UNIVERSITÀ DEGLI STUDI DI MILANO

Facoltà di Medicina e Chirurgia

Corso di Dottorato di Ricerca in Medicina molecolare

MED/26 Ciclo XXIII

POPULATION GENETIC APPROACHES FOR THE STUDY OF COMPLEX TRAITS: FOCUS ON INFECTIOUS AND AUTOIMMUNE DISEASES

Dott.ssa:

Manuela Sironi

Matr. n°R07736

Tutore: Prof. Giacomo P. Comi

Coordinatore: Prof.ssa Maria Luisa Villa

Anno accademico
2009/2010

ABSTRACT

It is commonly believed that infectious diseases have represented one of the major threats to human populations and have therefore acted as a powerful selective force. Therefore, several human genes have evolved in response to infectious agents. Indeed, it has been suggested that human populations may have adapted to pathogens to such a degree that the lower exposure to infectious agents in modern developed societies results in immune imbalances, with autoimmune and allergic conditions being the outcome (hygiene hypothesis). Quite obviously, the presence of a functional variant is a prerequisite for selection to act, and the identification of non-neutrally evolving genes has been regarded as a strategy complementary to classical clinical and epidemiological studies to provide insight into the mechanisms of host defense. Similarly, analysis of the evolutionary history of genes involved in immune defense might provide novel insights into the delicate balance between efficient response to pathogens and autoimmune/allergic manifestations. In this work population genetics approaches are applied to study the evolutionary patterns of specific groups of genes such as those encoding blood group antigens and interleukins/interleukin receptors. Several of these latter genes have evolved in response to parasitic worms, but a subset of disease alleles for inflammatory bowel disease and celiac disease have increased in frequency in response to non-helminthic pathogens (i.e. viruses and bacteria). At the genome-wide level, the identification of selective signatures was exploited to identify novel susceptibility variants for virus-, protozan-, and helminth-borne infections. These analyses allowed the identification of several variants that may modulate infection susceptibility and we noticed a partial overlap between genes involved in the response to helminths and those carrying susceptibility alleles for asthma/atopy; similarly, a number of genes subjected to virus-driven selective pressure have been involved in the pathogenesis of multiple sclerosis and type 1 diabetes. One of these, namely *IFIH1*, was studied in detail: we revealed a complex selective pattern in human populations distributed in different geographic locations. Nonetheless, the analysis of *IFIH1* variants involved in the susceptibility to type 1 diabetes indicated that they have evolved neutrally. Finally, we show that the identification of gene regions subjected to natural selection can provide information on the location of functional variants and these, in turn, may be regarded as strong candidates to prioritize on in case/control association studies. In the case of *ERAP2* we carried out one such study and verified that a nonsynonymous variant subjected to natural selection affects the natural resistance to HIV-1 infection. In summary we show that selective events leave a signature on human genes that can be detected using population genetics approaches and exploited for the identification of variants that influence complex phenotypic traits such as susceptibility to infections. These studies can also shed light on the relationship between past selective events and predisposition to common diseases in modern populations.

