations were performed using the "cosi" package (Schaffner et al. 2005) and its best-fit demographic parameters for AA, YRI, EU, and AS populations with 10,000 iterations. Coalescent simulations were conditioned on mutation rate, and recombination rate was derived from the University of California–Santa Cruz tables (http://genome.ucsc.edu/, snpRecombRateHamap table, last accessed November 3, 2012).

The maximum-likelihood-ratio Hudson–Kreitman–Aguadé (HKA) test was performed using the MLHKA software (Wright and Charlesworth 2004), as previously proposed (Fumagalli et al. 2009). For human populations, 16 reference loci were randomly selected among NIEHS loci shorter than 20 kb that have been resequenced in AA, YRI, EU, and AS; the only criterion was that Tajima's *D* did not suggest the action of natural selection (i.e., Tajima's *D* is higher than the 5th and lower than the 95th percentiles in the distribution of NIEHS genes).

Genotype data from 201 genes resequenced in AA and EU (SeattleSNPs Panel 1) were derived from the SeattleSNPs Variation Discovery Resource Program (http://pga.gs.washington.edu/, last accessed November 3, 2012).

Genotype data for 5-kb regions from 238 resequenced human genes were derived from the NIEHS SNPs Program web site (http://egp.gs.washington.edu, last accessed November 3, 2012). In particular, we selected genes that had been resequenced in populations of defined ethnicity including AA, YRI, EU, and AS (NIEHS Panel 2).

For the chimpanzee, the empirical comparison of θ_W and π parameters was performed using 16 resequenced regions in the same individuals as reference loci (Cagliani et al. 2012).

Haplotype Analysis and TMRCA Calculation

Haplotypes were inferred using PHASE (version 2.1) (Stephens et al. 2001; Stephens and Scheet 2005). The median-joining network to infer haplotype genealogy was constructed using NETWORK 4.5 (Bandelt et al. 1999). Estimate of the TMRCA derived from application of a maximum-likelihood coalescent method implemented in GENETREE (Griffiths and Tavare 1994, 1995). The method assumes an infinite-site model without recombination; therefore, GENETREE identifies sites that violate these assumptions: In the case of *KNG1*, nine variants (rs5029985, rs5029986, rs12635879, rs5029989, rs5029991, rs5029992, rs1656910, rs1829886, and rs1648714) were identified and we eliminated them from the data. Again, the mutation rate μ was obtained on the basis of the divergence between human and chimpanzee and under the assumption both that the species separation occurred 6 Ma (Glazko and Nei 2003) and of a generation time of 25 years. A constant population size model was used for simulations, and the migration matrix was derived from previous estimated migration rates (Schaffner et al. 2005). Using this μ and θ maximum likelihood (θ_{ML}), we estimated the effective population size parameter (N_e), which resulted equal to 30,545. With these assumptions, the coalescence time, scaled in $2N_e$ units, was converted into years. For the coalescence process, 10^6 simulations were performed.

LD analyzes were performed using Haploview (Barrett et al. 2005), and haplotypes blocks were identified through an implemented method (Gabriel et al. 2002).

Data for LD analysis were derived from resequencing data.

Supplementary Material

Supplementary figures S1–S3 and tables S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

References

Akira S, Isshiki H, Sugita T, Tanabe O, Kinoshita S, Nishio Y, Nakajima T, Hirano T, Kishimoto T. 1990. A nuclear factor for IL-6 expression (NF-IL6) is a member of a C/EBP family. *EMBO J.* 9:1897–1906.

Anisimova M, Bielawski JP, Yang Z. 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Mol Biol Evol.* 19:950–958.

Bagnaresi P, Barros NM, Assis DM, et al. (11 co-authors). 2012. Intracellular proteolysis of kininogen by malaria parasites promotes release of active kinins. *Malar J.* 11:156.

Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol.* 16:37–48.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21:263–265.

Ben Nasr A, Herwald H, Sjobring U, Renne T, Muller-Esterl W, Bjorck L. 1997. Absorption of kininogen from human plasma by *Streptococcus pyogenes* is followed by the release of bradykinin. *Biochem J.* 326(Pt 3):657–660.

Ben Nasr A, Olsen A, Sjobring U, Muller-Esterl W, Bjorck L. 1996. Assembly of human contact phase proteins and release of bradykinin at the surface of curli-expressing *Escherichia coli*. *Mol Microbiol.* 20:927–935.

Cagliani R, Fumagalli M, Biasin M, Piacentini L, Riva S, Pozzoli U, Bonaglia MC, Bresolin N, Clerici M, Sironi M. 2010. Long-term balancing selection maintains trans-specific polymorphisms in the human *TRIM5* gene. *Hum Genet.* 128:577–588.

Cagliani R, Fumagalli M, Riva S, Pozzoli U, Comi GP, Menozzi G, Bresolin N, Sironi M. 2008. The signature of long-standing balancing selection at the human defensin beta-1 promoter. *Genome Biol.* 9:R143.

Cagliani R, Guerini FR, Fumagalli M, et al. (22 co-authors). 2012. A trans-specific polymorphism in ZC3HAV1 is maintained by long-standing

balancing selection and may confer susceptibility to multiple sclerosis. *Mol Biol Evol.* 29:1599–1613.
Campbell DJ. 2001. The kallikrein-kinin system in humans. *Clin Exp Pharmacol Physiol.* 28:1060–1065.
Carvalho WS, Lopes CT, Juliano L, Coelho PM, Cunha-Melo JR, Beraldo WT, Pesquero JL. 1998. Purification and partial characterization of kininogenase activity from *Schistosoma mansoni* adult worms. *Parasitology* 117(Pt 4):311–319.
Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet.* 2:e64.
Colman RW, Schmaier AH. 1997. Contact system: a vascular biology modulator with anticoagulant, profibrinolytic, antiadhesive, and proinflammatory attributes. *Blood* 90:3819–3843.
Cordova M, Jara J, Del Nery E, Hirata IY, Araujo MS, Carmona AK, Juliano MA, Juliano L. 2001. Characterization of two cysteine proteinases secreted by *Fasciola hepatica* and demonstration of their kininogenase activity. *Mol Biochem Parasitol.* 116:109–115.
Del Nery E, Juliano MA, Lima AP, Scharfstein J, Juliano L. 1997. Kininogenase activity by the major cysteinyl proteinase (cruzipain) from *Trypanosoma cruzi*. *J Biol Chem.* 272:25713–25718.
Delport W, Poon AF, Frost SD, Kosakovsky Pond SL. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* 26:2455–2457.
ENCODE Project Consortium, Bernstein BE, Birney E, et al. (603 co-authors). 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.
Ferrer-Admetlla A, Bosch E, Sikora M, et al. (11 co-authors). 2008. Balancing selection is the main force shaping the evolution of innate immunity genes. *J Immunol.* 181:1315–1322.
Frick IM, Akesson P, Herwald H, Morgelin M, Malmsten M, Nagler DK, Bjorck L. 2006. The contact system—a novel branch of innate immunity generating antibacterial peptides. *EMBO J.* 25:5569–5578.
Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* 133:693–709.
Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi GP, Menozzi G, Bresolin N, Sironi M. 2009. Widespread balancing selection and pathogen-driven selection at blood group antigen genes. *Genome Res.* 19:199–212.
Gabriel SB, Schaffner SF, Nguyen H, et al. (18 co-authors). 2002. The structure of haplotype blocks in the human genome. *Science* 296:2225–2229.
Garrigan D, Hammer MF. 2006. Reconstructing human origins in the genomic era. *Nat Rev Genet.* 7:669–680.
Glazko GV, Nei M. 2003. Estimation of divergence times for major lineages of primate species. *Mol Biol Evol.* 20:424–434.
Griffiths RC, Tavare S. 1994. Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci.* 344:403–410.
Griffiths RC, Tavare S. 1995. Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Math Biosci.* 127:77–98.
Guindon S, Delsuc F, Dufayard JF, Gascuel O. 2009. Estimating maximum likelihood phylogenies with PhyML. *Methods Mol Biol.* 537:113–137.
Herwald H, Collin M, Muller-Esterl W, Bjorck L. 1996. Streptococcal cysteine proteinase releases kinins: a virulence mechanism. *J Exp Med.* 184:665–673.
Hollox EJ, Armour JA. 2008. Directional and balancing selection in human beta-defensins. *BMC Evol Biol.* 8:113.
Imamura T, Potempa J, Pike RN, Travis J. 1995. Dependence of vascular permeability enhancement on cysteine proteinases in vesicles of *Porphyromonas gingivalis*. *Infect Immun.* 63:1999–2003.
Imamura T, Tanase S, Hayashi I, Potempa J, Kozik A, Travis J. 2002. Release of a new vascular permeability enhancing peptide from kininogens by human neutrophil elastase. *Biochem Biophys Res Commun.* 294:423–428.
Imamura T, Tanase S, Szmyd G, Kozik A, Travis J, Potempa J. 2005. Induction of vascular leakage through release of bradykinin and a novel kinin by cysteine proteinases from *Staphylococcus aureus*. *J Exp Med.* 201:1669–1676.
Isordia-Salas I, Pixley RA, Parekh H, Kunapuli SP, Li F, Stadnicki A, Lin Y, Sartor RB, Colman RW. 2003. The mutation Ser511Asn leads to N-glycosylation and increases the cleavage of high molecular weight kininogen in rats genetically susceptible to inflammation. *Blood* 102:2835–2842.
Jiang Y, Doolittle RF. 2003. The evolution of vertebrate blood coagulation as viewed from a comparison of puffer fish and sea squirt genomes. *Proc Natl Acad Sci U S A.* 100:7527–7532.
Kahn R, Hellmark T, Leeb-Lundberg LM, et al. (12 co-authors). 2009. Neutrophil-derived proteinase 3 induces kallikrein-independent release of a novel vasoactive kinin. *J Immunol.* 182:7906–7915.
Karkowska-Kuleta J, Kedracka-Krok S, Rapala-Kozik M, Kamysz W, Bielinska S, Karafova A, Kozik A. 2011. Molecular determinants of the interaction between human high molecular weight kininogen and *Candida albicans* cell wall: identification of kininogen-binding proteins on fungal cell wall and mapping the cell wall-binding regions on kininogen molecule. *Peptides* 32:2488–2496.
Kosakovsky Pond SL, Frost SD. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol.* 22:1208–1222.
Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. 2006. Automated phylogenetic detection of recombination using a genetic algorithm. *Mol Biol Evol.* 23:1891–1901.
Muller-Esterl W. 1989. Kininogens, kinins and kinships. *Thromb Haemost.* 61:2–6.
Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. 2012. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* 8:e1002764.
Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A.* 76:5269–5273.
Nickel KF, Renne T. 2012. Crosstalk of the plasma contact system with bacteria. *Thromb Res.* 130(1 Suppl): S78–S83.
Nordahl EA, Rydengard V, Morgelin M, Schmidtchen A. 2005. Domain 5 of high molecular weight kininogen is antibacterial. *J Biol Chem.* 280:34832–34839.
Paegelow I, Trzeczak S, Bockmann S, Vietinghoff G. 2002. Migratory responses of polymorphonuclear leukocytes to kinin peptides. *Pharmacology* 66:153–161.
Parelho V, Hadjur S, Spivakov M, et al. (16 co-authors). 2008. Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* 132:422–433.
Pasupuleti M, Chalupka A, Morgelin M, Schmidtchen A, Malmsten M. 2009. Tryptophan end-tagging of antimicrobial peptides for increased potency against *Pseudomonas aeruginosa*. *Biochim Biophys Acta.* 1790:800–808.
Rapala-Kozik M, Karkowska-Kuleta J, Ryzanowska A, Golda A, Barbasz A, Faussner A, Kozik A. 2010. Degradation of human kininogens with the release of kinin peptides by extracellular proteinases of *Candida* spp. *Biol Chem.* 391:823–830.
Sato E, Koyama S, Nomura H, Kubo K, Sekiguchi M. 1996. Bradykinin stimulates alveolar macrophages to release neutrophil, monocyte, and eosinophil chemotactic activity. *J Immunol.* 157:3122–3129.
Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 15:1576–1583.
Scharfstein J, Schmitz V, Morandi V, Capella MM, Lima AP, Morrot A, Juliano L, Muller-Esterl W. 2000. Host cell invasion by *Trypanosoma cruzi* is potentiated by activation of bradykinin B(2) receptors. *J Exp Med.* 192:1289–1300.
Scharfstein J, Schmitz V, Svensjo E, Granato A, Monteiro AC. 2007. Kininogens coordinate adaptive immunity through the proteolytic release of bradykinin, an endogenous danger signal driving dendritic cell maturation. *Scand J Immunol.* 66:128–136.
Schmidtchen A, Pasupuleti M, Morgelin M, Davoudi M, Alenfall J, Chalupka A, Malmsten M. 2009. Boosting antimicrobial peptides

by hydrophobic oligopeptide end tags. *J Biol Chem.* 284: 17584–17594.
Sonesson A, Nordahl EA, Malmsten M, Schmidtchen A. 2011. Antifungal activities of peptides derived from domain 5 of high-molecular-weight kininogen. *Int J Pept.* 2011:761037.
Stephens M, Scheet P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet.* 76:449–462.
Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet.* 68:978–989.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.
Thornton K. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* 19:2325–2327.
Tishkoff SA, Verrelli BC. 2003. Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu Rev Genomics Hum Genet.* 4:293–340.
Wang X, Grus WE, Zhang J. 2006. Gene losses during human origins. *PLoS Biol.* 4:e52.
Watson RS, Carcillo JA. 2005. Scope and epidemiology of pediatric sepsis. *Pediatr Crit Care Med.* 6:S3–S5.
Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol.* 7: 256–276.
Wernersson R, Pedersen AG. 2003. RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res.* 31:3537–3539.
Wiuf C, Zhao K, Innan H, Nordborg M. 2004. The probability and chromosomal extent of trans-specific polymorphism. *Genetics* 168: 2363–2372.
Wollein Waldetoft K, Svensson L, Morgelin M, Olin AI, Nitsche-Schmitz DP, Bjorck L, Frick IM. 2012. Streptococcal surface proteins activate the contact system and control its antibacterial activity. *J Biol Chem.* 287:25010–25018.
Wright SI, Charlesworth B. 2004. The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. *Genetics* 168: 1071–1076.
Xue Y, Daly A, Yngvadottir B, Liu M, et al. (14 co-authors). 2006. Spread of an inactive form of caspase-12 in humans is due to recent positive selection. *Am J Hum Genet.* 78:659–670.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.
Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol.* 22: 1107–1118.
Zelezetsky I, Pontillo A, Puzzi L, Antcheva N, Segat L, Pacor S, Crovella S, Tossi A. 2006. Evolution of the primate cathelicidin. Correlation between structural variations and antimicrobial activity. *J Biol Chem.* 281:19861–19871.

3.4 Ancient and recent selective pressures shaped genetic diversity at AIM2-like nucleic acid sensors.

GBE

Genome Biology and Evolution

SMBE

Ancient and Recent Selective Pressures Shaped Genetic Diversity at AIM2-Like Nucleic Acid Sensors

Rachele Cagliani[1], Diego Forni[1], Mara Biasin[2], Manuel Comabella[3], Franca R. Guerini[4], Stefania Riva[1], Uberto Pozzoli[1], Cristina Agliardi[4], Domenico Caputo[4], Sunny Malhotra[3], Xavier Montalban[3], Nereo Bresolin[1,5], Mario Clerici[4,6], and Manuela Sironi[1,*]

[1]Bioinformatics Laboratory, Scientific Institute IRCCS E. Medea, Bosisio Parini (LC), Italy

[2]Department of Biomedical and Clinical Sciences, University of Milan, Italy

[3]Department of Neurology-Neuroimmunology, Centre d'Esclerosi Múltiple de Catalunya, Cemcat, Hospital Universitari Vall d'Hebron (HUVH), Barcelona, Spain

[4]Laboratory of Molecular Medicine, Don C. Gnocchi Foundation ONLUS, IRCCS, Milan, Italy

[5]Dino Ferrari Centre, Department of Physiopathology and Transplantation, University of Milan, Fondazione Ca' Granda IRCCS Ospedale Maggiore Policlinico, Milan, Italy

[6]Chair of Immunology, Department of Physiopathology and Transplantation, University of Milan, Italy

*Corresponding author: E-mail: manuela.sironi@bp.lnf.it.

Accepted: March 21, 2014

Data deposition: The *IFI16* coding sequences for *Macaca fascicularis* and *Chlorocebus aethiops* have been deposited at GenBank under the accessions KF154419 and KF154420.

Abstract

AIM2-like receptors (ALRs) are a family of nucleic acid sensors essential for innate immune responses against viruses and bacteria. We performed an evolutionary analysis of ALR genes (*MNDA*, *PYHIN1*, *IFI16*, and *AIM2*) by analyzing inter- and intraspecies diversity. Maximum-likelihood analyses indicated that *IFI16* and *AIM2* evolved adaptively in primates, with branch-specific selection at the catarrhini lineage for *IFI16*. Application of a population genetics–phylogenetics approach also allowed identification of positive selection events in the human lineage. Positive selection in primates targeted sites located at the DNA-binding interface in both IFI16 and AIM2. In IFI16, several sites positively selected in primates and in the human lineage were located in the PYD domain, which is involved in protein–protein interaction and is bound by a human cytomegalovirus immune evasion protein. Finally, positive selection was found to target nuclear localization signals in IFI16 and the spacer region separating the two HIN domains. Population genetic analysis in humans revealed that an *IFI16* genic region has been a target of long-standing balancing selection, possibly acting on two nonsynonymous polymorphisms located in the spacer region. Data herein indicate that ALRs have been repeatedly targeted by natural selection. The balancing selection region in *IFI16* carries a variant with opposite risk effect for distinct autoimmune diseases, suggesting antagonistic pleiotropy. We propose that the underlying scenario is the result of an ancestral and still ongoing host–pathogen arms race and that the maintenance of susceptibility alleles for autoimmune diseases at *IFI16* represents an evolutionary trade-off.

Key words: AIM2-like receptors, positive selection, long-standing balancing selection, IFI16.

Introduction

Mammalian nucleic acid-sensing receptors play essential roles in the recognition of infectious agents and in triggering innate and adaptive immune responses. Different classes of nucleic acid-sensing molecules have been identified; these molecules are classified on the basis of cellular localization, target specificity, and downstream signaling pathway. Among them, toll-like receptors (TLRs) are the best characterized class, and at least four members (TLR3, TLR7, TLR8, and TLR9), located at the endosomal membrane, are specialized in viral sensing (Desmet and Ishii 2012). TLRs signal through MyD88 or TRIF to induce the release either of inflammatory

cytokines or of type I interferons (IFNs). The cytoplasmic detection of viral RNA is also mediated by cytosolic RIG-I-like receptors (RIG-I, IFIH1, and DHX58), which elicit type I IFN responses through the mitochondrial antiviral signaling (MAVS) adaptor (Desmet and Ishii 2012). NOD2, a member of the NOD-like receptor (NLR) family, also impinges on MAVS upon sensing single-stranded RNA (ssRNA), whereas NLRP3 is activated by ssRNA or dsRNA resulting in the formation of the inflammasome complex, which is mediated by PYCARD (also known as ASC) (Desmet and Ishii 2012). Finally, the IFN-inducible HIN-200 gene family, also called PYHIN gene family, comprises a class of homologous viral sensor proteins characterized by the presence of an N-terminal pyrin-domain and a 200-amino acid signature motif (HIN-200 domain) (Veeranki and Choubey 2012).

In humans, four members of this family have been identified and are encoded by a cluster of genes (*MNDA*, *PYHIN1*, *IFI16*, and *AIM2*) located on chromosome 1. These proteins share a pyrin motif involved in protein–protein interactions, as well as one (AIM2, MNDA, and PYHIN1) or two (IFI16) HIN-200 domains that mediate binding to double-stranded DNA (dsDNA) (fig. 1) (Schattgen and Fitzgerald 2011). The best studied PYHIN family members are AIM2 and IFI16. The former is a sensor of cytosolic DNA, which triggers the inflammosome pathway through PYCARD resulting in caspase-1-mediated cleavage of IL-1β (Schattgen and Fitzgerald 2011). Conversely, IFI16 has a mainly nuclear activity (in analogy to MNDA and PYHIN1), although it can also sense dsDNA in the cytoplasm, as its nuclear-cytoplasmic shuttling is regulated by a multipartite nuclear localization signal (Li et al. 2012). IFI16 signals through STING-dependent pathways. PYHIN proteins, according to their function as innate DNA sensors, are also termed AIM2-like receptors (ALRs) (Unterholzner et al. 2010).

A recent analysis of ALR genes in mammals indicated that the cluster is extremely dynamic: Distinct species carry diverse sets of functional genes, suggesting that strong selective pressures have been acting on these loci (Brunette et al. 2012). Indeed, evolutionary analysis of other genes involved in nucleic acid sensing or in the downstream signaling pathways identified signatures of natural selection. Thus, *MAVS* evolved adaptively in primates, the underlying pressure being accounted for by hepaciviruses (Patel et al. 2012). Likewise, analysis of RIG-I-like receptors in human populations revealed signatures of local adaptation at the *IFIH1* and *DHX58* genes (Fumagalli et al. 2010; Vasseur et al. 2011; Quintana-Murci and Clark 2013). These observations are in line with viruses, and pathogens, in general, being a major determinant of molecular evolution in mammals and human populations (Kosiol et al. 2008; Fumagalli et al. 2011). Herein, we performed an evolutionary study of the ALR cluster by analyzing both inter- and intraspecies diversity.

Materials and Methods

Evolutionary Analysis in Mammals

Primate sequences for *MNDA*, *PYHIN1*, *IFI16*, and *AIM2* were retrieved from the Ensembl website (http://www.ensembl.org/index.html, last accessed January 30, 2014) and National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov/, last accessed January 30, 2014). *IFI16* coding sequencing information for *Macaca fascicularis* and *Chlorocebus aethiops* were obtained by real time polymerase chain reaction (PCR) amplification of RNA derived from CYNOM-K1 and COS1 cells. Primer sequences are available in supplementary table S1, Supplementary Material online. The species list for all genes is reported in supplementary table S2, Supplementary Material online.

DNA alignments were performed using the RevTrans 2.0 utility (Wernersson and Pedersen 2003), which uses the protein sequence alignment as a scaffold for constructing the corresponding DNA multiple alignment. This latter was checked and edited by hand to remove alignment uncertainties. For PAML analyses (Yang 2007), we used trees generated by maximum likelihood (ML) using the program PhyML (Guindon et al. 2009).

To detect selection, NSsite models that allow (M8, M2a) or disallow (M1a and M7) sites to evolve with $dN/dS > 1$ were fitted to the data using the F3x4 and F61 codon frequency model. Whenever ML trees showed differences (always minor) from the accepted primate phylogeny, analyses were repeated using the accepted tree, and the same results were obtained in all cases. Sites under selection with the M8 model were identified using Bayes empirical Bayes (BEB) analysis using a significance cutoff of 0.90 (Anisimova et al. 2002; Yang et al. 2005).

To explore possible variations in selective pressure among different lineages, we applied the free-ratio models implemented in the PAML package: The M0 model assumes all branches to have the same ω, whereas M1 allows each branch to have its own ω (Yang and Nielsen 1998). The models are compared through likelihood-ratio tests (degree of freedom = total number of branches − 1). To identify specific branches with a proportion of sites evolving with $\omega > 1$, we used branch-site-random effects likelihood (BS-REL; Kosakovsky Pond et al. 2011). Branches identified using this approach were cross-validated with the branch-site likelihood ratio tests (LRTs) from PAML (the so-called modified model A and model MA1, "test 2") (Zhang et al. 2005). The advantage of this method is that it also implements a BEB analysis analogous to that described earlier to calculate the posterior probabilities that each site belongs to the site class of positive selection on the foreground lineages. Thus, BEB allows identification of specific sites that evolve under positive selection on specific lineages, although it has limited statistical power (Zhang et al. 2005).

FIG. 1.—ALR gene domain representation and analysis of positively selected sites. (*A*) Domain structure of the four ALR genes. Positively selected sites in whole phylogeny (identified through both BEB and MEME) are shown in red; residues subject to positive selection in the human (green) and in the catarrhini (black: BEB sites; magenta: BEB and MEME sites) lineages are also indicated. (*B*) Three-dimensional (3D) structure of the AIM2 HIN:DNA complex (pdb ID 3RN2). Positively selected sites are shown in red, and yellow indicates key residues at the HIN:DNA interface. Positions refer to the human sequence. (*C*) Alignment of the IFI16 PYD domain. Positively selected sites in whole phylogeny are shown in red, whereas human-specific positively selected sites are marked by a green asterisks. A lysine residue conserved among PYD domains from different proteins is shown in violet. The extension of the six α-helices, inferred from structure analysis of human PYDs, is shown. (*D*) 3D structure of the IFI16 HIN:DNA complex (pdb ID 3RNU). The 615T positively selected site (red) is located at the DNA-binding interface formed by the 4 IFI16 monomers.

Genetic Algorithm Recombination Detection (GARD; Kosakovsky Pond et al. 2006), mixed effects model of evolution (MEME; Murrell et al. 2012), single likelihood ancestor counting (SLAC; Kosakovsky Pond and Frost 2005), and BS-REL analyses were performed through the DataMonkey server (Delport et al. 2010) (http://www.datamonkey.org, last accessed January 30, 2014).

Population Genetics–Phylogenetics Analysis

Data from the Pilot 1 phase of the 1000 Genomes Project were retrieved from the dedicated website (1000 Genomes Project Consortium et al. 2010). Low-coverage single-nucleotide polymorphism genotypes were organized in a MySQL database. A set of programs was developed to retrieve genotypes from the database and to analyze them according to selected regions/populations. These programs were developed in C++ using the GeCo++ (Cereda et al. 2011), the libsequence (Thornton 2003), and the mysqlpp libraries. Coding sequence information was obtained for the four ALR genes. To analyze the distribution of fitness effects (DFEs) for ALR genes, we used gammaMap (Wilson et al. 2011). We assumed θ (neutral mutation rate per site), k (transitions/transversions ratio), and T (branch length) to vary among genes following log-normal distributions. For each gene, we set the neutral frequencies of non-STOP codons (1/61) and the probability that adjacent codons share the same selection coefficient ($P = 0.02$). For selection coefficients, we considered a uniform Dirichlet distribution with the same prior weight for each selection class. For each gene, we run 10,000 iterations with a thinning interval of 10 iterations.

HapMap Samples and Sequencing

Human genomic DNA from 60 HapMap subjects (20 individuals for Yoruba [YRI], 20 Europeans [CEU], and 20 East Asian [AS] subjects) was obtained from the Coriell Institute for Medical Research (http://www.coriell.org/, last accessed January 30, 2014). The analyzed region (NCBI/hg18 chr1:157267539–157272405) was PCR amplified and directly sequenced. PCR products were treated with ExoSAP-IT (USB Corporation, Cleveland, OH), directly sequenced on both strands with a Big Dye Terminator sequencing Kit (v3.1, Life Technologies), and run on an Applied Biosystems ABI 3130 XL Genetic Analyzer (Life Technologies). Sequences were assembled using AutoAssembler version 1.4.0 (Life Technologies) and inspected manually by two distinct operators. Primer sequences are available in supplementary table S1, Supplementary Material online.

Primate Samples

The CYNOM-K1 and COS1 cells, as well as genomic DNA from *Gorilla gorilla* and *Pongo pygmaeus*, were obtained by the European Collection of Cell Cultures. The genomic DNA of eight *Pan troglodytes* was kindly provided by the Gene Bank of Primates, Primate Genetics, Germany (http://dpz.eu/index.php, last accessed January 30, 2014). These samples belong to the *Pan troglodytes verus* subspecies (Cagliani et al. 2012). The IFI16 coding sequences for *M. fascicularis* and *C. aethiops* have been submitted to GenBank (provisional IDs: KF154419 and KF154420).

Segmental Duplication Genotyping

The segmental duplication of exon 7 in *IFI16* was analyzed using a PCR-based method. In particular, PCR amplifications were performed with JumpStart AccuTaq LA DNA Polymerase (Sigma-Aldrich) and two sets of primers: one that amplifies only the duplicated form (F: GTCCTGTGCACCTTGTGTCA; R: CTGATGTATGGTGAGAGAGC) and one that flanks the segmental duplication (F: GTCCATTTCTGTAGCCATAGG; R: TCTGAGTTGTAGGAGAGCACT). The PCR products were electrophoretically separated on agarose gels.

F_{ST} Analysis

Human Genome Diversity Project (HGDP CEPH) panel data derive from a previous work (Li et al. 2008). Atypical or duplicated samples and pairs of close relatives were removed (Rosenberg 2006). Following previous indications (Fumagalli et al. 2009a, 2009b), Bantu individuals (South Africa) were considered as one population. F_{ST} was calculated for all HGDP-CEPH variants among continental groups using the R package HIERFSTAT (Goudet 2005); F_{ST} distributions were calculated after binning single-nucleotide polymorphisms (SNPs) into minor allele frequency (MAF) classes (50 quantile classes based on MAF calculated over the whole panel); outliers are defined as variants with an F_{ST} higher than the 95th percentile in the distribution of SNPs in the same MAF class.

As for the 1000 Genomes Project data, genotype information was obtained for the analyzed genomic region (NCBI/hg18, chr1:157063927–157317926) and for 2,000 randomly selected RefSeq genes. F_{ST} sliding window analysis was performed on overlapping 20 SNP windows moving with a step of three SNPs. The numbers of windows used to obtain a reference distribution (i.e., deriving from the 2,000 randomly selected genes) were 120,978 (YRI/CEU), 111,227 (YRY/AS), and 81,557 (CEU/AS).

Population Genetic Analyses

Tajima's D (Tajima 1989), Fu and Li's D^* and F^* (Fu and Li 1993) statistics, and diversity parameters θ_W (Watterson 1975) and π (Nei and Li 1979) were calculated using libsequence (Thornton 2003). Calibrated coalescent simulations were performed using the cosi package (Schaffner et al. 2005) and its best-fit parameters for YRI, CEU, and AS populations with 10,000 iterations. Coalescent simulations were conditioned on mutation rate, and recombination rate was derived

from UCSC tables (http://genome.ucsc.edu/, last accessed January 30, 2014, snpRecombRateHamap table).

The ML-ratio Hudson, Kreitman, and Aguadé (HKA) test was performed using the MLHKA software (Schaffner et al. 2005), as previously proposed (Fumagalli et al. 2009a). For human populations, 99 reference loci were randomly selected among National Institute of Environmental Health Sciences (NIEHS) loci that have been resequenced in YRI, CEU, and AS. Genotype data for 5-kb regions from 238 resequenced human genes were derived from the NIEHS SNPs Program web site (http://egp.gs.washington.edu, last accessed January 30, 2014). In particular, we only selected genes that had been resequenced in populations of defined ethnicity including YRI, CEU, and AS (NIEHS panel 2). After excluding windows with no SNPs and sequenced regions shorter than 5-kb, 211 windows were available (reference windows). The presence of resequencing gaps was accounted for in all calculations.

Haplotype Analysis and TMRCA Calculation

Haplotypes were inferred using PHASE (version 2.1) (Stephens et al. 2001; Stephens and Scheet 2005). Linkage disequilibrium (LD) analyses were performed using Haploview (v. 4.1) (Barrett et al. 2005), and blocks were identified through the confidence interval algorithm implemented in the software (Gabriel et al. 2002). Data for LD analysis were derived from resequencing data. Median-joining network to infer haplotype genealogy was constructed using NETWORK 4.6.1 (Bandelt et al. 1999). Estimate of the time to the most recent common ancestor (TMRCA) was obtained using an ML coalescent method implemented in GENETREE (Griffiths and Tavare 1994, 1995). The method assumes an infinite-site model without recombination; therefore, haplotypes and sites that violate these assumptions need to be removed: In the case of IFI16, we eliminated 1 variant. The mutation rate μ was obtained on the basis of the divergence between human and chimpanzee and under the assumption both that the species separation occurred 6 Ma (Glazko and Nei 2003) and of a generation time of 25 years. Using this μ and θ ML (θ_{ML}), we estimated the effective population size parameter (N_e), which resulted equal to 18,000. With these assumptions, the coalescence time, scaled in $2N_e$ units, was converted into years. For the coalescence process, 10^6 simulations were performed. A second TMRCA estimate was obtained by application of a method (Evans et al. 2005) that calculates the average pairwise difference between all chromosomes and the MRCA: This value was converted into years on the basis of mutation rate retrieved as above. The standard deviation (SD) for this estimate was calculated as described previously (Thomson et al. 2000). Using this method, the TMRCA was calculated for the *IFI16*-5 kb region and for 5-kb windows from NIEHS genes (one window/gene). In particular, windows were randomly selected with the only requirement that they did not contain any resequencing gap. After discarding X-linked loci, and windows containing no SNPs, 200 windows were used for TMRCA calculation.

Results

IFI16 and AIM2 Evolved Adaptively in Primates

To analyze the evolutionary history of ALR genes (*IFI16*, *AIM2*, *MNDA*, and *PYHIN1*) in primates, we obtained coding sequence information for 16 species from public databases or by sequencing (see Materials and Methods). In humans, the *IFI16* gene carries a polymorphic segmental duplication of exon 7, with the sequence of the two exons being identical (see later). We applied a PCR-based approach to determine the status of the exon 7 segmental duplication in different nonhuman primates, namely eight chimpanzees, one gorilla, and one orangutan. Results indicated the presence of the duplicated exon in gorilla and orangutan, whereas all chimpanzee samples carried a single copy of exon 7. Sequencing of the *IFI16* mRNA in *M. fascicularis* and *C. aethiops* also showed the presence of a single copy of exon 7. Thus, the nonduplicated gene sequence was used for multiple species alignment.

We calculated the average nonsynonymous substitution/synonymous substitution rate (dN/dS, also referred to as ω) for the four ALR genes using the single-likelihood ancestor counting (SLAC) method (Kosakovsky Pond and Frost 2005). This analysis indicated that *IFI16* might evolve under positive selection, as the average dN/dS was higher than 1, whereas *AIM2*, *MNDA*, and *PHYIN1* showed dN/dS values < 1 (supplementary table S3, Supplementary Material online), suggesting a role for purifying selection.

Nonetheless, positive selection might act on a few sites within a gene, which is elsewhere selectively constrained. To test this possibility, and to gain further insight into the evolutionary history of *IFI16*, we applied ML analyses implemented in the PAML package (Yang 1997, 2007). Because recombination might yield false-positive results when testing for selection (Anisimova et al. 2003), we first screened the alignments for evidence of recombination using GARD (Kosakovsky Pond et al. 2006); this analysis underscored no recombination breakpoint in any alignment. Thus, we used the codeml program to compare models of gene evolution that allow (NSsite models M2a and M8) or disallow (NSsite models M1a and M7) a class of codons to evolve with $dN/dS > 1$. For *IFI16* and *AIM2*, the two neutral models were rejected in favor of the positive selection models; this result was confirmed using different codon frequency models (F61 and F3x4) (table 1).

To identify specific sites subject to positive selection, we applied two methods, the BEB analysis (with a cutoff of 0.90) from M8 and the MEME (with the default cutoff of 0.1): Only sites detected using both methods were considered, and these are shown in figure 1.

We next mapped positively selected sites in *IFI16* and *AIM2* onto protein domain or three-dimensional structures.

Table 1
LRT Statistics for Models of Variable Selective Pressure among Sites (F3X4 and F61 Models of Codon Frequency)

Gene (Number of Codons)	df	$-2\Delta\ln L$	*P*	Percentage of Sites (Average dN/dS)
AIM2 (347)				
F3X4				
M1a vs. M2a	2	13.13	0.001	18.94 (2.4)
M7 vs. M8	2	15.54	0.0004	17.2 (2.5)
F61				
M1a vs. M2a	2	7.80	0.02	22.2 (2.0)
M7 vs. M8	2	9.16	0.01	22.0 (2.0)
IFI16 (731)				
F3X4				
M1a vs. M2a	2	39.32971	2.88×10^{-09}	14.8 (3.9)
M7 vs. M8	2	39.04452	3.32×10^{-09}	17.8 (3.6)
F61				
M1a vs. M2a	2	32.75	7.74×10^{-8}	22.74 (2.9)
M7 vs. M8	2	32.59	8.37×10^{-8}	29.46 (2.6)
MNDA (414)				
F3X4				
M1a vs. M2a	2	6.30	0.04	8.2 (3.0)
M7 vs. M8	2	6.31	0.04	10.3 (2.7)
F61				
M1a vs. M2a	2	3.33	0.19	—
M7 vs. M8	2	2.64	0.27	—
PYHIN1 (492)				
F3X4				
M1a vs. M2a	2	5.11	0.08	—
M7 vs. M8	2	5.12	0.08	—
F61				
M1a vs. M2a	2	2.88	0.24	—
M7 vs. M8	2	2.98	0.22	—

NOTE.—M1a is a nearly neutral model that assumes one ω class between 0 and 1 and one class with $\omega = 1$; M2a (positive selection model) is the same as M1a plus an extra class of $\omega > 1$. M7 (null model) assumes that $0 < \omega < 1$ is beta distributed among sites in 10 classes; M8 (selection model) has an extra class with $\omega \geq 1$; $2\Delta\ln L$: twice the difference of the natural logs of the maximum likelihood of the models being compared; *P*: *P* value of rejecting the neutral models (M1a or M7) in favor of the positive selection model (M2a or M8); percentage of sites (average *dN/dS*): estimated percentage of sites evolving under positive selection by M8 (*dN/dS* for these codons).

Three positively selected sites in AIM2 map to the HIN domain (fig. 1), and one is located in the PYD (helix α4). The crystal structure of the HIN domain has been solved, and it displays two oligonucleotide/oligosaccharide-binding (OB) folds forming the DNA-binding surface (Jin et al. 2012). Two of the positively selected residues in AIM2 are located in the region responsible for DNA binding, one in the OB1 region and one in the OB2 region (Jin et al. 2012) (fig. 1). In particular, 166T is immediately adjacent to a residue (165F) that, if mutated, strongly reduces DNA binding (Burckstummer et al. 2009). The third positively selected residue is located at the N terminus of the HIN domain, which is not directly involved in DNA binding.

In IFI16, we identified 14 positively selected sites. Three of them are located in the PYD domain; by comparison with the crystal structure of the homologous AIM2 domain (Jin et al. 2013), these residues are predicted to be located in the α1 (2G) and α6 (92P) helices and in the short loop connecting the α2 to the α3 helix (37L) (fig. 1) (Jin et al. 2013). The majority of positively selected sites in IFI16 are clustered in the interdomain regions (separating the two HIN or the PYD from the first HIN) (fig. 1). One of these sites (141G) is located within an accessory nuclear localization signal (referred to as motif-4) (Li et al. 2012). The positively selected 615 position in the second HIN domain could be mapped onto the crystal structure and was found to be located at the DNA-binding interface formed by the four IFI16 monomers (fig. 1) (Jin et al. 2012).

To explore possible variations in selective pressure among different lineages for *IFI16* and *AIM2*, we next tested whether a model that allows d*N*/d*S* to vary along branches (model M1) had significant better fit to the data than a model that assume one same d*N*/d*S* across the entire phylogeny (model M0) (Yang and Nielsen 1998). This was indeed the case for *IFI16* (table 2), indicating that different primates experienced variable levels of selective pressure at this gene. We thus used the BS-REL method (Kosakovsky Pond et al. 2011) to identify lineages on which a subset of sites has evolved under positive selection. BS-REL makes no a priori assumption about which lineages are more likely to represent selection targets; the method identified the internal branch leading to catarrhini in *IFI16* (fig. 2). This branch was cross-validated using the branch-site models implemented in PAML (Zhang et al. 2005), which apply a LRT to compare a model (MA) that allows positive selection on one or more lineages (foreground lineages) with a model (MA1) that does not allow such positive selection (table 2). Through BEB analysis, the PAML branch-site model allows identification of specific sites evolving under positive selection in the foreground branches; this procedure is accurate but has low statistical power (Zhang et al. 2005). Because MEME was specifically developed to detect episodic positive selection (in addition to pervasive selection), at least some lineage-specific BEB sites should have been identified by the MEME analysis we performed on the whole phylogeny. Indeed, BEB identified five sites for the catarrhini branch (fig. 1) and two were detected by MEME. All branch-specific sites are located in interdomain regions (fig. 1). Interestingly, the 131K residue is located within one of the two major nuclear localization signals (motif-2) (Li et al. 2012).

Evolution of ALR Genes in the Human Lineage

We next applied a recently developed population genetics–phylogenetics approach to study the evolution of ALR genes in the human lineage. Specifically, we applied the GammaMap program (Wilson et al. 2011) that jointly uses intraspecific variation and interspecific diversity to estimate the DFEs (i.e., selection coefficients, γ) along coding regions. To this

Table 2
LRT Statistics for Models of Positive Selection on Specific Branches

Gene	LRT[a] Model	Foreground Branch[a]	Codon Frequency Model	Degree of Freedom	$-2\Delta\ln L$[b]	P[c]
AIM2	M0 vs. M1[d]		F3x4	23	31.89	0.10
IFI16	M0 vs. M1[d]		F3x4	22	40,22	0.01
	MA1 vs. MA[e]	Catarrhini	F3x4	1	15.23	9.52×10^{-5}

[a]Likelihood ratio test (LRT).
[b]$2\Delta\ln L$: Twice the difference of the natural logs of the maximum likelihood of the models being compared.
[c]P value of rejecting the neutral model in favor of the positive selection model.
[d]M0 and M1 are free-ratio models, which assume all branches to have the same ω (M0) or allow each branch to have its own ω (M1).
[e]MA and MA1 are branch-site models that assume four classes of sites: the MA model allows a proportion of codons to have $\omega \geq 1$ on the foreground branches (those to be tested for selection), whereas the MA1 model does not.