SOMMARIO

E’ opinione comune che le malattie infettive abbiano rappresentato la principale minaccia per le popolazioni umane e che abbiano esercitato una potente pressione selettiva. Di conseguenza alcuni geni umani si sono evoluti in risposta a agenti infettivi. E’ stato anche suggerito che gli esseri umani si siano adattati ai loro patogeni a un livello tale per cui la minore esposizione a agenti infettivi nelle società industrializzate creerebbe un disequilibrio del sistema immunitario che risulterebbe poi nella predisposizione a malattie autoimmuni e allergie (ipotesi dell’igiene). Ovviamente, la presenza di una variante funzionale è necessaria perché la selezione possa agire e l’identificazione di geni che evolvono in modo non neutrale è considerata come un approccio complementare ai classici studi clinici e epidemiologici per chiarire i meccanismi di risposta dell’ospite. Similmente, l’analisi dal punto di vista evolutivo di geni coinvolti nella risposta immunitaria potrebbe fornire nuove conoscenze riguardo ai meccanismi che mantengono un equilibrio tra l’efficiente risposta ai patogeni e le manifestazioni allergiche o autoimmuni. In questo lavoro, abbiamo applicato approcci di genetica di popolazione allo studio del pattern evolutivo di gruppi specifici di geni come, ad esempio, quelli che codificano per gli antigeni dei gruppi sanguigni o per interleuchine e loro recettori. Molti di questi ultimi geni si sono evoluti in risposta ai vermi parassiti, ma una parte di alleli di suscettibilità per malattia celiaca e morbo di Crohn/colite ulcerosa sono aumentati in frequenza in risposta a patogeni quali virus e batteri. A livello di intero genoma, abbiamo sfruttato l’identificazione di tracce di selezione per identificare nuove varianti di suscettibilità per malattie trasmesse da virus, protozoi e elminti. Abbiamo notato una parziale sovrapposizione tra geni coinvolti nella risposta a elminti e quelli che portano varianti di suscettibilità per asma e atopia; similmente, un certo numero di geni che sono stati sottoposti a pressione selettiva esercitata da virus sono stati coinvolti nella patogenesi della sclerosi multipla e di diabete di tipo 1. Uno di questi, *IFIH1*, è stato studiato in dettaglio: abbiamo messo in luce un pattern evolutivo complesso in diverse popolazioni umane distribuite in diverse aree geografiche. Tuttavia, l’analisi delle varianti di suscettibilità per diabete di tipo 1 in *IFIH1* ha indicato che esse evolvono in modo neutrale. Infine, abbiamo dimostrato come l’identificazione di regioni geniche sottoposte a selezione naturale possa fornire indicazioni sulla localizzazione di varianti funzionali e queste, a loro volta, possono essere considerate ottimi candidati per studi di associazione caso/controllo. Nel caso di *ERAP2*, abbiamo appunto svolto uno studio caso/controllo e abbiamo verificato che una variante nonsinonima e sottoposta a selezione naturale altera la suscettibilità all’infezione da HIV-1. In conclusione, abbiamo mostrato che gli eventi selettivi lasciano un’impronta sui geni che può essere identificata attraverso analisi di genetica di popolazione e sfruttata per la scoperta di varianti che influenzano tratti fenotipici complessi come la suscettibilità alle infezioni. Questi studi possono anche fare luce sulla relazione tra eventi selettivi e predisposizione a malattie comuni nelle popolazioni moderne.

INDEX

SYMBOL LIST

SNP: single nucleotide polymorphisms
MAF: minor allele frequency
KY: kiloyears
MY: million years
STR: short tandem repeat
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
EHH: extended haplotype homozygosity
T: Kendall's rank correlation coefficient

1. INTRODUCTION

1.1 Human genetic variability and human adaptation

The complex patterns of genetic diversity in human populations are the product of many levels of demographic and evolutionary events acting on different time-scales, including colonizations, migrations, population expansions, mutation, genetic drift and selection.

Several recent studies have indicated that human demographic history (i.e. the pattern of human migrations throughout the world with subsequent expansions and bottlenecks) explains a large part of neutral genetic diversity among populations. Studies of neutral short tandem repeat (STR) loci from the HGDP-CEPH panel (Human Genome Diversity Panel-Centre d' Étude du Polymorphisme Humain [1]) showed that there is indeed a strong relationship between geography and various measures of genetic diversity at the worldwide scale. Geographic distances between populations (calculated along landmasses so as to possibly reflect out-of-Africa migratory routes) predict the respective genetic differentiation (Fig. 1a) and geographic distances from East Africa show a high negative correlation with measures of within-population diversity [2] (Fig.1b).