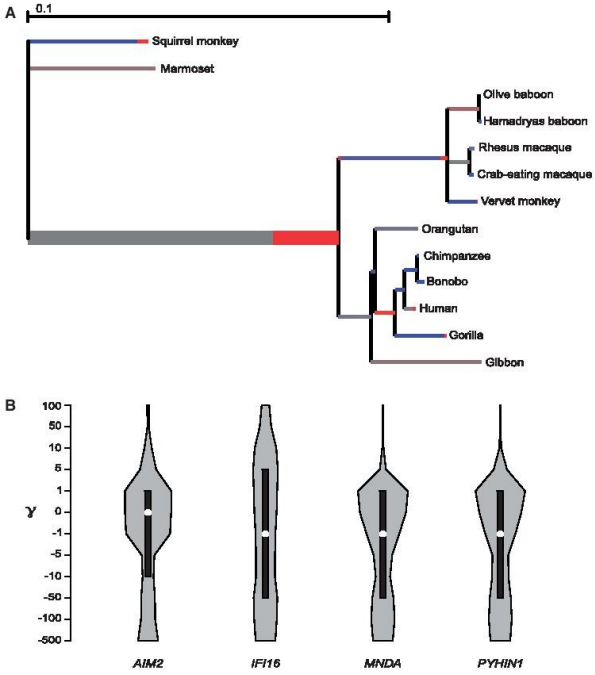


FIG. 2.—Lineage-specific selection and DFE analysis. (*A*) Branch-site analysis of positive selection in IFI16. Branch lengths are scaled to the expected number of substitutions per nucleotide, and branch colors indicate the strength of selection (ω). Red, positive selection ($\omega > 5$); blue, purifying selection ($\omega = 0$); gray, neutral evolution ($\omega = 1$). The proportion of each color represents the fraction of the sequence undergoing the corresponding class of selection. Thick branches indicate statistical support for evolution under episodic diversifying selection as determined by BS-REL. (*B*) Violin plot of selection coefficients (γ) for ALR genes (median, white dot; interquartile range, black bar). Selection coefficients are classified as strongly beneficial (100, 50), moderately beneficial (10, 5), weakly beneficial (1), neutral (0), weakly deleterious (−1), moderately deleterious (−5, −10), strongly deleterious (−50, −100), and inviable (−500).

aim, we exploited data from the 1000 Genomes Pilot project deriving from the low-coverage whole-genome sequencing of 179 individuals with different ancestry: CEU, YRI from Nigeria, and AS (Japanese plus Chinese) (1000 Genomes Project Consortium et al. 2010). The ancestral sequence was reconstructed by parsimony from the human, chimpanzee, orangutan and macaque sequences. We first applied GammaMap to obtain the overall distribution of selection coefficients along the four ALR genes. A general preponderance of codons evolving under negative selection ($\gamma < 0$) was observed for all genes excluding *AIM2*, which showed most codons to evolve with selection coefficients around neutrality (ranging from −1, weakly deleterious to 1, weakly beneficial). *IFI16* showed the highest proportion of codons with $\gamma > 5$ (fig. 2)

GammaMap allows to identify specific codons evolving under positive selection. Herein, we defined positively selected codons as those having a cumulative probability >0.80 of $\gamma \geq 1$. Five such codons were identified in *IFI16* and none in the remaining ALR genes. Two of the *IFI16* sites had previously been identified in the positive selection analysis we conducted on the whole mammalian phylogeny (supplementary table S4, Supplementary Material online). Three of the positively selected sites identified by GammaMap are located in the PYD domain (fig.1, supplementary table S4, Supplementary Material online). Two of them (67K and 70E) are within the α5 helix (Jin et al. 2013) (fig.1); the corresponding residues in the AIM2 PYD are predicted to be highly exposed (Jin et al. 2013).

Population Genetic Differentiation in Human Populations

We next addressed the role of natural selection in the shaping of genetic diversity at ALR genes in human populations. To this aim, we initially performed an analysis of population genetic differentiation, herein measured as F_{ST} (Wright 1950). High-F_{ST} values suggest that natural selection drives allele frequencies in distinct populations to differ more than expected on the basis of drift or demography alone. To analyze human population genetic differentiation along the ALR cluster, we exploited two partially independent sets of data. The first set is accounted for by genotype data generated by the 1000 Genomes Pilot project. Using these data, we calculated F_{ST} for the three pairwise comparisons (YRI/CEU, AS/CEU, and YRI/AS) (Wright 1950) in sliding windows moving along the genomic region where the four ALR genes are located. Sliding window analyses pose a multiple testing problem that is difficult to correct because of the nonindependence of windows. Moreover, the 1000 Genomes Pilot Project data suffer from a bias in the site frequency spectrum (SFS), with reduced power to detect low-frequency variants (1000 Genomes Project Consortium et al. 2010). To partially account for these limitations, we applied an outlier approach by obtaining F_{ST} distributions for the three pairwise comparisons in sliding windows from 2,000 randomly selected human genes. This allowed calculation of the 95th percentile and identification of regions in the ALR gene cluster above this threshold. A complementary set of data, namely the SNP genotypes from the Human Genome Diversity Panel (Li et al. 2008), was also used to analyze population genetic differentiation at the ALR gene cluster. For all HGDP SNPs within the cluster, we obtained F_{ST} values among continental groups; these values were compared with the distribution of F_{ST} calculated for HGDP variants in the same minor allele frequency (MAF) class. The HGDP panel includes 52 populations distributed worldwide and therefore represents a set of data largely independent from the 1000 Genomes Pilot Project. We thus focused on region of high F_{ST} identified using both data sets.

As shown in figure 3, three variants (rs856090:A>G, rs1614254:T>C, rs947275:T>C) were found to be outliers among HGDP continental groups in F_{ST} distribution values (ranks = 0.965, 0.951, and 0.981, respectively). Two of them are within the *IFI16* gene, and they are located in a peak of significantly high F_{ST} in all pairwise comparisons (YRI/CEU, YRI/AS, and CEU/AS), as assessed from the 1000 Genomes Pilot project data. Interestingly, susceptibility alleles for rheumatoid arthritis and for celiac disease (rs1772408:T>C) were identified in this region through a genome-wide association study (GWAS) (Zhernakova et al. 2011). Also, the F_{ST} outliers flank the exon 7 duplication allele (fig. 3).

Balancing Selection Maintains Genetic Diversity at the *IFI16* Gene in Human Populations

Given the results above, we decided to focus our attention on the *IFI16* gene region carrying the F_{ST} outlier SNPs. Because of the aforementioned SFS bias in the low-coverage 1000 Genomes data (1000 Genomes Project Consortium et al. 2010), we decided to resort to Sanger resequencing, so that further analyses could be performed within the framework of coalescent theory. Thus, we resequenced a 5-kb IFI16 region (*IFI16*-5 kb, fig. 3) encompassing the F_{ST} peaks and HGDP outliers in three HapMap populations, namely YRI, CEU, and AS. A PCR-based approach was also applied to these samples to determine allelic status for the exon 7 segmental duplication. The duplicated allele was detected in CEU only with a frequency of 7.5%.

Using Sanger-sequencing data, we first calculated F_{ST} for the entire *IFI16*-5 kb region in CEU/AS, CEU/YRI, and AS/YRI comparisons; these values were compared with the distribution of F_{ST} calculated for 5-kb windows (hereafter referred to as reference windows) deriving from 238 genes resequenced by the NIEHS Program in the same populations. For the YRI/CEU comparison, an F_{ST} of 0.38 was obtained, corresponding to a percentile rank of 0.972 in the distribution of reference windows, confirming high population differentiation at the *IFI16*-5 kb region.

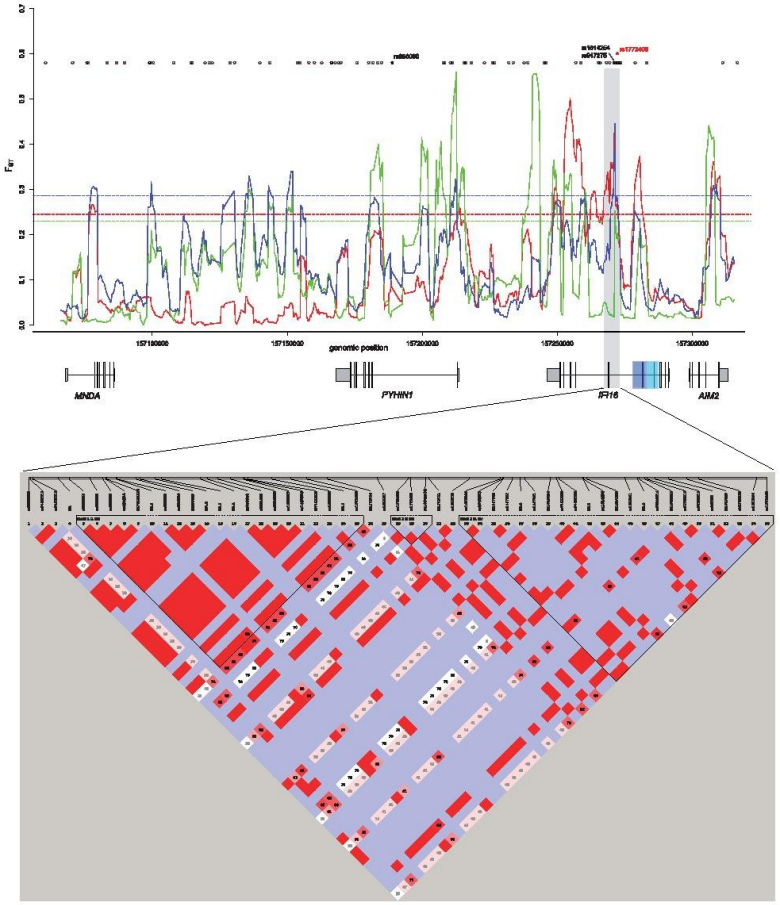


Fig. 3.—F_{ST} analysis of the ALR gene cluster. Data from the 1000 Genomes Pilot Project were used to calculate F_{ST} in sliding windows of 20 SNPs moving along the ALR gene cluster (NCBI/hg18, chr1:157063927–157317926) with a step of three SNPs (upper panel). Color codes refer to population comparisons: red, YRI/CEU; blue, YRI/AS; and green, CEU/AS. Horizontal dashed lines represent the 95th percentile in the distribution of F_{ST} calculated for sliding windows deriving from 2,000 randomly selected human genes. SNPs genotyped in the HGDP-CEPH panel are represented as gray circles (no unusual F_{ST} value among continental groups) or black circles (F_{ST} outliers); a SNP associated to rheumatoid arthritis and celiac disease is reported in red. The resequenced *IFI16* region is shaded in gray. The blue and cyan boxes represent the segmental duplication of exon 7. In the bottom panel, LD analysis for the *IFI16* resequenced region (5 kb) is shown. LD analysis was performed with the Haploview software using resequencing data, and blocks were identified through the implemented confidence interval algorithm (see Materials and Methods). Variants within the first LD block were used for Network and GENETREE analyses.

Table 3
Nucleotide Diversity and Neutrality Tests for the Analyzed *IFI16* Region

Population	N^a	S^b	Π ($\times 10^{-4}$)		Θ_W ($\times 10^{-4}$)		Tajima's *D*		Fu and Li's *D**		Fu and Li's *F**	
			Value	Rank[c]	Value	Rank[c]	Value (*P*)[d]	Rank[c]	Value (*P*)[d]	Rank[c]	Value (*P*)[d]	Rank[c]
YRI	40	47	28.22	**0.99**	22.70	**0.99**	0.86 (**0.028**)	**0.95**	1.06 (**0.014**)	**0.97**	1.18 (**0.01**)	**0.97**
CEU	40	33	15.31	0.94	15.94	**0.98**	−0.14 (0.44)	0.50	0.90 (0.11)	0.84	0.65 (0.23)	0.74
AS	40	26	21.84	**0.99**	14.01	**0.97**	1.92 (**0.031**)	**0.98**	1.78 (<**0.01**)	**0.99**	2.16 (<**0.01**)	**0.99**

NOTE.—Significant values are in bold.
[a]Sample size (chromosomes).
[b]Number of segregating sites.
[c]Percentile rank relative to a distribution of 238 5-kb windows from NIEHS genes.
[d]*P* value obtained by coalescent simulations.

Table 4
MLHKA Test for the *IFI16* Gene Region

Population	MLHKA	
	k^a	*P*
YRI	3.64	5.42×10^{-4}
CEU	3.63	1.644×10^{-2}
AS	2.21	5.05×10^{-3}

[a]Selection parameter ($k > 1$ indicates an excess of polymorphism compared with divergence; $k < 1$ indicates the opposite situation).

We next calculated θ_W (an estimate of the expected per site heterozygosity [Watterson 1975]) and π (the average number of pairwise sequence nucleotide differences between haplotypes [Nei and Li 1979]) and, again, compared the values with the distribution obtained from reference windows. The percentile ranks corresponding to *IFI16*-5 kb region in the distribution of NIEHS gene values indicate that the analyzed region displays high nucleotide diversity in all populations, although the rank of π in CEU did not reach the 95th percentile (table 3).

To confirm this observation, we applied a multilocus MLHKA test (Wright and Charlesworth 2004) by comparing polymorphism and divergence levels at the *IFI16*-5 kb with 99 randomly selected NIEHS genes resequenced in the same populations (YRI, CEU, and AS). Results, summarized in table 4, indicate that a significant excess of nucleotide diversity versus interspecies divergence is detectable in all populations for the *IFI16* study region.

Thus, these data indicate that high nucleotide diversity at the *IFI16*-5 kb region may be selectively maintained in human populations. This observation suggests the action of balancing selection, although this latter usually results in low rather than high F_{ST} (Charlesworth 2006) (see Discussion).

An effect of balancing selection is a distortion of the SFS toward intermediate frequency alleles. Common neutrality tests based on the SFS include Tajima's *D* (D_T) (Tajima 1989) and Fu and Li's *D** and *F** (Fu and Li 1993). Because population history beside affecting selective processes also influences the SFS, the significance of neutrality tests was evaluated by performing coalescent simulations with population genetics models that incorporate demographic scenarios (see Materials and Methods). We also applied an empirical comparison by calculating the percentile rank of D_T, F^*, and D^* in the *IFI16*-5 kb relative to 5-kb reference windows (obtained from 238 NIEHS genes). Neutrality tests indicated departure from neutrality with significantly positive values for all statistics in YRI and AS populations; conversely, no significant values were observed in CEU (table 3). Very similar results were obtained when different demographic models were used for coalescent simulations (Marth et al. 2004; Voight et al. 2005; Gutenkunst et al. 2009) (supplementary table S5, Supplementary Material online).

Further insight into the evolutionary history of a gene region can be gained by inferring haplotype genealogies. In particular, balancing selection is expected to result in two or more major clades with a deep coalescence time. Haplotype genealogies and inference of coalescent times may yield unreliable results in the presence of recombination. Thus, we selected a subregion based on LD; in particular, we used data from a 1.7-kb region (NCBI/hg18 chr1:157267850–157269530) with relatively high LD in all analyzed populations (fig. 3). As it is evident from both the median-joining network (Bandelt et al. 1999) and GENETREE analyses (Griffiths and Tavare 1995) (fig. 4), the haplotype genealogy is split into two major haplogroups (clades 1 and 2) separated by long branches. In line with the F_{ST} results, the major clade 2 haplotype is observed in CEU and AS but not in YRI. The time to the most recent common ancestor for the *IFI16* haplotype phylogeny was obtained using GENETREE (Griffiths and Tavare 1995) and amounted to 3.77 Myr (SD: 890 ky).

To obtain a second TMRCA estimate for the entire *IFI16*-5 kb region, we applied a previously described method (Evans et al. 2005) that calculates the average pairwise difference between all chromosomes and the MRCA; this value is then converted into years on the basis of the mutation rate (herein calculated on the basis of the number of fixed differences between chimpanzee and humans). Using this approach, we obtained a TMRCA of 4.6 Myr (SD:1.2 Myr) for the *IFI16*-5 kb region. As a comparison, TMRCA estimates were also

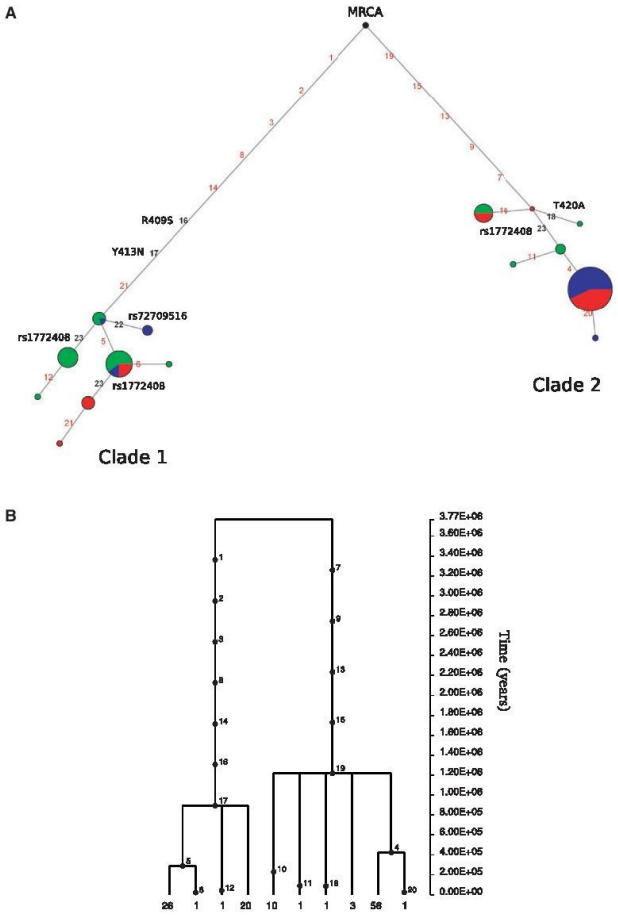


Fig. 4.—Haplotype analysis of *IFI16*. (*A*) Genealogy of haplotypes in the *IFI16* LD region (1.7-kb region (NCBI/hg18, chr1:157267850–157269530, see text) reconstructed through a median-joining network. Each node represents a different haplotype, with the size of the circle proportional to frequency. Nucleotide differences between haplotypes are indicated on the branches of the network. Color codes are as follows: YRI, green; CEU, blue; and AS, red. The most recent common ancestor (MRCA) is also shown. SNPs mentioned in the text are reported. (*B*) GENETREE for the LD subregion of *IFI16*. Variants are represented as black dots; the absolute frequency of each haplotype is reported.

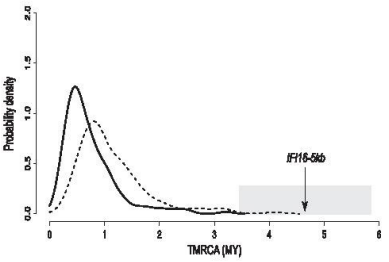


Fig. 5.—TMRCA estimates for the *IFI16*-5 kb region and for reference windows. Probability density plot of TMRCAs from 5-kb windows deriving from autosomal NIEHS genes (solid line); the upper (TMRCA+SD) estimates are also shown as hatched lines. The TMRCA estimate for the *IFI16*-5 kb region is indicated with upper- and lower bounds (gray shading).

obtained for 5-kb reference windows using haplotypes from YRI, CEU, and AS subjects, as derived from autosomal NIEHS genes. The mean TMRCA for these windows amounted to 0.76 Myr, in line with previous estimates for human autosomal loci (Garrigan and Hammer 2006). As shown in figure 5, the TMRCA estimate or the *IFI16*-5 kb region is a clear outlier in the TMRCA distribution obtained from reference windows.

Overall, these analyses indicate that the *IFI16* region has been a target of long-standing balancing selection.

To analyze the relationship among the major haplotype clades and the exon duplication allele, as well as the risk variant for autoimmune diseases, two variants located outside the LD region were included in the network: rs1772408:T>C (the GWAS SNP [Zhernakova et al. 2011]) and rs72709516:C>T, which is in full LD with the exon 7 duplication allele ($r^2 = 1$ in CEU). Analysis of the haplotype network indicated that, with the exception of few Asian and African haplotypes, rs1772408:T>C separates the two major clades; interestingly, two nonsynonymous variants located in exon 6, rs1057027:A>C (R409S) and rs1057028:A>T (Y413N), are also located on the major branches of the haplotype phylogeny; in line with this observation, the two SNPs are in full LD with the autoimmune risk variant in CEU ($r^2 = 1$ in both instances). Finally, a minority of CEU haplotypes in clade 1 are defined by the derived allele at rs72709516:C>T, which is in phase with the nonduplicated exon 7 allele. Analysis of 20 additional HapMap subjects of European descent confirmed full LD between the exon 7 segmental duplication polymorphism and rs72709516, and indicated a MAF of 0.05. This is in line with the reported MAF of 0.041 for rs72709516, as determined by the 1000 Genomes Project in CEU. Thus, either the exon 7 segmental duplication is neutrally evolving or it is subject to very recent/weak selection; its low frequency places it beyond the detection power of most tests based on haplotype homozygosity.

Discussion

The sensing of foreign genetic material is essential to trigger defensive responses that are important for organism survival to infections. An ever-increasing number of nucleic acid sensors are being identified in mammalian cells, revealing a complex machinery devoted to the detection of the invading pathogen (or of cell damage) and to the transduction of alert signals (Desmet and Ishii 2012). These cellular systems are expected to be engaged in a constant arms race with viruses and other microbial agents (Quintana-Murci and Clark 2013). Genetic conflicts leave signatures on the host genome, and protein regions directly contacting pathogen components are expected to evolve under the strongest diversifying selection. Therefore, evolutionary analyses can be applied both to study the history of host–pathogen interaction and, as recently demonstrated, to identify regions and residues directly involved in viral recognition or, more generally, in antiviral activity (Mitchell et al. 2012). On their side, viruses also evolve products that interfere with pathogen sensing, suggesting that positive selection on host proteins may also result from evolution away from viral recognition.

Results herein indicate that *AIM2* and *IFI16*, the best characterized members of the ALR family, evolved under positive selection in primates. Although only four selected residues were identified in *AIM2*, it should be noted that we defined positively selected sites by the combined use of two methods, BEB and MEME. Although this choice was taken to limit the number of false positive results, we most likely underestimated the number of selected sites, as these methods have different power to detect episodic and pervasive selection (Murrell et al. 2012). Three positively selected sites in AIM2 (166T and 333T) and in IFI16 (615V) are located at HIN domain/DNA-binding interface (fig. 1); one of them (166T) is adjacent to an AIM2 residue that has central importance in DNA binding (Burckstummer et al. 2009). Although AIM2 and IFI16 have been reported to bind dsDNA of both viral and bacterial origin, irrespective of GC content or sequence composition (Fernandes-Alnemri et al. 2009; Hornung et al. 2009; Unterholzner et al. 2010), these results suggest that positively selected residues in the HIN domains evolved to modulate recognition of specific substrates. Interestingly, recent evidence has indicated that, in addition to dsDNA, IFI16 can detect stem-rich secondary structures in ssDNA, which are produced during the replication cycle of lentiviruses (Jakobsen et al. 2013). These observations and the extreme plasticity of the ALR cluster in mammals (Brunette et al. 2012) indicate that diversification at ALR genes is evolutionary advantageous and possibly confers wider specificity in foreign nucleic acid recognition.

Overall, among ALR genes, *IFI16* was found to be the target of the strongest diversifying selection, also showing lineage-specific selection in catarrhini and humans.

Several positively selected sites in IFI16 are located in the PYD domain, which is found in all ALR proteins and in other molecules including PYCARD (ASC). Analysis of PYD-containing proteins indicated that the α3 helix is directly involved in PYD–PYD interactions, whereas the α6 helix is the most variable in length and sequence (Jin et al. 2013); the α5 helix comprises many basic residues, some of which are highly exposed at the protein surface. Of the three positively selected sites, we identified in the analysis on the whole phylogeny, one immediately flanks the α3 and another is located within the α6 helices; two of the human-specific selected sites are located in the α5 helix. Overall, these results suggest that diversity at these sites might modulate association with cellular or viral components. Although IFI16 mainly signals through STING, during Kaposi sarcoma-associated herpesvirus infection, it binds PYCARD through the PYD domain (Kerur et al. 2011); this same domain was shown to directly bind BRCA1, which also shows evidence of positive selection in primates (Pavlicek et al. 2004). Also, viral proteins might have evolved to bind IFI16 through either the PYD or other protein domains. Interestingly, the pUL83 tegument protein encoded by human cytomegalovirus (HCMV), a human-specific pathogen and a herpesvirus family member, interacts with the IFI16 PYD domain and blocks its oligomerization upon DNA sensing (Li et al. 2012). pUL83 is a central HCMV mediator of immune evasion and is predicted to establish extensive contacts with the IFI16 PYD domain. Likewise, the ICP0 protein of herpes simplex virus (HSV-1) has been reported to directly bind IFI16 and to target it for degradation (Orzalli et al. 2012). Unfortunately, the molecular details of the ICP0-IFI16 interaction are presently unknown. In general, IFI16 is thought to play a central role in the immune response to herpesviruses; this is likely achieved through the nuclear localization of this sensor, as the viral genome is protected in the cytoplasm by the capsid, but becomes exposed in the nucleus. In line with this view, HSV-1 eludes the surveillance of IFI16 mutants with cytoplasmic localization (Li et al. 2012). IFI16 displays a multipartite nuclear localization signal and its nuclear translocation is regulated by acetylation and phosphorylation (Li et al. 2012), suggesting that IFI16 localization is finely tuned, possibly in a cell-type and stimulus-dependent manner (Veeranki and Choubey 2012). We found one site positively selected in the catarrhini lineage to be located in one of the two motifs that play a nonredundant and essential role in determining IFI16 nuclear localization. One additional site subject to diversifying selection in the whole phylogeny maps to one accessory nuclear localization signal. It will be interesting to determine whether the selective pressure exerted on these motifs is related to specific viral interactors, to coevolution with cellular cofactors, or is secondary to the appearance and spread of viral species with particular cell-type tropism.

Finally, several sites subject to diversifying selection are located in the spacer separating the two IFI16 HIN domains; this region also shows length variation due to the segmental duplication of exon 7 and, at least in humans, to alternative splicing (Johnstone et al. 1998). An interesting possibility is that, by altering the structure of the spacer, selected sites determine subtle differences in HIN domain relative orientation, eventually affecting substrate recognition (Johnstone et al. 1998).

The spacer region also carries polymorphisms that might represent balancing selection targets in human populations. Application of different population genetic tests indicated that the region around exon 6 displays elevated nucleotide diversity, an excess of polymorphism relative to divergence, and a shift in the SFS toward intermediate-frequency alleles. In line with these findings, haplotype analysis indicated the presence of two clades separated by long branches with a deep TMRCA. Overall, these features represent strong molecular signatures of long-term balancing selection. Notably, when relatively constant in time and space, balancing selection may also result in low population genetic differentiation (Charlesworth 2006). Conversely, our data indicate that the *IFI16* region displays unusually high F_{ST} values. The possible explanations for these observations are manifold and depend on the underlying reason for the maintenance of the balanced polymorphism/haplotype. Balancing selection may result from different effects, including variable environmental conditions and frequency-dependent selection (Charlesworth 2006). Because of the dynamic nature of these processes, distinct populations may experience variable pressures and, consequently, different relative frequencies of the selected allele(s), resulting in high differentiation. Also, it should be noted that the sliding-window F_{ST} analysis revealed other peaks of high genetic differentiation. We focused on the *IFI16* region, as it was detected by both the 1000 Genomes and by the HGDP genotype data, and because the region carries an autoimmune disease susceptibility variant. Nonetheless, we do not imply that all other ALR gene regions are neutrally evolving in human populations.

Analysis of the *IFI16* haplotype phylogeny suggested that the exon 7 segmental duplication polymorphism does not represent a balancing selection target; in line with the estimated TMRCA, the presence of the duplicated allele in other nonhuman primates (orangutan and gorilla) most likely results from inherent instability (whereby the duplication undergoes nonallelic homologous recombination) rather than from active maintenance due to selection. Conversely, two nonsynonymous variants (R409S and Y413N) in exon 6 separate the major branches of the haplotype network, as would be expected if they represented the selection targets in human populations. R409S and Y413N polymorphisms affect positions conserved among primates, and in CEU, these variants are in full LD with the autoimmune risk variant (rs1772408:T>C), suggesting that they might represent the

causal polymorphisms for RA and celiac disease. Interestingly, the rs1772408:T>C variant displays an opposite risk profile: The ancestral allele predisposes to celiac disease but is protective for RA (Zhernakova et al. 2011). This observation suggests that the balancing selection regime results from antagonistic pleiotropy. This is a situation where one locus is associated with more than one trait, with both beneficial and detrimental effects for fitness. Although variants with opposite risk effects on autoimmune diseases are relatively common (Sirota et al. 2009; Wang et al. 2010), few of these have been demonstrated to be maintained as balanced polymorphisms (Cagliani et al. 2011). This is possibly the result of the weak selective effect of autoimmune diseases, which often become clinically relevant at postreproduction ages (Sironi and Clerici 2010). In this case, although celiac disease presents early in life, the widespread use of gluten-containing foods has likely appeared too recently in human populations to account for the long-standing balancing selection signature we describe herein. Therefore, one possible explanation is that functionally different *IFI16* variants were originally maintained by antagonistic pleiotropy related to immune response against pathogens, with differential susceptibility to autoimmune diseases being a consequence. This hypothesis is in line with recent analyses of genes subject to long-standing balancing selection in humans (Cagliani et al. 2010, 2012; Segurel et al. 2012; Leffler et al. 2013) and with theoretical modelling of host–pathogen arms races (Tellier and Brown 2007).

In summary, our data indicate that, in analogy to other nucleic acid sensors (Fumagalli et al. 2010; Vasseur et al. 2011; Patel et al. 2012; Quintana-Murci and Clark 2013), AIM2 and IFI16 have evolved adaptively in primates; in particular a continuum of selective pressure acting on *IFI16* is observed as the gene also represents a selection target in human populations. We suggest that the underlying scenario is the result of an ancestral and still ongoing host–pathogen arms race and that the maintenance of susceptibility alleles for autoimmune diseases at *IFI16* represents an evolutionary tradeoff. Ultimately, our results provide evolutionary and functional information about candidate ALR gene variants that might affect immunologic phenotypes.

Supplementary Material

Supplementary tables S1–S5 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by a fellowship of the Doctorate School of Molecular Medicine, University of Milan to D.F.

Literature Cited

1000 Genomes Project Consortium, et al. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061–1073.
Anisimova M, Bielawski JP, Yang Z. 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol Biol Evol. 19:950–958.
Anisimova M, Nielsen R, Yang Z. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164:1229–1236.
Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol. 16:37–48.
Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265.
Brunette RL, et al. 2012. Extensive evolutionary and functional diversity among mammalian AIM2-like receptors. J Exp Med. 209:1969–1983.
Burckstummer T, et al. 2009. An orthogonal proteomic-genomic screen identifies AIM2 as a cytoplasmic DNA sensor for the inflammasome. Nat Immunol. 10:266–272.
Cagliani R, et al. 2010. Long-term balancing selection maintains trans-specific polymorphisms in the human TRIM5 gene. Hum Genet. 128:577–588.
Cagliani R, et al. 2011. Balancing selection is common in the extended MHC region but most alleles with opposite risk profile for autoimmune diseases are neutrally evolving. BMC Evol Biol. 11:171.
Cagliani R, et al. 2012. A trans-specific polymorphism in ZC3HAV1 is maintained by long-standing balancing selection and may confer susceptibility to multiple sclerosis. Mol Biol Evol. 29:1599–1613.
Cereda M, Sironi M, Cavalleri M, Pozzoli U. 2011. GeCo++: a C++ library for genomic features computation and annotation in the presence of variants. Bioinformatics 27:1313–1315.
Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet. 2:e64.
Delport W, Poon AF, Frost SD, Kosakovsky Pond SL. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 26:2455–2457.
Desmet CJ, Ishii KJ. 2012. Nucleic acid sensing at the interface between innate and adaptive immunity in vaccination. Nat Rev Immunol. 12: 479–491.
Evans PD, et al. 2005. Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. Science 309:1717–1720.
Fernandes-Alnemri T, Yu JW, Datta P, Wu J, Alnemri ES. 2009. AIM2 activates the inflammasome and cell death in response to cytoplasmic DNA. Nature 458:509–513.
Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. Genetics 133:693–709.
Fumagalli M, et al. 2009a. Widespread balancing selection and pathogen-driven selection at blood group antigen genes. Genome Res. 19: 199–212.
Fumagalli M, et al. 2009b. Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions. J Exp Med. 206:1395–1408.
Fumagalli M, et al. 2010. Population genetics of IFIH1: ancient population structure, local selection and implications for susceptibility to type 1 diabetes. Mol Biol Evol. 27:2555–2566.
Fumagalli M, et al. 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genet. 7:e1002355.
Gabriel SB, et al. 2002. The structure of haplotype blocks in the human genome. Science 296:2225–2229.
Garrigan D, Hammer MF. 2006. Reconstructing human origins in the genomic era. Nat Rev Genet. 7:669–680.
Glazko GV, Nei M. 2003. Estimation of divergence times for major lineages of primate species. Mol Biol Evol. 20:424–434.
Goudet J. 2005. Hierfstat, a package for R to compute and test hierarchical F-statistics. Mol Ecol Notes. 5(1):184–186.
Griffiths RC, Tavare S. 1994. Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc Lond B Biol Sci. 344:403–410.

Griffiths RC, Tavare S. 1995. Unrooted genealogical tree probabilities in the infinitely-many-sites model. Math Biosci. 127:77–98.
Guindon S, Delsuc F, Dufayard JF, Gascuel O. 2009. Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol. 537:113–137.
Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5:e1000695.
Hornung V, et al. 2009. AIM2 recognizes cytosolic dsDNA and forms a caspase-1-activating inflammasome with ASC. Nature 458:514–518.
Jakobsen MR, et al. 2013. IFI16 senses DNA forms of the lentiviral replication cycle and controls HIV-1 replication. Proc Natl Acad Sci U S A. 110:E4571–E4580.
Jin T, Perry A, Smith P, Jiang J, Xiao TS. 2013. Structure of the absent in melanoma 2 (AIM2) pyrin domain provides insights into the mechanisms of AIM2 autoinhibition and inflammasome assembly. J Biol Chem. 288:13225–13235.
Jin T, et al. 2012. Structures of the HIN domain:DNA complexes reveal ligand binding and activation mechanisms of the AIM2 inflammasome and IFI16 receptor. Immunity 36:561–571.
Johnstone RW, Kershaw MH, Trapani JA. 1998. Isotypic variants of the interferon-inducible transcriptional repressor IFI 16 arise through differential mRNA splicing. Biochemistry 37:11924–11931.
Kerur N, et al. 2011. IFI16 acts as a nuclear pathogen sensor to induce the inflammasome in response to kaposi sarcoma-associated herpesvirus infection. Cell Host Microbe. 9:363–375.
Kosakovsky Pond SL, Frost SD. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 22:1208–1222.
Kosakovsky Pond SL, et al. 2006. Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol. 23:1891–1901.
Kosakovsky Pond SL, et al. 2011. A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol. 28:3033–3043.
Kosiol C, et al. 2008. Patterns of positive selection in six mammalian genomes. PLoS Genet. 4:e1000144.
Leffler EM, et al. 2013. Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science 339:1578–1582.
Li JZ, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–1104.
Li T, Diner BA, Chen J, Cristea IM. 2012. Acetylation modulates cellular distribution and DNA sensing ability of interferon-inducible protein IFI16. Proc Natl Acad Sci U S A. 109:10558–10563.
Marth GT, Czabarka E, Murvai J, Sherry ST. 2004. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics 166:351–372.
Mitchell PS, et al. 2012. Evolution-guided identification of antiviral specificity determinants in the broadly acting interferon-induced innate immunity factor MxA. Cell Host Microbe. 12:598–604.
Murrell B, et al. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 8:e1002764.
Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A. 76:5269–5273.
Orzalli MH, DeLuca NA, Knipe DM. 2012. Nuclear IFI16 induction of IRF-3 signaling during herpesviral infection and degradation of IFI16 by the viral ICP0 protein. Proc Natl Acad Sci U S A. 109:E3008–E3017.
Patel MR, Loo YM, Horner SM, Gale M Jr, Malik HS. 2012. Convergent evolution of escape from hepaciviral antagonism in primates. PLoS Biol. 10:e1001282.
Pavlicek A, et al. 2004. Evolution of the tumor suppressor BRCA1 locus in primates: implications for cancer predisposition. Hum Mol Genet. 13:2737–2751.
Quintana-Murci L, Clark AG. 2013. Population genetic tools for dissecting innate immunity in humans. Nat Rev Immunol. 13:280–293.
Rosenberg NA. 2006. Standardized subsets of the HGDP-CEPH human genome diversity cell line panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet. 70:841–847.
Schaffner SF, et al. 2005. Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 15:1576–1583.
Schattgen SA, Fitzgerald KA. 2011. The PYHIN protein family as mediators of host defenses. Immunol Rev. 243:109–118.
Segurel L, et al. 2012. The ABO blood group is a trans-species polymorphism in primates. Proc Natl Acad Sci U S A. 109:18493–18498.
Sironi M, Clerici M. 2010. The hygiene hypothesis: an evolutionary perspective. Microbes Infect. 12:421–427.
Sirota M, Schaub MA, Batzoglou S, Robinson WH, Butte AJ. 2009. Autoimmune disease classification by inverse association with SNP alleles. PLoS Genet. 5:e1000792.
Stephens M, Smith N, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet. 68:978–989.
Stephens M, Scheet P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet. 76:449–462.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.
Tellier A, Brown JK. 2007. Stability of genetic polymorphism in host-parasite interactions. Proc Biol Sci. 274:809–817.
Thomson R, Pritchard JK, Shen P, Oefner PJ, Feldman MW. 2000. Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. Proc Natl Acad Sci U S A. 97:7360–7365.
Thornton K. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics 19:2325–2327.
Unterholzner L, et al. 2010. IFI16 is an innate immune sensor for intracellular DNA. Nat Immunol. 11:997–1004.
Vasseur E, et al. 2011. The selective footprints of viral pressures at the human RIG-I-like receptor family. Hum Mol Genet. 20:4462–4474.
Veeranki S, Choubey D. 2012. Interferon-inducible p200-family protein IFI16, an innate immune sensor for cytosolic and nuclear double-stranded DNA: regulation of subcellular localization. Mol Immunol. 49:567–571.
Voight BF, et al. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc Natl Acad Sci U S A. 102:18508–18513.
Wang K, et al. 2010. Comparative genetic analysis of inflammatory bowel disease and type 1 diabetes implicates multiple loci with opposite effects. Hum Mol Genet. 19:2059–2067.
Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 7:256–276.
Wernersson R, Pedersen AG. 2003. RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res. 31:3537–3539.
Wilson DJ, Hernandez RD, Andolfatto P, Przeworski M. 2011. A population genetics-phylogenetics approach to inferring natural selection in coding sequences. PLoS Genet. 7:e1002395.
Wright S. 1950. Genetical structure of populations. Nature 166:247–249.
Wright SI, Charlesworth B. 2004. The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. Genetics 168:1071–1076.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555–556.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.
Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol. 46:409–418.

Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol. 22: 1107–1118.
Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 22:2472–2479.
Zhernakova A, et al. 2011. Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. PLoS Genet. 7:e1002004.

Associate editor: Gunter Wagner

3.5 RIG-I-like receptors evolved adaptively in mammals, with parallel evolution at LGP2 and RIG-I

Article

RIG-I-Like Receptors Evolved Adaptively in Mammals, with Parallel Evolution at *LGP2* and *RIG-I*

Rachele Cagliani[1,†], Diego Forni[1,†], Claudia Tresoldi[1], Uberto Pozzoli[1], Giulia Filippi[2], Veronica Rainone[3], Luca De Gioia[2], Mario Clerici[4,5] and Manuela Sironi[1]

*1 - **Scientific Institute,** IRCCS Eugenio Medea, 23842 Bosisio Parini LC, Italy*
*2 - **Department of Biotechnology and Biosciences,** University of Milan Bicocca, 20126 Milan, Italy*
*3 - **Department of Biomedical and Clinical Sciences,** University of Milan, 20100 Milan, Italy*
*4 - **Department of Physiopathology and Transplantation,** University of Milan, 20100 Milan, Italy*
*5 - **Don C. Gnocchi ONLUS Foundation,** IRCCS, 20100 Milan, Italy*

Correspondence to Manuela Sironi: *manuela.sironi@bp.lnf.it*
http://dx.doi.org/10.1016/j.jmb.2013.10.040
Edited by E. Freed and M. Gale

Abstract

RIG-I-like receptors (RLRs) are nucleic acid sensors that activate antiviral innate immune response. These molecules recognize diverse non-self RNA substrates and are antagonized by several viral inhibitors. We performed an evolutionary analysis of RLR genes (*RIG-I*, *MDA5*, and *LGP2*) in mammals. Results indicated that purifying selection had a dominant role in driving the evolution of RLRs. However, application of maximum-likelihood analyses identified several positions that evolved adaptively. Positively selected sites are located in all domains of *MDA5* and *RIG-I*, whereas in *LGP2* they are confined to the helicase domain. In both *MDA5* and *RIG-I*, the linkers separating the caspase activation and recruitment domain and the helicase domain represented preferential targets of positive selection. Independent selective events in RIG-I and LGP2 targeted the corresponding site (Asp421 and Asp179, respectively) within a protruding α-helix that grips the V-shaped structure formed by the pincer. Most of the positively selected sites in *MDA5* are in regions unique to this RLR, including a characteristic insertion within the helicase domain. Additional selected sites are located at the contact interface between MDA5 monomers, in spatial proximity to a positively selected human polymorphism (Arg843His) and immediately external to the parainfluenza virus 5 V protein binding region. Structural analyses suggested that the positively selected His834 residue is involved in parainfluenza virus 5 V protein binding. Data herein suggest that RLRs have been engaged in host-virus genetic conflict leading to diversifying selection and indicate parallel evolution at the same site in RIG-I and LGP2, a position likely to be of central importance in antiviral responses.