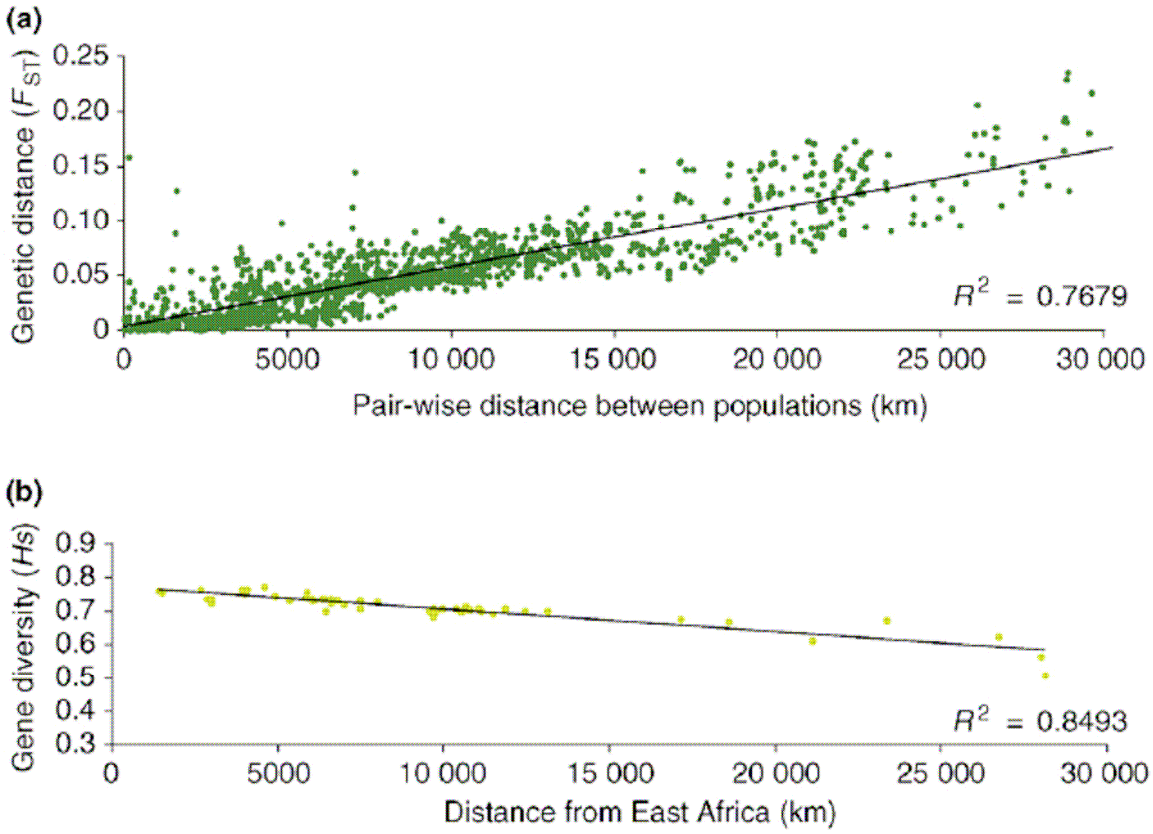


Figure 1. (a) Plot of pairwise genetic distance between populations in the HGDP-CEPH panel against pairwise geographic distance. (b) Gene diversity

within the HGDP-CEPH populations is plotted againstgeographic distance from East Africa. Figure taken from [2]

More recently, similar findings have been reported using a large number of SNPs genotyped in 52 populations distributed worldwide (Human Genetic Diversity Panel, HGDP-CEPH) [3].
These studies strongly support an Africa origin for modern humans. Indeed, genetic data indicate that the alleles found outside Africa are often a subset of the African allele pool [4-5], and that continent-specific alleles or haplotypes are rare in general, but are far more common in Africa than in any other continent [4-5]. An analysis of HGDP-CEPH populations supported a serial founder model, in which non-African populations form a sequential chain of colonies [3].
These observations indicate that human genetic diversity has largely been shaped by phenomena occurring in geographic space, and, therefore, that these effects must be accounted for in order to identify loci that have been involved in adaptation. In order to demonstrate genetic adaptation in humans, or in any organism, it is necessary to acquire evidence substantiating that natural selection underlies the evolution of a particular trait. Because demography affects all genes equally, whereas selection acts upon specific genome regions, it is usually implied that genes/gene regions deviating from the general distribution of genetic variation may represent targets of natural selection.