Introduction

The innate immune response represents the first-line defense against invading pathogens. The mechanisms responsible for the activation of innate immunity have been at least partially clarified in recent years. Thus, pathogen-associated molecular patterns, non-self structure shared by groups of pathogens, are recognized by pattern recognition receptors (PRRs). The interaction between pathogen-associated molecular patterns and PRRs activates signal transduction in immune and non-immune cells leading to the triggering of the inflammasome, the production of type 1 interferons, and the up-regulation of a vast family of antimicrobial factors known as interferon-stimulated genes. At least three different families of PRRs have been recognized: Toll-like receptors, NOD-like receptors, and RIG-I-like receptors (RLRs) [1]. PRRs are present on an extended variety of immune and non-immune cell types and are able to preferentially recognize specific families of pathogens; RLRs, in particular, seem to be better fit at sensing viral genetic material. The RLR family comprises three

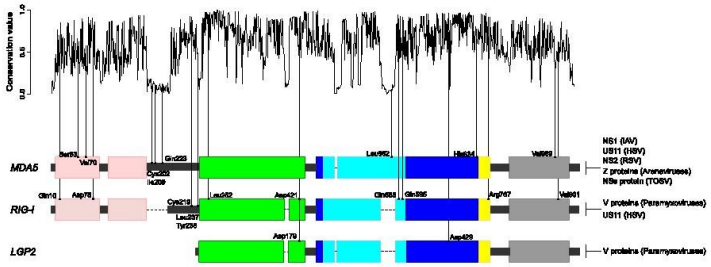


Fig. 1. Conservation score and schematic domain representation of the three RLRs. A normalized multi-gene and multi-species conservation score is shown for the RLR alignment. The domain structure of MDA5, RIG-I, and LGP2 is color-coded as follows: CARD, pink; HEL1, green; HEL2i, cyan; HEL2, blue; pincer, yellow; CTD, gray. Horizontal hatched lines denote gap regions in the alignment (i.e., inserted region in one RLR). Positively selected sites are mapped onto the conservation score and the domain structures; positions are referred to human sequences. A list of known viral inhibitors is reported for each RLR: NS1 from IAV (*i*nfluenza *A* virus) [70]; US11 from HSV (*h*erpes *s*implex *v*irus) [57]; NS2 from RSV (*r*espiratory *s*yncytial *v*irus) [55]; Z proteins from Arenaviruses [13]; NSs from TOSV (*Tos*cana *v*irus) [71]; V proteins from Paramyxovirus [12].

main sensors: retinoic acid-inducible gene I (*RIG-I* or *DDX58*), melanoma differentiation-associated gene 5 (*MDA5* or *IFIH1*), and laboratory of genetics and physiology 2 (*LGP2* or *DHX58*). These receptors are characterized by a cytoplasmic localization, recognize viral RNA, and induce signal transduction, upon binding to exogenous RNA, via the interaction with the adaptor molecule MAVS (or IPS-1), ultimately resulting in NF-κB- and IRF-3-dependent production of type I interferons and proinflammatory cytokines [2].

The three RLRs display a conserved core structure consisting of two DExD/H box helicase domains (HEL1 and HEL2) that are similar in structure to the helicases superfamily 1 and 2 (SF1 and SF2) but differ by the presence of an insertion domain in HEL2 (HEL2i), which plays an important role in the specific recognition of RNA [3]. The HEL2 domain of all three RLRs is connected to the C-terminal regulatory domain (CTD) by a so-called "pincer" or "bridge" [3,4] (Fig. 1). Both MDA5 and RIG-I display two N-terminal caspase activation and recruitment domains (CARDs) responsible for downstream signal transduction (Fig. 1). This makes the structure of RLRs unique in its domain composition, and contrasting hypotheses concerning the origin of such domain arrangements have been proposed [5,6].

Despite their structural similarity, MDA5 and RIG-I can activate immune responses against distinct viruses, possibly because of different specificities in substrate recognition. In fact, MDA5 has a preference for long double-stranded RNA (dsRNA), although higher-order structure or branched RNA seems necessary for optimal activation of this RLR [7]. Conversely, RNA molecules with 5′ triphosphorylated (5′ppp) ends and short dsRNAs activate RIG-I [2]. The function and specificity of LGP2 is less clear, but a recent work suggested that it binds blunt-ended dsRNA of different lengths [8]. Thus, some flaviviridae such as dengue and West Nile viruses are sensed by both MDA5 and RIG-I; however, RIG-I has the ability to detect a broader range of viral species including paramyxoviruses (e.g., measles, respiratory syncytial virus, Sendai virus), rhabdoviruses (e.g., rabies), influenza viruses, and flaviviruses (e.g., HCV) [2]. Conversely, the specific viral substrates of MDA5 are essentially restricted to picornaviruses (e.g., encephalomyocarditis virus) and vaccinia anuvirus [2]. Therefore, RLR receptors play a non-redundant role in virus sensing, suggesting that they evolved diverse specificities to broaden antiviral responses. As expected, viruses also managed to develop molecules that, by counteracting the action of RLRs, favor immune evasion. Most paramyxoviruses encode V proteins that bind the helicase domain of both mammalian and avian MDA5 molecules with an inhibitory effect [9,10]; V proteins also interact with LGP2 but are unable to bind and inactivate RIG-I [11,12]. The opposite is true for Z proteins encoded by New World arenaviruses, which block the function of RIG-I but are unable to bind MDA5 [13]. Interestingly, Z proteins from phylogenetically related Old World arenaviruses are ineffective against RIG-I [13]. As detailed in Fig. 1, additional viral products target MDA5 and/or RIG-I, suggesting that these molecules have been engaged in a constant

Table 1. Average dN/dS for the three RLR genes and number of positively selected sites

Gene	Number of species	Recombination breakpoints (position)	Average dN/dS (confidence intervals)	Number of positively selected sites		
				BEB	MEME	BEB and MEME
MDA5	46	—	0.293 (0.284,0.302)	19	74	8
RIG-I	42	—	0.403 (0.390,0.416)	30	78	11
LGP2	46	1 (470)	0.221 (0.213,0.230)	6[a]	19[a]	2[a]

[a] These numbers refer to the second region (alignment positions 471–678) of *LGP2* (see also Table 2).

genetic conflict with viruses and that they may evolve under virus-driven selection. In line with this view, analyses of genetic diversity in human populations indicated that amino-acid-replacing polymorphisms in *MDA5* and *LGP2* have represented targets of natural selection [14,15], whereas purifying selection was the major force acting on *RIG-I* [15]. Studies on inter-species genetic diversity provide insight into the long-standing evolutionary history of genes and may be exploited to identify protein sites or domains that directly interact with viral components or account for antiviral specificity [16]. Herein we analyzed the evolutionary history of RLRs in mammals and defined amino acid residues and specific lineages evolving under diversifying selection.

Results

Adaptive evolution of RLRs in mammals

To analyze the evolutionary history of RLR genes (*MDA5*, *RIG-I*, and *LGP2*) in mammals, we obtained coding sequence information for all available species from public databases (Supplementary Table S1). At least 42 species were available for each gene, including metatheria and eutheria and representing about 175 million years of mammalian history (Table 1) [17].

Previous studies have indicated that recombination can largely inflate type I error rates when models of positive selection are applied [18]. This is because

Table 2. Likelihood ratio test statistics for models of variable selective pressure among sites (F3x4 and F61 models of codon frequency)

Gene	Model	Degrees of freedom	−2ΔLnL	*p* Value (corrected *p* value)	Percentage of sites (average dN/dS)
MDA5	F3x4				
	M1a *versus* M2a	2	80.83	2.81×10^{-18}	2.6 (2.3)
	M7 *versus* M8	2	130.33	5.00×10^{-29}	9.5 (1.4)
	F61				
	M1a *versus* M2a	2	62.73	2.39×10^{-14}	2.6 (2.1)
	M7 *versus* M8	2	111.64	5.72×10^{-25}	10.2 (1.4)
LGP2, region 1	F3x4				
	M1a *versus* M2a	2	0	1 (1)	—
	M7 *versus* M8	2	63.80	1.40×10^{-14} (2.80×10^{-14})	—
LGP2, region 2	F3x4				
	M1a *versus* M2a	2	31.97	1.14×10^{-7} (2.28×10^{-7})	0.8 (3.2)
	M7 *versus* M8	2	36.73	1.05×10^{-8} (2.10×10^{-8})	2.8 (1.5)
	F61				
	M1a *versus* M2a	2	39.58	2.54×10^{-9} (5.08×10^{-9})	3.4 (2.1)
	M7 *versus* M8	2	58.33	2.15×10^{-13} (4.30×10^{-13})	7.6 (1.5)
RIG-I	F3x4				
	M1a *versus* M2a	2	131.83	2.36×10^{-29}	4.4 (2.6)
	M7 *versus* M8	2	150.90	1.72×10^{-33}	6.0 (1.9)
	F61				
	M1a *versus* M2a	2	88.79	5.27×10^{-20}	4.0 (2.3)
	M7 *versus* M8	2	113.85	1.90×10^{-25}	7.8 (1.7)

Note: M1a is a nearly neutral model that assumes one ω class between 0 and 1 and one class with $\omega = 1$; M2a (positive selection model) is the same as M1a plus an extra class of $\omega > 1$. M7 is a null model assuming that $0 < \omega < 1$ is beta distributed among sites; M8 (positive selection model) is the same as M7 but also includes an extra category of sites with $\omega > 1$. 2ΔLnL is twice the difference of the natural logs of the maximum likelihood of the models being compared; *p* value is the *p* value of rejecting the neutral models in favor of the positive selection model; for *LGP2*, the *p* value has been Bonferroni corrected (for two tested regions); percentage of sites (average dN/dS) is the estimated percentage of sites evolving under positive selection by M2a and M8 (dN/dS for these codons).

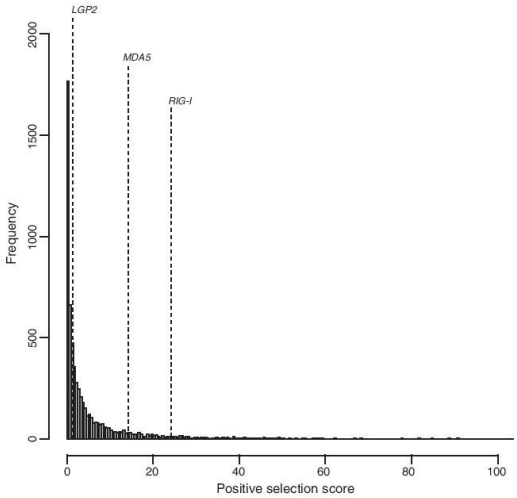


Fig. 2. Analysis of positive selection scores. Positive selection score distribution for 6426 genes (see the text). The position of *LGP2*, *MDA5*, and *RIG-I* is shown.

most methods used to infer positive selection assume that the phylogenetic tree and branch lengths are constant across all sites in the alignment, a tenet that is invalid in the presence of recombination. Indeed, recombination may introduce apparent substitution rate heterogeneity among sites [19] and may cause the estimated phylogeny to have excessively long terminal branches [20]. Thus, we first screened the three alignments for the presence of recombination breakpoints using genetic algorithm recombination detection (GARD) [21]. This program uses phylogenetic incongruence among segments of a sequence alignment to detect the best-fit number and location of recombination breakpoints. No evidence of recombination was detected for *MDA5* and *RIG-I*, whereas GARD detected one breakpoint in *LGP2* (at alignment position 470) (Table 1). Taking this information into account, we calculated the average non-synonymous substitution/synonymous substitution rate (dN/dS, also referred to as ω) for the three genes using the single-likelihood ancestor counting (SLAC) method [22]. SLAC estimates branch lengths and substitution rates using maximum likelihood and, in the presence of recombination, allows fitting of branch lengths separately for each segment. The algorithm then uses these estimates to fit a codon model and to obtain a global dN/dS ratio with confidence intervals; thus, the calculated ω accounts for recombination. For the three genes, the dN/dS rate was much lower than 1 (Table 1), indicating a major role for purifying selection in shaping genetic diversity at RLRs. However, whereas constraints on protein function and structure might result in overall purifying selection being the primary evolutionary force acting on gene regions, diversifying selection might involve specific sites or domains. To test this possibility, we applied maximum-likelihood analyses implemented in the phylogenetic analysis by maximum likelihood (PAML) package [23,24]. Specifically, we used the *codeml* program to compare models of gene evolution that allow (NSsite models M2a and M8, positive selection models) or disallow (NSsite models M1a and M7, null models) a class of codons to evolve with $dN/dS > 1$ using the F3x4 model of codon frequency. This analysis was performed for the *MDA5* and *RIG-I* alignments, as well as independently for the two regions of *LGP2*, according to the recombination breakpoint. As reported in Table 2, for both *MDA5* and *RIG-I*, both null models were rejected in favor of the positive selection models; the same occurred for the second region of *LGP2* (after Bonferroni correction for two tests, to account for alignment splitting). These

Table 3. Likelihood ratio test statistics for models of variable selective pressure along branches and branch-site tests (F3x4 model)

Gene	Model	−2ΔLnL	Degrees of freedom	*p* Value
MDA5	**M0** *versus* **M1**	261.84	90	1.09×10^{-18}
LGP2, region 2	**M0** *versus* **M1**	114.42	89	0.036
RIG-I	**M0** *versus* **M1**	224.63	81	1.95×10^{-15}

Single branch analysis

Gene	Foreground branch (MA *versus* MA1)	−2ΔLnL	Degrees of freedom	*p* Value (FDR corrected *p* value)
MDA5	Tasmanian devil	10.08	1	7.15×10^{-6} (4.29×10^{-5})
	Tree shrew	9.44	1	0.0021 (0.0042)
	Guinea pig	9.43	1	0.0021 (0.011)
	Alpaca	7.21	1	0.0072 (0.011)
	Insectivora	0.19	1	0.535 (0.642)
RIG-I	Xenarthra	36.53	1	1.5×10^{-9} (9.00×10^{-9});
	Squirrel	11.92	1	5.5×10^{-4} (1.65×10^{-3})
	Golden hamster	4.17	1	0.041 (0.082)
	Dog	0	1	1 (1)
	Fereuungulata	2.25	1	0.13 (0.20)
	Horse	0	1	1 (1)

Note: M0 and M1 are free-ratio models that assume all branches to have the same ω (M0) or allow each branch to have its own ω (M1); MA and MA1 are branch-site models that assume four classes of sites: the MA model allows a proportion of codons to have $\omega \geq 1$ on the foreground branches, whereas the MA1 model does not. 2ΔLnL is twice the difference of the natural logs of the maximum likelihood of the models being compared. Similar results were obtained using the F61 codon model (data not shown).

results were confirmed using the F61 model of codon frequency (Table 2).

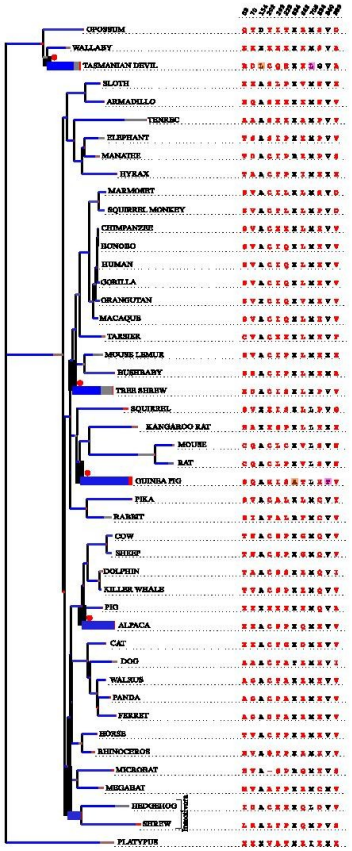
Overall, these analyses indicate that the three RLR genes evolved adaptively. The power to detect positive selection is influenced by the number and phylogenetic distance of the species being analyzed [25]. Few large scale with as many species as those analyzed herein are available. Thus, to evaluate the strength of selection acting on RLRs compared to other genes, we exploited data from a genome-wide analysis of 29 mammalian genomes [26]. In particular, we obtained positive selection scores (see Materials and Methods) for the three RLRs and for all genes ($n = 6426$) with a similar number of available species as RLRs [26]. Relative to the distribution of these 6426 genes, positive selection scores for *LGP2*, *MDA5*, and *RIG-I* corresponded to the 44.1th, 85.9th, and 91.3th percentiles, respectively (Fig. 2). The probability of randomly sampling three genes with at least one percentile value equal to or higher than each of those observed for the RLRs amounts to 0.0251, confirming that RLRs represented preferential selection targets during mammalian evolution.

Identification of positively selected sites and branch-specific selection

In order to identify specific sites subject to positive selection, we applied the Bayes empirical Bayes (BEB) analysis (with a cutoff of 0.90) from M8 [27,28] and the mixed effects model of evolution (MEME) (with a default cutoff of 0.1) [29]. Only sites detected using both methods were considered targets of positive selection: we identified 8, 11, and 2 positively selected sites in MDA5, RIG-I, and LGP2, respectively (Table 1 and Fig. 1).

We next analyzed the location of positively selected sites in the three genes relative to domain organization and inter-gene homology. Thus, we performed a multi-species, multi-gene alignment and calculated the degree of conservation at each position using Scorecons [30]. As expected, sequence stretches showing the strongest conservation were observed in the HEL1 and HEL2 domains (although MDA5 displays a unique 29-amino-acid insertion in HEL2i; Fig. 1). The positively selected sites are located in all domains of MDA5 and RIG-I, whereas they are confined to the LGP2 helicase domain (Fig. 1). Notably, four out of eight positively selected sites in MDA5 (Cys202, Ile209, Gly223, and Leu662) fall in regions that are unique to this RLR, including the 29-amino-acid insertion in HEL2i. Conversely, several targets of diversifying selection in RIG-I are located within sequence stretches of high homology among RLRs. Finally, positive selection was detected at the Asp421 and Asp179 residues in RIG-I and LGP2, respectively; these sites correspond to the same position in the two RLRs (Fig. 1), indicating independent selective events at the same residue.

In order to explore possible variations in selective pressure among different mammals and to identify sites subject to episodic selection (i.e., selection along one or few lineages), we first tested whether models that allow dN/dS to vary along branches (model M1) had significant better fit to the data than



models that assume one same dN/dS across the entire phylogeny (model M0) [31]. This condition was verified for the three RLRs (Table 3). We thus used the branch site-random effects likelihood (BS-REL) method [32] to identify lineages on which a subset of sites have evolved under positive selection. One advantage of BS-REL is that it makes no *a priori* assumption about which lineages are more likely to represent selection targets. Whereas no significant branch was detected for *LGP2* (region 2), BS-REL identified five and six branches with evidence of positive selection in *MDA5* and *RIG-I*, respectively (Figs. 3 and 4). These results were cross-validated using the branch-site models implemented in PAML [33], which apply a likelihood ratio test to compare a model (MA) that allows positive selection on one or more lineages (foreground lineages) with a model (MA1) that does not allow such positive selection (Table 3). As suggested [34], a false discovery rate (FDR) correction was applied to these *p* values, as multiple hypotheses are being tested on the same dataset. PAML analysis confirmed two of the *RIG-I* branches identified by BS-REL, with the hamster lineage being borderline after FDR (Table 3 and Fig. 4). As for *MDA5*, all branches were detected by both methods with the exception of the insectivora (Table 3 and Fig. 3).

The PAML branch-site models offer the possibility of identifying specific sites that evolve under positive selection in the foreground branches; this is achieved through implementation of a BEB analysis, which is accurate but has low statistical power [33]. We reasoned that, because MEME was specifically developed to detect episodic positive selection (in addition to pervasive selection), most lineage-specific BEB sites should have been identified by the MEME analysis we performed on the whole phylogeny. As shown in Fig. 4, BEB identified sites subject to episodic positive selection in *RIG-I* for the Xenarthra and squirrel branches, and all these sites were also detected by MEME. Likewise, for *MDA5*, two of the four sites identified by BEB (in the Tasmanian devil and guinea pig lineages) were among those found by MEME in the initial analysis (Fig. 3). The position of branch-specific sites relative to RLR domains is reported in Supplementary Fig. S1.

Fig. 3. Branch-site analysis of positive selection for MDA5. Branch lengths are scaled to the expected number of substitutions per nucleotide, and branch colors indicate the strength of selection (dN/dS or ω). Red, positive selection (ω > 5); blue, purifying selection (ω = 0); gray, neutral evolution (ω = 1). The proportion of each color represents the fraction of the sequence undergoing the corresponding class of selection. Thick branches indicate statistical support for evolution under episodic diversifying selection as determined by BS-REL. Dots denote branches that were also detected to be under positive selection using the PAML branch-site models (after FDR correction for multiple tests). Positively selected sites are also shown (positions refer to the human sequence). X symbols denote missing information. Red, sites subject to positive selection in the whole phylogeny; magenta shading, lineage-specific positively selected sites as detected by BEB; orange shading, lineage-specific positively selected sites as detected by both BEB and MEME.

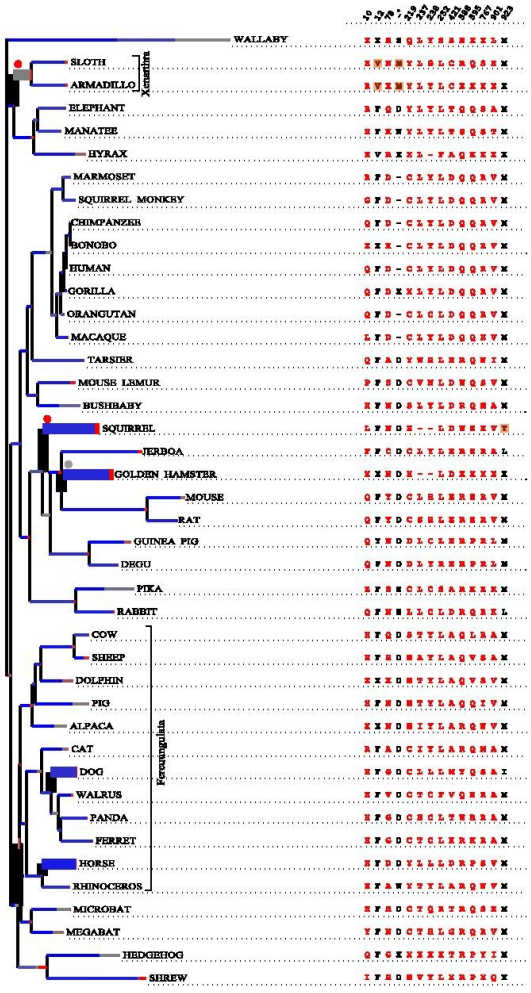


Fig. 4. Branch-site analysis of positive selection for RIG-I. The same as in Fig. 3. Gray dots denote branches that were detected to be under positive selection using the PAML branch-site models but did not withstand FDR correction for multiple tests. The site marked with an asterisk represents a single amino acid insertion in a minority of species, which is positively selected in Xenarthra.

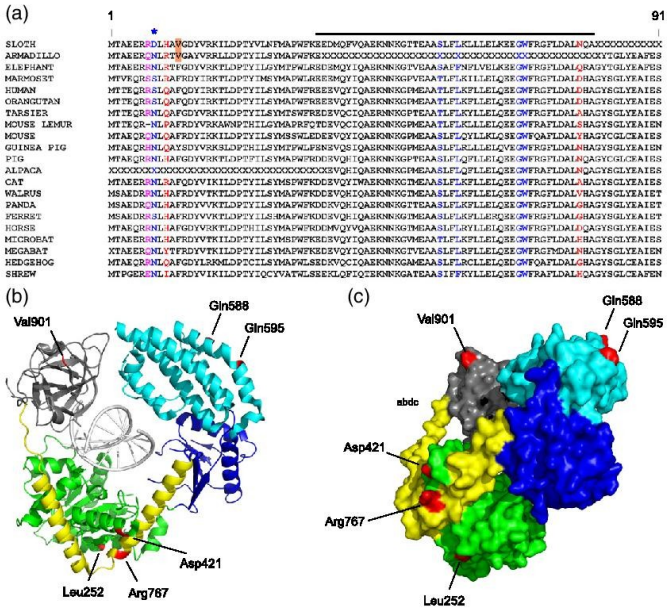


Fig. 5. Analysis of positively selected sites in RIG-I. Positively selected sites are always shown in red. (a) Multiple alignment of the first CARD domain of RIG-I for a few representative mammalian species. Color codes are as follows: blue, mutations that affect RIG-I function; orange shading, positively selected site in the Xenarthra lineage; magenta, human polymorphic site (Arg7Cys). The asterisks denoted the site that undergoes phosphorylation when a serine residue is present. The horizontal line denotes amino acids that are removed as a result of alternative splicing (exon 2 skipping). (b) Ribbon diagram of RIG-I (ΔCARD) bound to dsRNA; domains are color-coded as in Fig. 1. (c) Surface of RIG-I (slightly rotated compared to A to show all positively selected sites within these domains).

Analysis of positively selected sites and of human polymorphisms

As shown in Fig. 1, in both MDA5 and RIG-I, three positively selected sites are located in the spacers separating the second CARD domain from the helicase domain. These spacers are relatively short (91 and 55 residues for MAD5 and RIG-I, respectively), and by random uniform sampling, we determined that they represent preferential targets of positive selection in both genes (i.e., assuming a random distribution of selected sites, the likelihood of identifying three in the spacers amounts to 0.027 and 0.023 for MDA5 and RIG-I, respectively). In line with this analysis, a fourth site positively selected in Xenarthra was identified in the RIG-I spacer region (Fig. 4 and Supplementary Fig. S1).

As for CARDs, all sites we identified (with the exception of one residue specifically selected in the Tasmanian devil lineage; Supplementary Fig. S1) were located in the first domain. In the case of RIG-I, this domain is essential for binding to TRIM25 [35], an ubiquitin ligase that delivers K63-linked polyubiquitin moieties to lysine residues in the two CARDs [36]. Ubiquitination up-regulates RIG-I activity [36]. Several mutations in the first CARD domain that abolish CARD ubiquitination have been generated (Fig. 5a) [35,37], and an alternative splicing event that removes amino acids 36–80 in the protein product (thus encompassing the positively selected Asp78

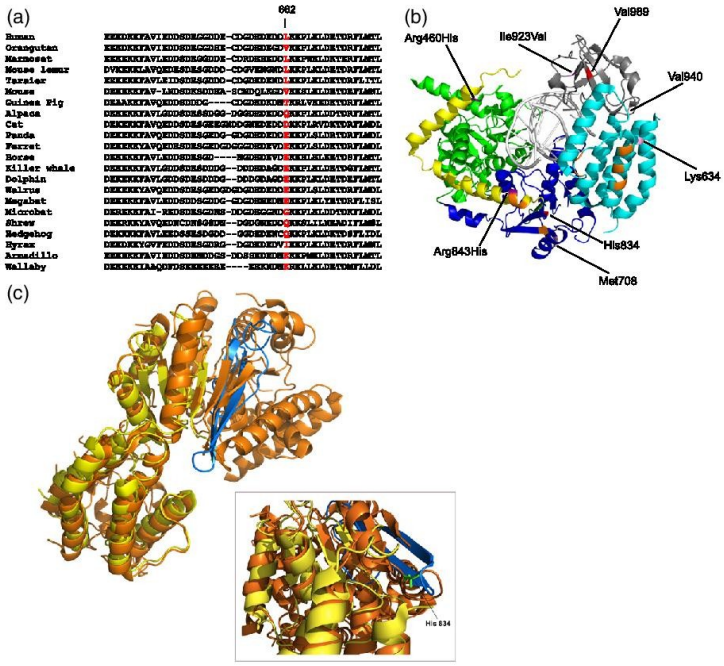


Fig. 6. Analysis of positively selected sites in MDA5. Positively selected sites are always shown in red. (a) Multiple alignment of the unique MDA5 HEL2i insertion for a few representative mammalian species. (b) Ribbon diagram of MDA5; domains are color-coded as in Fig. 1. Sites in pink are positively selected in specific lineages. Positions in orange form the head surface of the monomer interaction surface; the dark-green β-sheet directly interacts with paramyxovirus V proteins; positions in violet represent human amino-acid-replacement polymorphisms. (c) Superimposition of the structure of the HEL2 domain of MDA5 (orange) in complex with PIV5 (blue) (PDB code: 4I1S) and the whole MDA5 (yellow) (PDB code: 4GL2). In the enlargement, the position of His834 is shown in green.

site) is unable to bind TRIM25 [35]. Likewise, phosphorylation of the Ser8 residue decreases TRIM25 binding and RIG-I activity [38], and the Arg7Cys human polymorphism affects RIG-I ubiquitination levels [39]. Interestingly, the serine residue at position 8 is almost exclusively present in primates (Fig. 5); we detected two positively selected sites at positions 10 and 12 (this latter specific for the Xenarthra) (Figs. 4 and 5a), suggesting that variants in this N-terminal region might modulate RIG-I regulation in a species-specific fashion.

For both RIG-I and MDA5, crystal structures have been solved without CARD domains (ΔCARD). Thus, to gain further insight into the role of positively selected sites, we mapped these onto available RLR three-dimensional (3D) structures. In RIG-I, Asp421 is located in an helix (α17) protruding from HEL1 and forming a shaft that grips the V-shaped structure formed by the pincer region [3] (Fig. 5b). Notably, another positively selected site in RIG-I (Arg767) is located in the pincer and faces the Asp421 residue (Fig. 5b). As for the two selected sites in HEL2i, these are located in an α-helix external to the cavity that binds RNA. In general, all positively selected residues that could be mapped onto the 3D structure of RIG-I involve highly exposed positions (Fig. 5c).

The crystal structure of MDA5 (ΔCARD) lacks information for a portion of the unique insertion in HEL2i, including the positively selected Leu662 residue and the most C-terminal tail [40]. The HEL2i domain (which is absent in the other RLRs and in SF1/SF2 helicases) has a characteristic sequence composition with acidic N-terminal and basic C-terminal residues (Fig. 6a). Analysis of different mammalian species indicated that the positively selected 662 site can be occupied by neutral, as well as positively and negatively charged, amino acids (Fig. 6a).

One of the positively selected sites identified in MDA5 (Val989) is located within the β9 sheet of the CTD, which is known to directly interact with dsRNA [41]. Likewise, site 940, positively selected in the guinea pig lineage, is immediately adjacent to a residue that is in direct contact with the nucleic acid [41]. Interestingly, a functional human polymorphism associated with type 1 diabetes susceptibility [42,43], Ile923Val (rs35667974), is also located in the CTD and involves a site directly contacting dsRNA [41] (Fig. 6b). The CTD harbors a second human variant associated with different autoimmune diseases: the derived allele of the common Ala946Thr (rs1990760) polymorphism alters the expression of several immune response genes and predisposes to type 1 diabetes, systemic lupus erythematosus, rheumatoid arthritis, and multiple sclerosis [44,45]. This variant could not be visualized because it is located in a region that was not solved in the 3D structure. Overall, these data suggest that variation at CTD residues has the potential to influence MDA5 activity and to translate into phenotypic diversity.

Two additional human polymorphisms segregate in human populations and have been described to have been targets of selection during the recent evolution of the human species [14,15]. The His460Arg variant (Fig. 6b) was shown to affect MDA5 function by increasing apoptosis, and it predisposes to systemic lupus erythematosus [45]. No positively selected site is located in its proximity. The other variant, Arg843His, with unknown functional significance, is in the pincer, within a region where residues forming the head surface of the contact interface between MDA5 monomers are located (Fig. 6b). Likewise, the 708 residue, which is positively selected in the Tasmanian devil lineage, is also in proximity of the interaction surface (Fig. 6b). In the same region, we detected a site positively selected in the whole mammalian phylogeny, as well. In fact, 834His is at the C-terminal extremity of HEL2, close to a buried β-sheet, and serves as the interaction partner for paramyxovirus V proteins (Fig. 6b). Motz *et al.* have solved the crystal structure of the MDA5 HEL2 domain in complex with the C-terminal portion of the parainfluenza virus 5 V protein (PIV5) [46]. We superimposed the structure of the MDA5 HEL2 domain with PIV5 and the whole MDA5: as expected from the work of Motz *et al.*, the C-terminal domain of PIV5 occupied the position of two MDA5 β-sheets [46]. Indeed, PIV5 can disrupt the fold of HEL2, hence the impossibility to perform modeling studies using protein–protein docking. Nonetheless, from the superimposition of the two structures, it is clear that His834 is in a crucial position for the interaction of the two proteins since it is located in the region subject to unfolding upon interaction (Fig. 6c).

Discussion

We exploited the availability of an ever-increasing number of resequenced mammalian genomes, which allows high power in natural selection tests, to analyze the evolutionary history of RLRs in mammals and to identify selected sites. To minimize false positives, we applied a conservative approach, by screening for recombination and by requiring all neutral models to be rejected in favor of the positive selection models. Likewise, individual sites subject to positive selection were defined by the use of two methods, BEB and MEME. Again, this choice was motivated by the desire to limit the number of false positive results, although we possibly underestimated the number of selected sites. Indeed, MEME allows the distribution of dN/dS to vary from branch to branch at an individual site, resulting in the ability to detect both episodic and pervasive positive selection [29]; conversely, sites evolving under episodic selection are likely to be missed by BEB. Consistently, BEB analyses from the branch-site models detected seven sites that evolved under positive selection along specific lineages and five of these were identified by MEME without *a priori* branch specification, providing methodological validation. It is worth mentioning that, although we used two methods (BS-REL and the PAML branch-site models) to identify lineages subject to positive selection and we checked alignment quality, minor errors in the reference genome sequences of specific lineages might yield false positive results, especially for terminal branches. Clearly, the presence of lineage-specific selective events suggests that species-specific viral pathogens acted as selective pressures. Nonetheless, current information concerning the diversity and host range of animal viruses may be too limited [47] to allow speculation on which viral species have acted as major selective forces; also, some of the species detected herein such as the Xenarthra, which represent the more convincing evidence of episodic selection, are poorly studied. Finally, it should be noted that the fact that some branches display statistical evidence of episodic positive selection does not imply that the remaining lineages are neutrally evolving. Indeed, maximum-likelihood analyses (both M1 *versus* M2 and M7 *versus* M8) after

removal of significant branches still supported the action of positive selection for both *MDA5* and *RIG-I* (data not shown). Overall, our data indicate that RLRs have represented preferential selection targets during mammalian evolution, as also assessed from large-scale estimates of positive selection scores [26]. Indeed, even previous works that analyzed gene families involved in immune response such as T cell regulatory molecules [48] and interleukins [49], that used an approach similar to that applied herein, and that included several mammalian species identified a proportion of positively selected genes not higher than 60%.

Even accounting for its lack of CARD domains, we detected fewer positively selected sites in *LGP2* compared to *RIG-I* and *MDA5*. The role of this sensor is still poorly understood; initial evidences suggesting that it mainly functions as a negative regulator of the activity of other RLRs [50,51] have been challenged by the observation that LGP2 can positively participate to the antiviral responses triggered by RIG-I and MDA5 [52,53]. Most likely, however, LGP2 acts more as a modulator than as an active participant in antiviral response [54], possibly explaining weaker selective pressure on this RLR. Nonetheless, cross-gene comparison indicated that the same position in RIG-I (Asp421) and LGP2 (Asp179) has been independently targeted by selection, suggesting that this site plays a central role in processes involving both RLRs. In RIG-I, the Asp421 residue is highly exposed and located within a helix that functions as a shaft for the pincer, which, in turn, is thought to modulate the relative orientation of the helicase domain and the CTD [3]. Thus, variants in the shaft of the pincer (i.e., Arg767 in RIG-I) might affect substrate recognition by inducing minor changes in domain positioning. An alternative possibility is that the Asp421 and Asp179 residues in RIG-I and LGP2 are part of a surface recognized by one or more viral proteins. Indeed, RIG-I is targeted by many viral products (Fig. 1) [13,55–57], although the specific molecular details of these interactions are unknown. Beside position 421, most of the positively selected sites detected in RIG-I involve solvent-exposed residues that may function as anchors for viral protein binding. Exposed residues in host proteins are obvious targets for the binding of viral products [58], and the interaction between V proteins and MDA5 likely represents an exception to this rule. Indeed, Motz *et al.* showed that PIV5 recognizes a β-sheet in MDA5 that is normally buried within the helicase fold, suggesting that PIV5 has the ability to unfold the RLR [46]. Even if, due to the unfolding event, we were unable to perform protein–protein docking analyses, the position of His834 strongly suggests that it modulates the interaction between MDA5 and V proteins, confirming the view whereby at least some positively selected sites lie at host-pathogen interaction surfaces.

Some sites positively selected in mammals, as well as a human polymorphism previously described as a selection target [15], mapped to a region that forms part of the interaction surface between MDA5 monomers. Indeed, MDA5 forms filaments of monomers arranged in head to tail around dsRNA that are necessary to nucleate the formation of the active fibrillar form of MAVS [40,59]. By disrupting the MDA5 fold, PIV5 inhibits its polymeric arrangement [46] and other viral products might target the monomer interaction surface. Recent evidences indicated that RIG-I also arranges in filaments along dsRNA, and for both RLRs, this process allows the formation of CARD domain oligomeric patches along the fibril. In turn, this is made possible by the presence of a linker region (assumed to have a random-coil structure) that separates CARDs from HEL domains. Interestingly, data herein show that, for both MDA5 and RIG-I, these linker regions have been preferential targets of positive selection. One possible interpretation of this observation is that, being exposed on the fibril surface, linkers might represent extremely good candidates for virus-induced cleavage or misfolding. CARD domains represent the signaling units of MDA5 and RIG-I, by establishing homotypic interactions with the CARD of MAVS, which also carries sites positively selected in primates [60]; in addition, CARD domains regulate RLR activity. Indeed, the human Arg7Cys polymorphism affects RIG-I ubiquitination [36] and modulates the humoral response to rubella vaccination [61]. Overall, these data suggest that the positively selected sites we detected in CARDs might result from co-evolution with the homologous MAVS domain or with other interactors such as TRIM25, a possibility that could be addressed once crystal structures have been solved. Indeed, the presence of a phosphorylation site (Ser8) [38] only in primates suggests that distinct species adopt diverse regulatory mechanisms.

Finally, many of the positively selected sites in MDA5 are in regions that are unique to this RLR, such as the longer spacer separating the CARD and helicase domains and the characteristic insertion in HEL2i. This latter could not be solved in the 3D structure, but a similar insertion in the bacterial Hef helicase mediates branched DNA processing [62]. It is thus tempting to speculate that this insert in MDA5 modulates the recognition of specific substrates, possibly branched or higher-order RNA structures [7], and that charge variations at the positively selected Leu662 position might contribute to further diversification of the range of detected substrates. More generally, distinct RLRs might have evolved innovations relative to the other family members, as demonstrated by their non-redundant roles, and diversifying selection within these regions might contribute additional functional diversification in terms of either viral recognition or inhibitor antagonism.

Materials and Methods

Detection of positive selection

Mammalian sequences for *RIG-I*, *LGP2*, and *MDA5* were retrieved from the Ensembl[‡] and National Center for Biotechnology Information[§] databases. The list of species for each gene is reported in Supplementary Table S1. DNA alignments were performed using the RevTrans 2.0 utility [63], which uses the protein sequence alignment as a scaffold for constructing the corresponding DNA multiple alignment. This latter was checked and edited by hand to remove alignment uncertainties. For PAML analyses [24], we used trees generated by maximum-likelihood using the program PhyML [64].

The site models implemented in PAML have been developed to detect positive selection affecting only a few amino acid residues in a protein. These models treat the dN/dS (ω) ratio for any codon in the gene as a random variable from a statistical distribution, thus allowing ω to vary from site to site, assuming a constant rate at synonymous sites. To detect selection, we fitted site models that allow (M2a, M8) or disallow (M1a, M7) a class of sites to evolve with $\omega > 1$ to the data using the F3x4 and the F61 codon frequency model. Positively selected sites were identified using the BEB analysis (with a cutoff of 0.90), which calculates the posterior probability that each codon is from the site class of positive selection (under model M8) [27]. A second method, MEME (with the default cutoff of 0.1) [29], was applied to identify positively selected sites. MEME allows the distribution of ω to vary from site to site and from branch to branch at a site, therefore allowing the detection of both pervasive and episodic positive selection; the method has been shown to have more power than methods that assume constant ω across lineages [29].

We used GARD [21] to screen the alignments for recombination; GARD uses phylogenetic incongruence among fragments of a sequence alignment to detect recombination events; the application of a genetic algorithm allows searching for multiple breakpoints; the probability that each breakpoint is due to recombination (rather than, e.g., heterotachy) is assessed through Kishino–Hasegawa tests [65]. Thus, the method reaches high power and good accuracy in most evolutionary scenarios [21,66].

GARD, MEME [29], and SLAC [22] analyses were performed through the Datamonkey server[‖] [67].

To explore possible variations in selective pressure among different lineages, we applied the free-ratio models implemented in the PAML package: the M0 model assumes all branches to have the same ω, whereas M1 allows each branch to have its own ω [31]. The models are compared through likelihood ratio tests (degrees of freedom = total number of branches, −1). In order to identify specific branches with a proportion of sites evolving with $\omega > 1$, we used BS-REL [32]. This method implements branch-site models that simultaneously allow ω variation across branches and sites. BS-REL requires no prior knowledge about which lineages are of interest (i.e., are more likely have experienced episodic diversifying selection) and uses sequential likelihood ratio tests to identify significant branches (with final Holm's multiple testing correction). Branches identified using this approach were cross-validated using the branch-site likelihood ratio tests from PAML (the so-called modified model A and model MA1, "test 2") [33]. In this test, branches are divided *a priori* into foreground (those to be analyzed for positive selection) and background lineages, and a likelihood ratio test is applied to compare a model that allows positive selection on the foreground lineages with a model that does not allow such positive selection. An FDR correction was applied to account for multiple hypothesis testing (i.e., we corrected for the number of tested lineages), as previously suggested [34]. The advantage of this method is that it also implements a BEB analysis analogous to that described above to calculate the posterior probabilities that each site belongs to the site class of positive selection on the foreground lineages. Thus, BEB allows identification of specific sites that evolve under positive selection on specific lineages, although it has limited statistical power [33].

Positive selection scores for 11,827 genes were retrieved from a previous work [26]. Scores were calculated by combining (Fisher's method) nominal *p* values for all sites with $dN/dS > 0$ in each gene [26]. Data for LGP2, MDA5, and RIG-I derived from the analysis of 27, 30, and 26 species, respectively. Thus, we only considered genes with 26–31 available species ($n = 6426$).

Sequence and 3D structure analysis

Multi-species, multi gene alignments were performed using MAFFT version 7 [68] and used as an input for Scorecons [30]. The conservation score was calculated using the "valdar01" method, which accounts for amino acid frequency and stereochemical diversity and normalizes against redundancy in the alignment [69]. Because CARD domains are not present in LGP2, CARD conservation scores were calculated using a multiple alignment that comprises MDA5 and RIG-I sequences only.

Protein 3D structures for MDA5 (PDB ID: 4GL2) and RIG-I (PDB ID: 2YKG) were derived from the Protein Data Bank (PDB). Sites were mapped into structures using PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.2 Schrödinger, LLC). PyMOL files are available as supplementary material (3DModel_RIG-I and 3DModel_MDA5).

Acknowledgements

D.F. is supported by a fellowship of the Doctorate School of Molecular Medicine, University of Milan.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jmb.2013.10.040.