Given this premise, it is clear enough that, along their evolutionary history, humans, as all other living organisms, have adapted to their environment. Specifically, several environmental factors have possibly represented strong selective pressures for human populations but major effects are generally ascribed to pathogens, climate and diet.

1.1.1 Adaptation to pathogens

This is extensively discussed in the hygiene hypothesis review included below (1.4).

1.1.2 Climatic adaptation

Humans have spent the longest part of their evolutionary history in Africa, that is in a hot climate, with low availability of salt and with high UV irradiation. Migration out of Africa (Fig. 2) exposed ancestral populations to colder environments, with less incident sunlight.

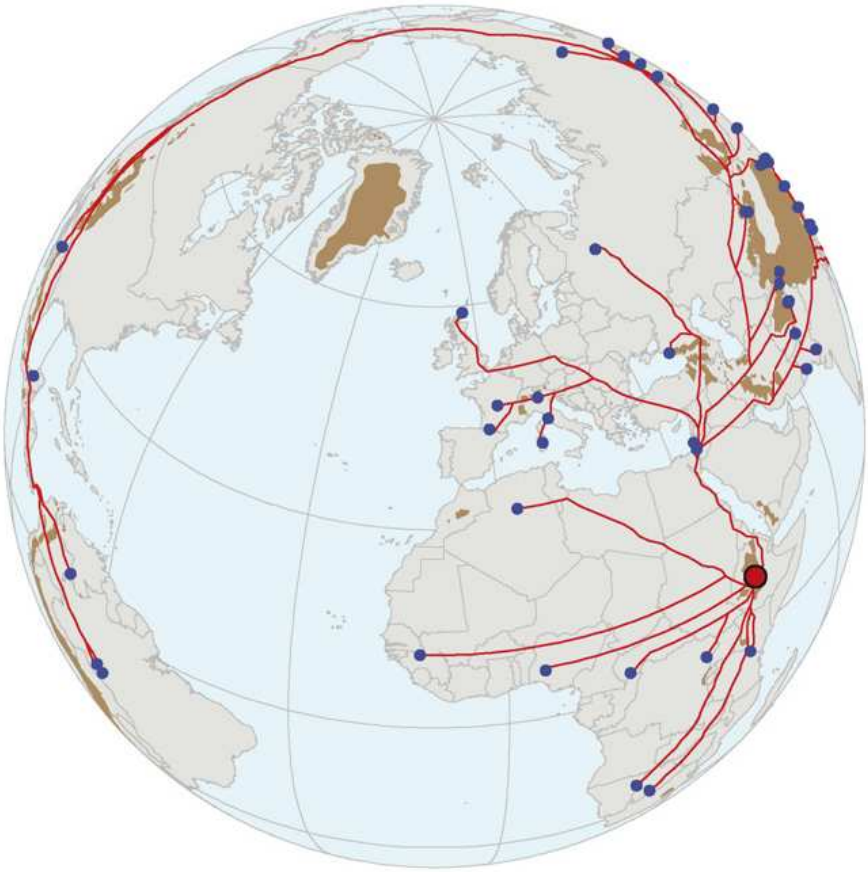


Figure 2. Likely colonization routes (red lines) between populations in the HGDP-CEPH panel (blue spots) assuming an origin of modern humans in East Africa (Addis Ababa, red spot). The map was calculated by by forcing migrations through landmasses and avoiding altitudes over 2000 m. Figure taken from [2]

The most obvious adaptive response to these novel environments is in skin pigmentation, which is basically due to the quantity, type, and distribution of melanin. Skin colour is strongly geographically differentiated, with darker-skinned populations concentrated in the tropics, and lighter-skinned populations in more northerly latitudes. Dark skins in topical climates protects against UV-induced damage, while where sunlight is low, depigmentation may be favoured because UV penetration is necessary for vitamin D synthesis [6-7]. A balance between these factors can largely explain the global pattern of pigmentation [6-7].