Received 16 July 2013;
Received in revised form 11 October 2013;
Accepted 30 October 2013
Available online 7 November 2013

Keywords:
RIG-I-like receptors;
MDA5;
RIG-I;
LGP2;
positive selection

† R.C. and D.F. contributed equally to this work.
‡ http://www.ensembl.org/index.html
§ http://www.ncbi.nlm.nih.gov/
‖ http://www.datamonkey.org

Abbreviations used:
RLR, RIG-I-like receptor; PRR, pattern recognition receptor; PAML, phylogenetic analysis by maximum likelihood; BS-REL, branch site-random effects likelihood; MEME, mixed effects model of evolution; BEB, Bayes empirical Bayes; SLAC, single-likelihood ancestor counting; GARD, genetic algorithm recombination detection; FDR, false discovery rate; CTD, C-terminal regulatory domain; CARD, caspase activation and recruitment domain; PIV5, parainfluenza virus 5 V protein; dsRNA, double-stranded RNA; 3D, three-dimensional.

References

[1] Takeuchi O, Akira S. Pattern recognition receptors and inflammation. Cell 2010;140:805–20.
[2] Loo YM, Gale M. Immune signaling by RIG-I-like receptors. Immunity 2011;34:680–92.
[3] Luo D, Ding SC, Vela A, Kohlway A, Lindenbach BD, Pyle AM. Structural insights into RNA recognition by RIG-I. Cell 2011;147:409–22.
[4] Kowalinski E, Lunardi T, McCarthy AA, Louber J, Brunel J, Grigorov B, et al. Structural basis for the activation of innate immune pattern-recognition receptor RIG-I by viral RNA. Cell 2011;147:423–35.
[5] Sarkar D, Desalle R, Fisher PB. Evolution of MDA-5/RIG-I-dependent innate immunity: independent evolution by domain grafting. Proc Natl Acad Sci USA 2008;105:17040–5.
[6] Zou J, Chang M, Nie P, Secombes CJ. Origin and evolution of the RIG-I like RNA helicase gene family. BMC Evol Biol 2009;9:85.
[7] Pichlmair A, Schulz O, Tan CP, Rehwinkel J, Kato H, Takeuchi O, et al. Activation of MDA5 requires higher-order RNA structures generated during virus infection. J Virol 2009;83:10761–9.
[8] Li X, Ranjith-Kumar CT, Brooks MT, Dharmaiah S, Herr AB, Kao C, et al. The RIG-I-like receptor LGP2 recognizes the termini of double-stranded RNA. J Biol Chem 2009;284:13881–91.
[9] Andrejeva J, Childs KS, Young DF, Carlos TS, Stock N, Goodbourn S, et al. The V proteins of paramyxoviruses bind the IFN-inducible RNA helicase, MDA-5, and inhibit its activation of the IFN-beta promoter. Proc Natl Acad Sci USA 2004;101:17264–9.
[10] Childs KS, Andrejeva J, Randall RE, Goodbourn S. Mechanism of MDA-5 Inhibition by paramyxovirus V proteins. J Virol 2009;83:1465–73.
[11] Childs K, Randall R, Goodbourn S. Paramyxovirus V proteins interact with the RNA helicase LGP2 to inhibit RIG-I-dependent interferon induction. J Virol 2012;86:3411–21.
[12] Parisien JP, Bamming D, Komuro A, Ramachandran A, Rodriguez JJ, Barber G, et al. A shared interface mediates paramyxovirus interference with antiviral RNA helicases MDA5 and LGP2. J Virol 2009;83:7252–60.
[13] Fan L, Briese T, Lipkin WI, Fan L, Briese T, Lipkin WI. Z proteins of New World arenaviruses bind RIG-I and interfere with type I interferon induction. J Virol 2010;84:1785–91.
[14] Fumagalli M, Cagliani R, Riva S, Pozzoli U, Biasin M, Piacentini L, et al. Population genetics of IFIH1: ancient population structure, local selection and implications for susceptibility to type 1 diabetes. Mol Biol Evol 2010;27:2555–66.
[15] Vasseur E, Patin E, Laval G, Pajon S, Fornarino S, Crouau-Roy B, et al. The selective footprints of viral pressures at the human RIG-I-like receptor family. Hum Mol Genet 2011;20:4462–74.
[16] Mitchell PS, Patzina C, Emerman M, Haller O, Malik HS, Kochs G. Evolution-guided identification of antiviral specificity determinants in the broadly acting interferon-induced innate immunity factor MxA. Cell Host Microbe 2012;12:598–604.
[17] Madsen O. Mammals (mammalia). In: Hedges SB, Kumar S, editors. New York: Oxford University Press; 2009. p. 459–61.
[18] Anisimova M, Nielsen R, Yang Z. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 2003;164:1229–36.
[19] Worobey M. A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria, and mitochondria. Mol Biol Evol 2001;18:1425–34.
[20] Schierup MH, Hein J. Consequences of recombination on traditional phylogenetic analysis. Genetics 2000;156:879–91.
[21] Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol 2006;23:1891–901.
[22] Kosakovsky Pond SL, Frost SD. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 2005;22:1208–22.
[23] Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 1997;13:555–6.
[24] Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 2007;24:1586–91.
[25] Anisimova M, Bielawski JP, Yang Z. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol 2001;18:1585–92.
[26] Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 2011;478:476–82.
[27] Anisimova M, Bielawski JP, Yang Z. Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol Biol Evol 2002;19:950–8.
[28] Yang Z, Wong WS, Nielsen R. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 2005;22:1107–18.
[29] Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. Detecting individual sites subject to episodic diversifying selection. PLoS Genet 2012;8:e1002764.
[30] Valdar WS. Scoring residue conservation. Proteins 2002;48:227–41.
[31] Yang Z, Nielsen R. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 1998;46:409–18.

[32] Kosakovsky Pond SL, Murrell B, Fourment M, Frost SD, Delport W, Scheffler K. A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol 2011;28:3033–43.
[33] Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 2005;22:2472–9.
[34] Anisimova M, Yang Z. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol 2007;24:1219–28.
[35] Gack MU, Kirchhofer A, Shin YC, Inn KS, Liang C, Cui S, et al. Roles of RIG-I N-terminal tandem CARD and splice variant in TRIM25-mediated antiviral signal transduction. Proc Natl Acad Sci USA 2008;105:16743–8.
[36] Oshiumi H, Matsumoto M, Seya T. Ubiquitin-mediated modulation of the cytoplasmic viral RNA sensor RIG-I. J Biochem 2012;151:5–11.
[37] Jiang X, Kinch LN, Brautigam CA, Chen X, Du F, Grishin NV, et al. Ubiquitin-induced oligomerization of the RNA sensors RIG-I and MDA5 activates antiviral innate immune response. Immunity 2012;36:959–73.
[38] Nistal-Villan E, Gack MU, Martinez-Delgado G, Maharaj NP, Inn KS, Yang H, et al. Negative role of RIG-I serine 8 phosphorylation in the regulation of interferon-beta production. J Biol Chem 2010;285:20252–61.
[39] Hu J, Nistal-Villan E, Voho A, Ganee A, Kumar M, Ding Y, et al. A common polymorphism in the caspase recruitment domain of RIG-I modifies the innate immune response of human dendritic cells. J Immunol 2010;185:424–32.
[40] Wu B, Peisley A, Richards C, Yao H, Zeng X, Lin C, et al. Structural basis for dsRNA recognition, filament formation, and antiviral signal activation by MDA5. Cell 2013;152:276–89.
[41] Li X, Lu C, Stewart M, Xu H, Strong RK, Igumenova T, et al. Structural basis of double-stranded RNA recognition by the RIG-I like receptor MDA5. Arch Biochem Biophys 2009;488:23–33.
[42] Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 2009;324:387–9.
[43] Shigemoto T, Kageyama M, Hirai R, Zheng J, Yoneyama M, Fujita T. Identification of loss of function mutations in human genes encoding RIG-I and MDA5: implications for resistance to type I diabetes. J Biol Chem 2009;284:13348–54.
[44] Cen H, Wang W, Leng RX, Wang TY, Pan HF, Fan YG, et al. Association of IFIH1 rs1990760 polymorphism with susceptibility to autoimmune diseases: a meta-analysis. Autoimmunity 2013;46:455–62.
[45] Molineros JE, Maiti AK, Sun C, Looger LL, Han S, Kim-Howard X, et al. Admixture mapping in lupus identifies multiple functional variants within IFIH1 associated with apoptosis, inflammation, and autoantibody production. PLoS Genet 2013;9:e1003222.
[46] Motz C, Schuhmann KM, Kirchhofer A, Moldt M, Witte G, Conzelmann KK, et al. Paramyxovirus V proteins disrupt the fold of the RNA sensor MDA5 to inhibit antiviral signaling. Science 2013;339:690–3.
[47] Anthony SJ, Epstein JH, Murray KA, Navarrete-Macias I, Zambrana-Torrelio CM, Solovyov A, et al. A strategy to estimate unknown viral diversity in mammals. MBio 2013;4:e00598-005913.
[48] Forni D, Cagliani R, Pozzoli U, Colleoni M, Riva S, Biasin M, et al. A 175 million year history of T cell regulatory molecules reveals widespread selection, with adaptive evolution of disease alleles. Immunity 2013;38:1129–41.
[49] Neves F, Abrantes J, Steinke JW, Esteves PJ. Maximum-likelihood approaches reveal signatures of positive selection in IL genes in mammals. Innate Immun 2013. http://dx.doi.org/10.1177/1753425913486687.
[50] Komuro A, Horvath CM. RNA- and virus-independent inhibition of antiviral signaling by RNA helicase LGP2. J Virol 2006;80:12332–42.
[51] Rothenfusser S, Goutagny N, DiPerna G, Gong M, Monks BG, Schoenemeyer A, et al. The RNA helicase Lgp2 inhibits TLR-independent sensing of viral replication by retinoic acid-inducible gene-I. J Immunol 2005;175:5260–8.
[52] Satoh T, Kato H, Kumagai Y, Yoneyama M, Sato S, Matsushita K, et al. LGP2 is a positive regulator of RIG-I- and MDA5-mediated antiviral responses. Proc Natl Acad Sci USA 2010;107:1512–7.
[53] Venkataraman T, Valdes M, Elsby R, Kakuta S, Caceres G, Saijo S, et al. Loss of DExD/H box RNA helicase LGP2 manifests disparate antiviral responses. J Immunol 2007;178:6444–55.
[54] Pippig DA, Hellmuth JC, Cui S, Kirchhofer A, Lammens K, Lammens A, et al. The regulatory domain of the RIG-I family ATPase LGP2 senses double-stranded RNA. Nucleic Acids Res 2009;37:2014–25.
[55] Ling Z, Tran KC, Teng MN. Human respiratory syncytial virus nonstructural protein NS2 antagonizes the activation of beta interferon transcription by interacting with RIG-I. J Virol 2009;83:3734–42.
[56] Mibayashi M, Martinez-Sobrido L, Loo YM, Cardenas WB, Gale M, Garcia-Sastre A. Inhibition of retinoic acid-inducible gene I-mediated induction of beta interferon by the NS1 protein of influenza A virus. J Virol 2007;81:514–24.
[57] Xing J, Wang S, Lin R, Mossman KL, Zheng C. Herpes simplex virus 1 tegument protein US11 downmodulates the RLR signaling pathway via direct interaction with RIG-I and MDA-5. J Virol 2012;86:3528–40.
[58] Franzosa EA, Xia Y. Structural principles within the human-virus protein–protein interaction network. Proc Natl Acad Sci USA 2011;108:10538–43.
[59] Berke IC, Yu X, Modis Y, Egelman EH. MDA5 assembles into a polar helical filament on dsRNA. Proc Natl Acad Sci USA 2012;109:18437–41.
[60] Patel MR, Loo YM, Horner SM, Gale M, Malik HS. Convergent evolution of escape from hepaciviral antagonism in primates. PLoS Biol 2012;10:e1001282.
[61] Ovsyannikova IG, Dhiman N, Haralambieva IH, Vierkant RA, O'Byrne MM, Jacobson RM, et al. Rubella vaccine-induced cellular immunity: evidence of associations with polymorphisms in the Toll-like, vitamin A and D receptors, and innate immune response genes. Hum Genet 2010;127:207–21.
[62] Nishino T, Komori K, Tsuchiya D, Ishino Y, Morikawa K. Crystal structure and functional implications of *Pyrococcus furiosus* Hef helicase domain involved in branched DNA processing. Structure 2005;13:143–53.
[63] Wernersson R, Pedersen AG. RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res 2003;31:3537–9.
[64] Guindon S, Delsuc F, Dufayard JF, Gascuel O. Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol 2009;537:113–37.
[65] Kishino H, Hasegawa M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA

sequence data, and the branching order in hominoidea. J Mol Evol 1989;29:170–9.
[66] Bay RA, Bielawski JP. Recombination detection under evolutionary scenarios relevant to functional divergence. J Mol Evol 2011;73:273–86.
[67] Delport W, Poon AF, Frost SD, Kosakovsky Pond SL. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 2010;26:2455–7.
[68] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 2013;30:772–80.
[69] Valdar WS, Thornton JM. Protein–protein interfaces: analysis of amino acid conservation in homodimers. Proteins 2001;42:108–24.
[70] Ruckle A, Haasbach E, Julkunen I, Planz O, Ehrhardt C, Ludwig S. The NS1 protein of influenza A virus blocks RIG-I-mediated activation of the noncanonical NF-kappaB pathway and p52/RelB-dependent gene expression in lung epithelial cells. J Virol 2012;86:10211–7.
[71] Gori-Savellini G, Valentini M, Cusi MG. Toscana virus NSs protein inhibits the induction of type I interferon by interacting with RIG-I. J Virol 2013;87:6660–7.

4 CONCLUSIONS

The study and identification of natural selection patterns in humans and, in general, in different species have different purposes. First of all,they improve our understanding of the evolutionary processes that shaped diversity in human populations; but they can also help to highlight specific variants that have functional roles or influence phenotypes.

The identification of non neutrally-evolving genomic regions and genes can also help understand the evolutionary history of modern complex diseases. In fact, variation in susceptibility genes for human common traits may be due to changes in selective pressure during the evolutionary history of our species.

Interspecific evolutionary studies can highlight how infectious agents have represented a selective force acting on hosts genes. In fact pathogens may either develop strategies to modulate the transcript level of hosts genes, with the aim of evading the host immune system, or they may encode molecules that directly bind host proteins and alter their function. That was the case of CD86: the Kaposi sarcoma–associated herpesvirus MIR2 protein directly binds CD86. Docking analysis indicated that the two positively selected sites identified by the interspecific study (positions 260 and 268) and located in the transmembrane region of CD86 are crucial for the interaction with MIR2; changing the aminoacid at these sites alters the binding pose of the two proteins and most likely affects the binding efficiency. Therefore, the selective pressure exerted by MIR2/MIR2-related proteins might have driven the evolution of the CD86 transmembrane region to decrease binding by viral-encoded ubiquitinases or to displace the ubiquitine ligase.

Pathogens have been shown to be also a major selective force in human populations: in particular I addressed the question whether adaptive events

have affected the spread of human disease alleles. I found that the disease variants described herein as selection target are preferentially located in regulatory regions; thus, selection at disease alleles is likely to have operated by modulating expression levels, which need to be finely tuned between efficient response to infection and maintenance of self-tolerance. In this respect, one of the most convincing examples I show here concerns an inflammatory bowel disease risk variant (rs762421): the risk allele at rs762421 strongly correlates with bacterial diversity, suggesting that it increased in frequency due to its conferring increased resistance to bacterial diseases. Indeed, the same allele increases the expression of ICOSLG in response to a bacterial superantigen. These results support the idea that the presence in the past of a number of different pathogens has influenced the spread of some autoimmune risk alleles in human populations, that conferred resistance to those pathogens.

Evolutionary analyses both at the intra- and inter- specific level can also be useful to identify a continuum in selective pressure acting on different timescales on a particular gene. For example I found that in CD207, a positively selected site in mammalian species immediately flanks a human polymorphic position representing a balancing selection target with known effect on sugar binding [69]. This indicates that natural selection has maintained variability in langerin forms that differ in binding specificity and may recognize distinct microbial glycan structures,ultimately affecting the susceptibility to specific infections. Given this premise, I provided a preliminary evidence for this hypothesis by showing that the CD207 haplotype defined by the selected variants is associated with protection to sexually transmitted HIV-1 infection.

The same situation is verified for the contact system, with the *KNG1* gene being a target of an extremely diverse array of pathogen species on different timescales. The results showed here reinforce the hypothesis that

KNG1 has a specific role in the modulation of immune response

Another fundamental aspect of evolutionary studies is that their results suggest caution when extrapolating information from specific experiments in model organisms, as a portion of genetic diversity at a specific gene could have accumulated not as a result of neutral processes but in response to adaptive events. An example of this is from the human and mouse HAVCR2 orthologs. In HAVCR2, two positively selected sites identified in the mammalian phylogeny are located at the top of pocket that accommodates PtdSer. This molecule is a central signal exposed by apoptotic cells and it is exploited by intracellular pathogens such as Leishmania and Toxoplasma to dampen the host response [70, 71]. In vitro experiments have indicated that substitution of the mouse residues at this pocket with the corresponding human aminoacids significantly decreases PtdSer binding [72], providing a direct evidence that positive selection at these sites affected the functional properties of HAVCR2.

Finally evolutionary studies can provide information about sites and domains that determine antiviral specificity as well as sensitivity to viral inhibitors. For example the evolutionary analyses of nucleic acids receptors (i.e. AIM-2 like and RIG-I like molecules) showed an ancestral and still ongoing host-virus arms race, providing information on the location and nature of adaptive changes, and highlighting the presence of functional variation.

5 BIBLIOGRAPHY

1. Cann H. M., de Toma C., Cazes L., Legrand M. F., Morel V., Piouffre L., Bodmer J., Bodmer W. F., Bonne-Tamir B., Cambon-Thomsen A., et al, "A human genome diversity cell line panel", *Science*, Vol. 296, no. 5566, 2002, pp. 261-262.

2. Handley L. J., Manica A., Goudet J., Balloux F., "Going the distance: Human population genetics in a clinal world", *Trends Genet*, Vol. 23, no. 9, 2007, pp. 432-439.

3. Fumagalli M., Sironi M., Pozzoli U., Ferrer-Admettla A., Pattini L., Nielsen R., "Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution", *PLoS Genet*, Vol. 7, no. 11, 2011, pp. e1002355.

4. Luijckx P., Fienberg H., Duneau D., Ebert D., "A matching-allele model explains host resistance to parasites", *Curr Biol*, Vol. 23, no. 12, 2013, pp. 1085-1088.

5. Paterson S., Vogwill T., Buckling A., Benmayor R., Spiers A. J., Thomson N. R., Quail M., Smith F., Walker D., Libberton B., et al, "Antagonistic coevolution accelerates molecular evolution", *Nature*, Vol. 464, no. 7286, 2010, pp. 275-278.

6. Strachan D. P., "Hay fever, hygiene, and household size", *Bmj*, Vol. 299, no. 6710, 1989, pp. 1259-1260.

7. Castillo-Davis C. I., Kondrashov F. A., Hartl D. L., Kulathinal R. J., "The functional genomic distribution of protein divergence in two animal phyla: Coevolution, genomic conflict, and constraint", *Genome Res*, Vol. 14, no. 5, 2004, pp. 802-811.

8. Sironi M., Menozzi G., Comi G. P., Cagliani R., Bresolin N., Pozzoli U., "Analysis of intronic conserved elements indicates that functional complexity might represent a major source of negative selection on non-coding sequences", *Hum Mol Genet*, Vol. 14, no. 17, 2005, pp. 2533-2546.

9. Kosiol C., Vinar T., da Fonseca R. R., Hubisz M. J., Bustamante C. D., Nielsen R., Siepel A., "Patterns of positive selection in six mammalian genomes", *PLoS Genet*, Vol. 4, no. 8, 2008, pp. e1000144.

10. Nielsen R., Hellmann I., Hubisz M., Bustamante C., Clark A. G., "Recent and ongoing selection in the human genome", *Nat Rev Genet*, Vol. 8, no. 11, 2007, pp. 857-868.

11. Charlesworth D., "Balancing selection and its effects on sequences in nearby genome regions", *PLoS Genet*, Vol. 2, no. 4, 2006, pp. e64.

12. Kimura M. (1983) The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

13. Tajima F., "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism", *Genetics*, Vol. 123, no. 3, 1989, pp. 585-595.

14. Watterson G. A., "On the number of segregating sites in genetical models without recombination", *Theor Popul Biol*, Vol. 7, no. 2, 1975, pp. 256-276.

15. Nei M., Li W. H., "Mathematical model for studying genetic variation in terms of restriction endonucleases", *Proc Natl Acad Sci U S A*, Vol. 76, no. 10, 1979, pp. 5269-5273.

16. Fay J. C., Wu C. I., "Hitchhiking under positive darwinian selection", *Genetics*, Vol. 155, no. 3, 2000, pp. 1405-1413.

17. Fu Y. X., Li W. H., "Statistical tests of neutrality of mutations", *Genetics*, Vol. 133, no. 3, 1993, pp. 693-709.

18. Schaffner S. F., Foo C., Gabriel S., Reich D., Daly M. J., Altshuler D., "Calibrating a coalescent simulation of human genome sequence variation", *Genome Res*, Vol. 15, no. 11, 2005, pp. 1576-1583.