Residence of ancestral populations in tropical Africa also necessitated heat adaptation, including cooling through efficient sweating. The considerable salt loss, combined with low dietary salt intake, possibly led to selection for salt retention; at the same time, there was likely selection for increased arterial tone and cardiac contraction force when blood volume was depleted by water loss (reviewed in [8]). Initially, heat-adapted African populations expanded northward, undergoing selection for cold adaptation. Subsequently, cold-adapted north Asians expanded

southward into the Americas less than 20 KY ago, undergoing selection for heat adaptation, so that Native Americans show similar salt retention and cardiovascular phenotypes to Africans living at the same latitudes.

1.1.3 Dietary adaptation

Humans have spent the longest part of their evolutionary history in Africa, that is in In terms of influence on diet, the most important development in human prehistory was agriculture, beginning around 10 KY ago in the Middle East. Agriculture is likely to have changed substantially the diet of early humans who had mainly relied on a hunter–gatherer lifestyle. Agriculture increased the intake of starch and sugar. A common view is that this change happened in the context of a genetic background adapted to a lifestyle of feast and famine by favouring the maintenance of “thrifty alleles”, and that the arrival of agriculture lead to high incidences of type 2 diabetes (reviewed in [8]) (see also par. 1.3). With agriculture pastoralism also begun, resulting in the availability of milk as a source of energy. In this respect the evolutionary history of *LCT* alleles permitting persistence of lactase expression into adulthood is emblematic.

Lactose tolerance is extremely high in North Europe, for example, and in general the frequency of a *LCT* persistence allele in Europe correlates well with that of populations with a history of cattle domestication and milk drinking [9]. In the HapMap samples, *LCT* in Europeans shows one of the strongest signals of positive selection [10], reflecting a powerful advantage that may be related to milk as a source of uninfected water rather than as a source of nutrition. Conversely, lactase-persistence is relatively rare in Africa and lactase-persistent populations in this continent do not carry the same *LCT* allele described in Europeans. Studies of Tanzanians, Kenyans and Sudanese [11] revealed that different *LCT* variants cause lactase persistence and the three African variants arose independently of each other and of the European variant. Together with skin pigmentation, this represents one of the best known examples of convergent evolution in humans.

1.2 Natural selection: molecular signatures

The detection of natural selection from genetic data is not simple. One major difficulty is disentangling signals of selection from the signal left by the demographic history of our species. Helpful in this regard has been the increasing availability of large genome-wide data sets of human DNA polymorphism and resequenced gene regions that allow inferences about our demographic history. It is expected that demographic phenomena such as population expansions,subdivision and bottlenecks will affect variation at all genes. On the other hand, natural selection is a locus-specific force.
The neutral model of molecular evolution postulates that most evolutionary change is a consequence of random genetic drift, rather than adaptive evolution. It should be noted, however, that the neutral model does not exclude the role of natural selection: natural selection is assumed to remove deleterious mutations (i.e. purifying selection) and fix the rarely arising advantageous mutation [12]. Also, natural selection maintains fluctuating polymorphisms if these are favourable under different environmental conditions or as a result of overdominant selection (heterozygote advantage). Under the neutral model, specific theoretical predictions

can be made regarding the relationship between the rate of mutation and evolutionary parameters: (i) polymorphism within a species is a function of the mutation rate and population size; (ii) the rate at which mutational differences accumulate as two species diverge (i.e. the substitution rate) is the same as the rate at which neutral mutations arise [13]; (iii) the expected frequency of alleles in a sample is a function of the population size [14]. Statistical tests designed to detect natural selection take advantage from the relative ease with which predictions made by the neutral theory can be verified against empirical data.
Therefore, the study of natural selection using genetic data is based upon tests of the null hypothesis of neutrality, rather than tests of natural selection. Thus, the tests employed in the study of natural selection are more adequately defined as tests of the null hypothesis of neutrality, or neutrality tests.

Different forms of selection shape genetic variation within and between species. Positive selection refers to the cases in which a novel DNA variant has a selective advantage over others, and consequently rises in frequency. Negative (or purifying selection) refers to the cases in which novel DNA variants have a selective disadvantage with respect to others, and tend to remain at low frequencies or be removed. Balancing selection refers to selective regimes that increase genetic variation within a species.
Natural selection is expected to directly affect the genetic variants that alter an individual's survival probability. However, due to linkage disequilibrium (LD), the effects of selection may not be restricted to the causal variant associated with the selection target and, depending on different factors, the consequences of natural selection can extend over very small (e.g. long-standing balancing selection) or very long (e.g. recent selective sweeps) genomic regions surrounding the selection target.

1.2.1 Tests based on the frequency of polymorphisms

As mentioned above, under neutrality, genetic variation accumulates in a population as a function of the population size and mutation rate. Different forms of selection will impact the frequency spectrum in distinct ways. If a mutation is favoured by natural selection (i.e. positive selection) it will rise to high frequency in a population carrying along all linked neutral variants. This process is known as a selective sweep and results in an overall decrease in the genetic variation at the selected site, as well as at sites linked to it [15]. (Fig. 3). The mutational process will introduce new mutations after the selective sweep. These novel DNA variants will initially be present at low frequencies. Thus, shortly after a selective sweep, we expect to observe a large proportion of low frequency variants in a sample. Moreover, derived alleles (which usually display lower frequency compared to ancestral alleles) linked to the selected variant will also rise in frequency. Therefore, another signature of a selective sweep in an excess of high-frequency derived alleles (reviewed in [16]).