19. Wright S., "Genetical structure of populations", *Nature*, Vol. 166, no. 4215, 1950, pp. 247-249.

20. Voight B. F., Kudaravalli S., Wen X., Pritchard J. K., "A map of recent

positive selection in the human genome", *PLoS Biol*, Vol. 4, no. 3, 2006, pp. e72.

21. Sabeti P. C., Schaffner S. F., Fry B., Lohmueller J., Varilly P., Shamovsky O., Palma A., Mikkelsen T. S., Altshuler D., Lander E. S., "Positive natural selection in the human lineage", *Science*, Vol. 312, no. 5780, 2006, pp. 1614-1620.

22. 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, "Evolutionary dynamics of human toll-like receptors and their different contributions to host defense", *PLoS Genet*, Vol. 5, no. 7, 2009, pp. e1000562.

23. Tang K., Thornton K. R., Stoneking M., "A new approach for using genome scans to detect recent positive selection in the human genome", *PLoS Biol*, Vol. 5, no. 7, 2007, pp. e171.

24. Hudson R. R., Kreitman M., Aguade M., "A test of neutral molecular evolution based on nucleotide data", *Genetics*, Vol. 116, no. 1, 1987, pp. 153-159.

25. Anisimova M., Nielsen R., Yang Z., "Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites", *Genetics*, Vol. 164, no. 3, 2003, pp. 1229-1236.

26. Yang Z., dos Reis M., "Statistical properties of the branch-site test of positive selection", *Mol Biol Evol*, Vol. 28, no. 3, 2011, pp. 1217-1228.

27. Yang Z., "PAML: A program package for phylogenetic analysis by maximum likelihood", *Comput Appl Biosci*, Vol. 13, no. 5, 1997, pp. 555-556.

28. Yang Z., "PAML 4: Phylogenetic analysis by maximum likelihood", *Mol Biol Evol*, Vol. 24, no. 8, 2007, pp. 1586-1591.

29. Yang Z., Wong W. S., Nielsen R., "Bayes empirical bayes inference of amino acid sites under positive selection", *Mol Biol Evol*, Vol. 22, no. 4, 2005, pp. 1107-1118.

30. Murrell B., Wertheim J. O., Moola S., Weighill T., Scheffler K.,

Kosakovsky Pond S. L., "Detecting individual sites subject to episodic diversifying selection", *PLoS Genet,* Vol. 8, no. 7, 2012, pp. e1002764.

31. Yang Z., Nielsen R., "Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages", *Mol Biol Evol,* Vol. 19, no. 6, 2002, pp. 908-917.

32. Zhang J., Nielsen R., Yang Z., "Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level", *Mol Biol Evol,* Vol. 22, no. 12, 2005, pp. 2472-2479.

33. Kosakovsky Pond S. L., Murrell B., Fourment M., Frost S. D., Delport W., Scheffler K., "A random effects branch-site model for detecting episodic diversifying selection", *Mol Biol Evol,* Vol. 28, no. 11, 2011, pp. 3033-3043.

34. Wernersson R., Pedersen A. G., "RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences", *Nucleic Acids Res,* Vol. 31, no. 13, 2003, pp. 3537-3539.

35. Kosakovsky Pond S. L., Posada D., Gravenor M. B., Woelk C. H., Frost S. D., "Automated phylogenetic detection of recombination using a genetic algorithm", *Mol Biol Evol,* Vol. 23, no. 10, 2006, pp. 1891-1901.

36. Pond S. L., Frost S. D., Muse S. V., "HyPhy: Hypothesis testing using phylogenies", *Bioinformatics,* Vol. 21, no. 5, 2005, pp. 676-679.

37. Kishino H., Hasegawa M., "Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea", *J Mol Evol,* Vol. 29, no. 2, 1989, pp. 170-179.

38. Kosakovsky Pond S. L., Frost S. D., "Not so different after all: A comparison of methods for detecting amino acid sites under selection", *Mol Biol Evol,* Vol. 22, no. 5, 2005, pp. 1208-1222.

39. Anisimova M., Bielawski J. P., Yang Z., "Accuracy and power of bayes prediction of amino acid sites under positive selection", *Mol Biol Evol,* Vol. 19, no. 6, 2002, pp. 950-958.

40. Yang Z., Nielsen R., "Synonymous and nonsynonymous rate variation in nuclear genes of mammals", *J Mol Evol*, Vol. 46, no. 4, 1998, pp. 409-418.

41. Anisimova M., Yang Z., "Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites", *Mol Biol Evol*, Vol. 24, no. 5, 2007, pp. 1219-1228.

42. Schymkowitz J., Borg J., Stricher F., Nys R., Rousseau F., Serrano L., "The FoldX web server: An online force field", *Nucleic Acids Res*, Vol. 33, no. Web Server issue, 2005, pp. W382-8.

43. Capriotti E., Fariselli P., Casadio R., "I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure", *Nucleic Acids Res*, Vol. 33, no. Web Server issue, 2005, pp. W306-10.

44. Buchan D. W., Ward S. M., Lobley A. E., Nugent T. C., Bryson K., Jones D. T., "Protein annotation and modelling servers at university college london", *Nucleic Acids Res*, Vol. 38, no. Web Server issue, 2010, pp. W563-8.

45. Xu D., Zhang Y., "Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field", *Proteins*, Vol. 80, no. 7, 2012, pp. 1715-1735.

46. Lyskov S., Gray J. J., "The RosettaDock server for local protein-protein docking", *Nucleic Acids Res*, Vol. 36, no. Web Server issue, 2008, pp. W233-8.

47. Comeau S. R., Gatchell D. W., Vajda S., Camacho C. J., "ClusPro: A fully automated algorithm for protein-protein docking", *Nucleic Acids Res*, Vol. 32, no. Web Server issue, 2004, pp. W96-9.

48. Wilson D. J., Hernandez R. D., Andolfatto P., Przeworski M., "A population genetics-phylogenetics approach to inferring natural selection in coding sequences", *PLoS Genet*, Vol. 7, no. 12, 2011, pp. e1002395.

49. Cereda M., Sironi M., Cavalleri M., Pozzoli U., "GeCo++: A C++

library for genomic features computation and annotation in the presence of variants", *Bioinformatics,* Vol. 27, no. 9, 2011, pp. 1313-1315.

50. Thornton K., "Libsequence: A C++ class library for evolutionary genetic analysis", *Bioinformatics,* Vol. 19, no. 17, 2003, pp. 2325-2327.

51. Fu Y. X., Li W. H., "Statistical tests of neutrality of mutations", *Genetics,* Vol. 133, no. 3, 1993, pp. 693-709.

52. Zeng K., Fu Y. X., Shi S., Wu C. I., "Statistical tests for detecting positive selection by utilizing high-frequency variants", *Genetics,* Vol. 174, no. 3, 2006, pp. 1431-1439.

53. Stephens M., Smith N. J., Donnelly P., "A new statistical method for haplotype reconstruction from population data", *Am J Hum Genet,* Vol. 68, no. 4, 2001, pp. 978-989.

54. Stephens M., Scheet P., "Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation", *Am J Hum Genet,* Vol. 76, no. 3, 2005, pp. 449-462.

55. Bandelt H. J., Forster P., Rohl A., "Median-joining networks for inferring intraspecific phylogenies", *Mol Biol Evol,* Vol. 16, no. 1, 1999, pp. 37-48.

56. Griffiths R. C., Tavare S., "Unrooted genealogical tree probabilities in the infinitely-many-sites model", *Math Biosci,* Vol. 127, no. 1, 1995, pp. 77-98.

57. Griffiths R. C., Tavare S., "Sampling theory for neutral alleles in a varying environment", *Philos Trans R Soc Lond B Biol Sci,* Vol. 344, no. 1310, 1994, pp. 403-410.

58. Evans P. D., Gilbert S. L., Mekel-Bobrov N., Vallender E. J., Anderson J. R., Vaez-Azizi L. M., Tishkoff S. A., Hudson R. R., Lahn B. T., "Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans", *Science,* Vol. 309, no. 5741, 2005, pp. 1717-1720.

59. Glazko G. V., Nei M., "Estimation of divergence times for major

lineages of primate species", *Mol Biol Evol*, Vol. 20, no. 3, 2003, pp. 424-434.

60. Hudson R. R., "Two-locus sampling distributions and their application", *Genetics*, Vol. 159, no. 4, 2001, pp. 1805-1817.

61. Wright S. I., Charlesworth B., "The HKA test revisited: A maximum-likelihood-ratio test of the standard neutral model", *Genetics*, Vol. 168, no. 2, 2004, pp. 1071-1076.

62. Li J. Z., Absher D. M., Tang H., Southwick A. M., Casto A. M., Ramachandran S., Cann H. M., Barsh G. S., Feldman M., Cavalli-Sforza L. L., et al, "Worldwide human relationships inferred from genome-wide patterns of variation", *Science*, Vol. 319, no. 5866, 2008, pp. 1100-1104.

63. Fumagalli M., Pozzoli U., Cagliani R., Comi G. P., Bresolin N., Clerici M., Sironi M., "Genome-wide identification of susceptibility alleles for viral infections through a population genetics approach", *PLoS Genet*, Vol. 6, no. 2, 2010, pp. e1000849.

64. Fumagalli M., Pozzoli U., Cagliani R., Comi G. P., Bresolin N., Clerici M., Sironi M., "The landscape of human genes involved in the immune response to parasitic worms", *BMC Evol Biol*, Vol. 10, 2010, pp. 264.

65. Pozzoli U., Fumagalli M., Cagliani R., Comi G. P., Bresolin N., Clerici M., Sironi M., "The role of protozoa-driven selection in shaping human genetic variability", *Trends Genet*, 2010,.

66. Fumagalli M., Pozzoli U., Cagliani R., Comi G. P., Riva S., Clerici M., Bresolin N., Sironi M., "Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions", *J Exp Med*, Vol. 206, no. 6, 2009, pp. 1395-1408.

67. 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, "A draft sequence of the neandertal genome", *Science*, Vol. 328, no. 5979, 2010, pp. 710-722.

Y., Viola B., Briggs A. W., Stenzel U., Johnson P. L., et al, "Genetic history of an archaic hominin group from denisova cave in siberia", *Nature,* Vol. 468, no. 7327, 2010, pp. 1053-1060.

69. Ward E. M., Stambach N. S., Drickamer K., Taylor M. E., "Polymorphisms in human langerin affect stability and sugar binding activity", *J Biol Chem,* Vol. 281, no. 22, 2006, pp. 15450-15456.

70. Seabra S. H., de Souza W., Damatta R. A., "Toxoplasma gondii exposes phosphatidylserine inducing a TGF-beta1 autocrine effect orchestrating macrophage evasion", *Biochem Biophys Res Commun,* Vol. 324, no. 2, 2004, pp. 744-752.

71. Mendes Wanderley J. L., Costa J. F., Borges V. M., Barcinski M., "Subversion of immunity by leishmania amazonensis parasites: Possible role of phosphatidylserine as a main regulator", *J Parasitol Res,* Vol. 2012, 2012, pp. 981686.

72. DeKruyff R. H., Bu X., Ballesteros A., Santiago C., Chim Y. L., Lee H. H., Karisola P., Pichavant M., Kaplan G. G., Umetsu D. T., et al, "T cell/transmembrane, ig, and mucin-3 allelic variants differentially recognize phosphatidylserine and mediate phagocytosis of apoptotic cells", *J Immunol,* Vol. 184, no. 4, 2010, pp. 1918-1930.

Manuscripts Published

1: Forni D, Pozzoli U, Cagliani R, Tresoldi C, Menozzi G,Riva S, Guerini FR, Comi GP, Bolognesi E, Bresolin N, Clerici M, Sironi M. **Genetic adaptation of the human circadian clock to day-length latitudinal variations and relevance for affective disorders.** *Genome Biology.* IN PRESS.

2: Mozzi A, Forni D, Cagliani R, Pozzoli U, Vertemara J, Bresolin N, Sironi M. Albuminoid genes: evolving at the interface of dispensability and selection. *Genome Biol Evol*. IN PRESS.

3: Sironi M, Biasin M, Gnudi F, Cagliani R, Saulle I, **Forni D**, Rainone V, Trabattoni D, Garziano M, Mazzotta F, Real LM, Rivero-Juarez A, Caruz A, Caputo SL, Clerici M. A **Regulatory Polymorphism in HAVCR2 Modulates Susceptibility to HIV-1 Infection**. *PLoS One*. 2014 Sep 2;9(9):e106442. doi:10.1371/journal.pone.0106442. eCollection 2014.

4: Sironi M, Biasin M, Cagliani R, Gnudi F, Saulle I, Ibba S, Filippi G, Yahyaei S, Tresoldi C, Riva S, Trabattoni D, De Gioia L, Lo Caputo S, Mazzotta F, Forni D, Pontremoli C, Pineda JA, Pozzoli U, Rivero-Juarez A, Caruz A, Clerici M. **Evolutionary analysis identifies an MX2 haplotype associated with natural resistance to HIV-1 infection**. *Mol Biol Evol*. 2014 Sep;31(9):2402-14. doi:10.1093/molbev/msu193. Epub 2014 Jun 14.

5: Guerini FR, Clerici M, Cagliani R, Malhotra S, Montalban X, Forni D, Agliardi C, Riva S, Caputo D, Galimberti D, Asselta R, Fenoglio C, Scarpini E, Comi GP, Bresolin N, Comabella M, Sironi M. **No association of IFI16 (interferon-inducible protein 16) variants with susceptibility to multiple sclerosis.** *J Neuroimmunol*.2014 Jun 15;271(1-2):49-52. doi: 10.1016/j.jneuroim.2014.04.006. Epub 2014 Apr 16.

6: Cagliani R, Forni D, Biasin M, Comabella M, Guerini FR, Riva S, Pozzoli U, Agliardi C, Caputo D, Malhotra S, Montalban X, Bresolin N, Clerici M, Sironi M. **Ancient and recent selective pressures shaped genetic diversity at AIM2-like nucleic acid sensors.** *Genome Biol Evol*. 2014 Apr;6(4):830-45. doi:10.1093/gbe/evu066.

7: Forni D, Cagliani R, Tresoldi C, Pozzoli U, De Gioia L, Filippi G, Riva S, Menozzi G, Colleoni M, Biasin M, Lo Caputo S, Mazzotta F, Comi GP, Bresolin N, Clerici M, Sironi M. **An evolutionary analysis of antigen processing and presentation across different timescales reveals pervasive selection.** *PLoS Genet*. 2014 Mar 27;10(3):e1004189. Doi: 10.1371/journal.pgen.1004189. ECollection 2014 Mar.

8: Forni D, Cleynen I, Ferrante M, Cassinotti A, Cagliani R, Ardizzone S, Vermeire S, Fichera M, Lombardini M, Maconi G, de Franchis R, Asselta R, BiasinM, Clerici M, Sironi M. **ABO histo-blood group might modulate predisposition to Crohn's disease and affect disease behavior.** *J Crohns Colitis*. 2014 Jun 1;8(6):489-94. doi: 10.1016/j.crohns.2013.10.014. Epub 2013 Nov 21.

9: Cagliani R, Forni D, Tresoldi C, Pozzoli U, Filippi G, Rainone V, De Gioia L, Clerici M, Sironi M. **RIG-I-like receptors evolved adaptively in mammals, with parallel evolution at LGP2 and RIG-I.** *J Mol Biol*. 2014 Mar 20;426(6):1351-65.doi: 10.1016/j.jmb.2013.10.040. Epub 2013 Nov 7.

10: Forni D, Cagliani R, Pozzoli U, Colleoni M, Riva S, Biasin M, Filippi G, De Gioia L, Gnudi F, Comi GP, Bresolin N, Clerici M, Sironi M. **A 175 million year history of T cell regulatory molecules reveals widespread selection, with adaptive evolution of disease alleles.** *Immunity*. 2013 Jun 27;38(6):1129-41. doi:10.1016/j.immuni.2013.04.008. Epub 2013 May 23.

11: Cagliani R, Forni D, Riva S, Pozzoli U, Colleoni M, Bresolin N, Clerici M, Sironi M. **Evolutionary analysis of the contact system indicates that kininogen evolved adaptively in mammals and in human populations.** *Mol Biol Evol.* 2013 Jun;30(6):1397-408. doi:10.1093/molbev/mst054. Epub 2013 Mar 16.

12: Al-Daghri NM, Clerici M, Al-Attas O, Forni D, Alokail MS, Alkharfy KM, Sabico S, Mohammed AK, Cagliani R, Sironi M. **A nonsense polymorphism (R392X) in TLR5 protects from obesity but predisposes to diabetes**. *J Immunol*. 2013 Apr1;190(7):3716-20. doi: 10.4049/jimmunol.1202936. Epub 2013 Mar 1.

13: Biasin M, Sironi M, Saulle I, de Luca M, la Rosa F, Cagliani R, Forni D, Agliardi C, lo Caputo S, Mazzotta F, Trabattoni D, Macias J, Pineda JA, Caruz A, Clerici M. **Endoplasmic reticulum aminopeptidase 2 haplotypes play a role in modulating susceptibility to HIV infection**. AIDS. 2013 Jul 17;27(11):1697-706.doi: 10.1097/QAD.0b013e3283601cee.

14: Cagliani R, Guerini FR, Rubio-Acero R, Baglio F, Forni D, Agliardi C, Griffanti L, Fumagalli M, Pozzoli U, Riva S, Calabrese E, Sikora M, Casals F, Comi GP, Bresolin N, Cáceres M, Clerici M, Sironi M. **Long-standing balancing selection in the THBS4 gene: influence on sex-specific brain expression and gray matter volumes in Alzheimer disease.** *Hum Mutat.* 2013 May;34(5):743-53. doi:10.1002/humu.22301. Epub 2013 Apr 2.

15: Cagliani R, Pozzoli U, Forni D, Cassinotti A, Fumagalli M, Giani M, Fichera M, Lombardini M, Ardizzone S, Asselta R, de Franchis R, Riva S, Biasin M, Comi GP, Bresolin N, Clerici M, Sironi M. **Crohn's disease loci are common targets of protozoa-driven selection**. *Mol Biol Evol*. 2013 May;30(5):1077-87. doi:10.1093/molbev/mst020. Epub 2013 Feb 6.

16: Al-Daghri NM, Cagliani R, Forni D, Alokail MS, Pozzoli U, Alkharfy KM, Sabico S, Clerici M, Sironi M. **Mammalian NPC1 genes may undergo positive selection and human polymorphisms associate with type 2 diabetes.** *BMC Med*. 2012 Nov 15;10:140. doi: 10.1186/1741-7015-10-140.

17: Fumagalli M, Fracassetti M, Cagliani R, Forni D, Pozzoli U, Comi GP, Marini F, Bresolin N, Clerici M, Sironi M. **An evolutionary history of the selectin gene cluster in humans**. *Heredity* (Edinb). 2012 Aug;109(2):117-26. doi:10.1038/hdy.2012.20. Epub 2012 May 2.

18: Sironi M, Biasin M, Forni D, Cagliani R, De Luca M, Saulle I, Caputo SL,Mazzotta F, Macias J, Pineda JA, Caruz A, Clerici M. **Genetic variability at the TREX1 locus is not associated with natural resistance to HIV-1 infection**. *AIDS*.2012 Jul 17;26(11):1443-5. doi:10.1097/QAD.0b013e328354b3c2.

19: Cagliani R, Guerini FR, Fumagalli M, Riva S, Agliardi C, Galimberti D, Pozzoli U, Goris A, Dubois B, Fenoglio C, Forni D, Sanna S, Zara I, Pitzalis M, Zoledziewska M, Cucca F, Marini F, Comi GP, Scarpini E, Bresolin N, Clerici M, Sironi M. **A trans-specific polymorphism in ZC3HAV1 is maintained by long-standing balancing selection and may confer susceptibility to multiple sclerosis**. *Mol Biol Evol.* 2012 Jun;29(6):1599-613. doi: 10.1093/molbev/mss002. Epub 2012 Jan 6.

20: Guerini FR, Cagliani R, Forni D, Agliardi C, Caputo D, Cassinotti A, Galimberti D, Fenoglio C, Biasin M, Asselta R, Scarpini E, Comi GP, Bresolin N, Clerici M, Sironi M. **A functional variant in ERAP1 predisposes to multiple sclerosis**. *PLoS One.* 2012;7(1):e29931. doi: 10.1371/journal.pone.0029931. Epub 2012 Jan 12.

21: Cagliani R, Riva S, Marino C, Fumagalli M, D'Angelo MG, Riva V, Comi GP, Pozzoli U, Forni D, Cáceres M, Bresolin N, Clerici M, Sironi M. **Variants in SNAP25 are targets of natural selection and influence verbal performances in women**. *Cell Mol Life Sci.* 2012 May;69(10):1705-15. doi: 10.1007/s00018-011-0896-y. Epub 2011 Dec 23.

22: Sironi M, Biasin M, Cagliani R, Forni D, De Luca M, Saulle I, Lo Caputo S, Mazzotta F, Macías J, Pineda JA, Caruz A, Clerici M. **A common polymorphism in TLR3 confers natural resistance to HIV-1 infection.** *J Immunol.* 2012 Jan 15;188(2):818-23. doi: 10.4049/jimmunol.1102179. Epub 2011 Dec 14.