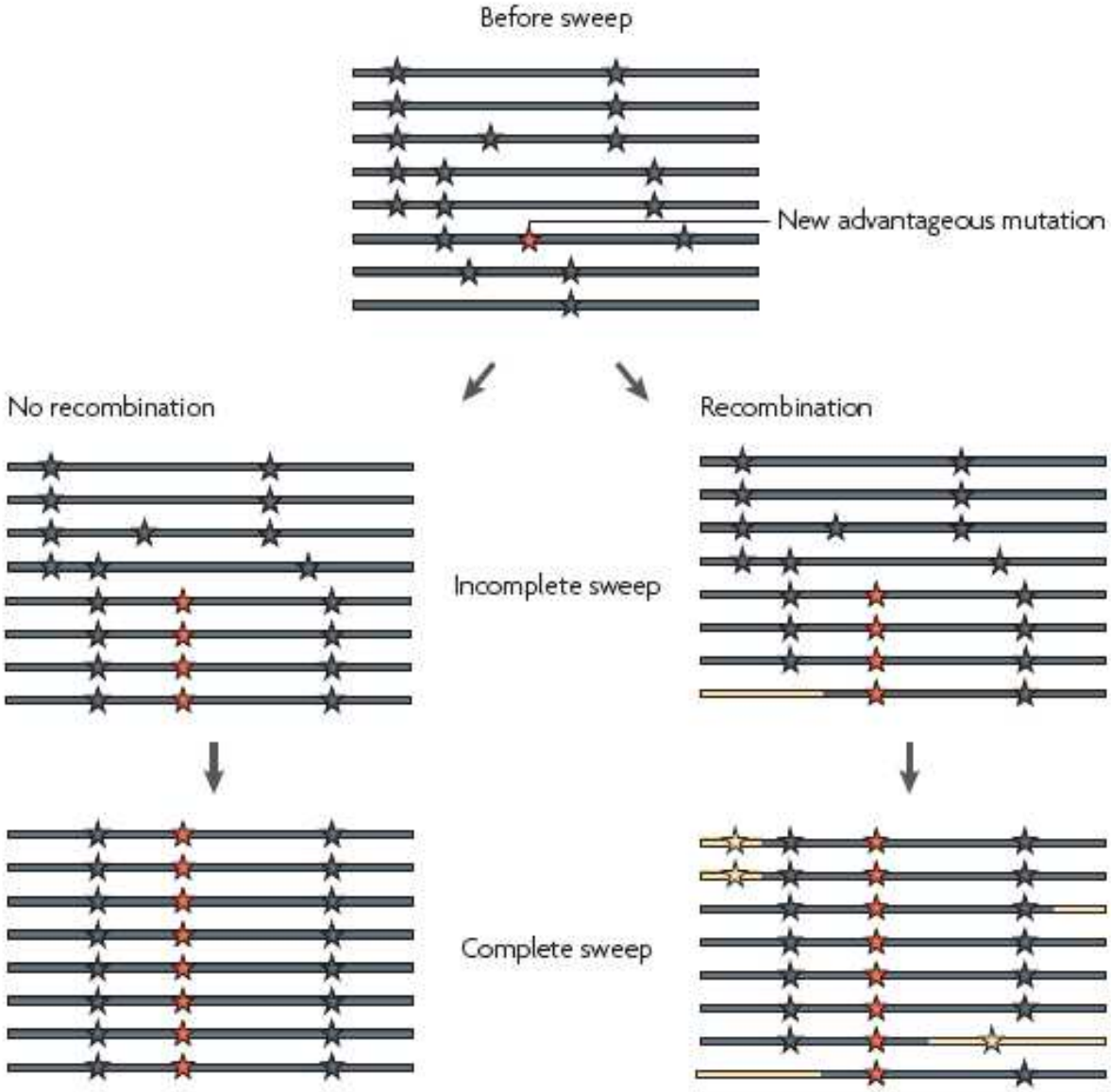


Figure 3. Selective sweep. Haplotypes are indicated as horizontal lines; derived SNP alleles are depicted as stars. A new advantageous mutation is represented with a red star and initially appears on one haplotype. In the absence of recombination (left): neutral SNP alleles on the chromosome as the advantageous mutation will increase in frequency (incomplete sweep) and eventually be fixed (complete sweep) with the selected variant. Conversely, alleles that do not occur on this chromosome will be lost, so that variability will be severly reduced after the sweep. With recombination (right): haplotypes can emerge through recombination, allowing some of the neutral mutations that are linked to the advantageous mutation to segregate after a completed selective sweep. As the rate of recombination depends on the physical distance among sites, the effect of a selective sweep on variation in the genomic regions around it diminishes with distance from the site that is under selection. Chromosomal segments that are linked to advantageous mutations through recombination during the selective sweep are coloured yellow. Figure taken from [16]

Purifying selection is also expected to result in an increase in the proportion of low-frequency variants. This can be understood as a consequence of the fact that novel mutations that enter the population generally remain at low frequencies, because of their deleterious effects. Thus, in the case of purifying selection the frequency of derived alleles is expected to be very low. In contrast, balancing selection will increase the proportion of variants at intermediate frequencies, since this selection regime favours the maintenance of variation of multiple alleles. Moreover, under balancing selection linked variants will be maintained together with the selected alleles, resulting in an excess of nucleotide diversity in the region (Fig.4) (reviewed in [17]).

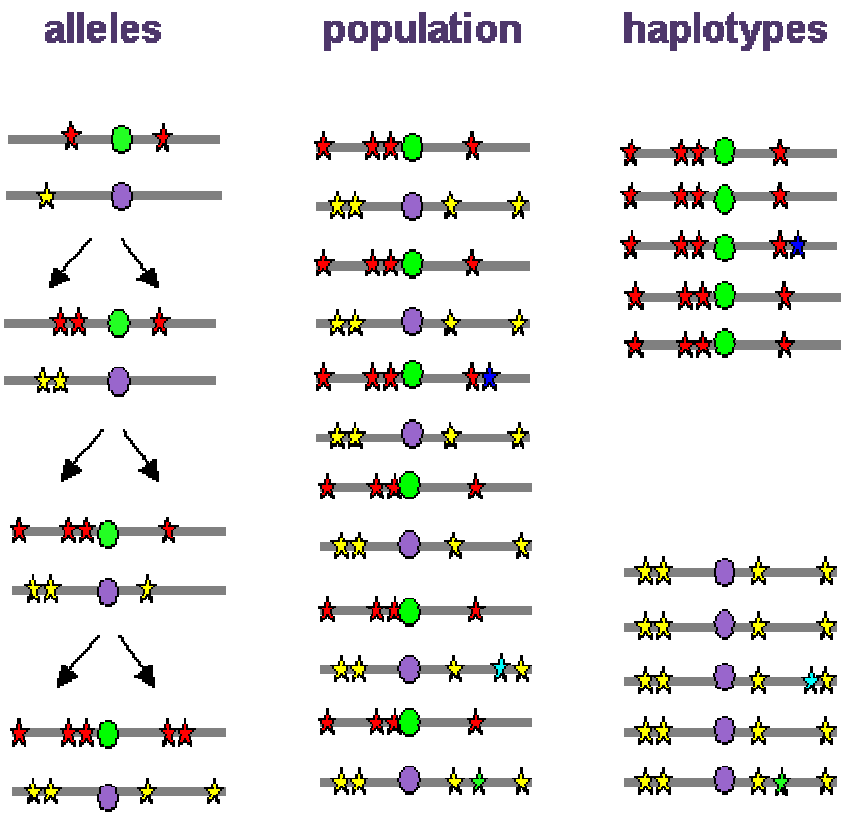


Figure 4. Balancing selection. Haplotypes are indicated as horizontal lines; neutral SNP alleles are depicted as stars. A biallelic balanced polymorphism is shown with the violet and green circles (alleles). The two balanced alleles are located on different haplotypes carrying neutral variants (red and yellow stars). Neutral SNP alleles are maintained with the two selected alleles together with new mutations that arise over time. Since polymorphisms, due to their linkage with the selected alleles, tend to be maintained longer than expected under neutrality, the effect is higher genetic diversity in the region carrying the balanced alleles (higher density of polymorphisms). If the two balanced alleles are maintained at an intermediate frequency, all

neutral variation will also remain at intermediate frequency (excess of intermediate-frequency alleles). Moreover, after a certain amount of time, haplotypes carrying the balanced alleles will accumulate different neutral alleles. The result is highly differentiated haplotype lineages (or haplotype clades).

Different quantitative tests have been developed to interpret whether the frequency spectrum of a population sample reveals the action of one or another of these forms of selection. The most widely used test that explores the frequency spectrum was proposed by Tajima [18]. This test is based on the comparison of two measures of the neutral parameter θ_W, an estimate of the expected per site heterozigosity [19], and π [20], the average number of pairwise sequence nucleotide differences [$D= (\pi- \theta_W)/sd(\pi- \theta_W)$]. Since π depends on the frequency of sampled alleles, D will be negative when an excess of rare alleles is observed and positive when many intermediate alleles occur. As described above, a selective sweep can result in a large proportion of low-frequency variants. This will result in negative D values. Similarly, under purifying selection (i.e. negative selection), we have an excess of low-frequency variants, resulting in a low value of Tajima's D. Conversely, balancing selection favours the maintenance of different alleles in the population, resulting in an excess of intermediate frequency alleles, and thus Tajima's $D > 0$.

Fay and Wu [21] proposed a test of the frequency spectrum that offers a solution to the challenge of distinguishing among different selective processes that result in negative values of Tajima's D (i.e. purifying selection or positive selection). These authors exploited the fact that when positive selection takes place, it can drive derived alleles, found at nearby locations on the chromosome, to high frequencies. Fay and Wu [21] therefore proposed a test based on contrasting two diversity estimates; an estimate based on the nucleotide diversity (π), and a measure of diversity that is sensitive to high-frequency derived mutations (θ_H). Under neutral-equilibrium conditions, H has an expected value of zero. When a recent selective sweep has taken place the excess of high-frequency derived mutations results in a negative value for H [21].

Fu and Li [22] also developed neutrality tests based on the allele frequency spectrum. These tests are conceptually similar to Tajima's D but they also take into account whether mutations occur in external (new mutations) or internal (old mutations) branches of a gene genealogy.

These tests are powerful but suffer from the complication that demographic events also result in deviations from the site frequency spectrum. For example a population bottelneck is also expected to result in an increase of intermediate frequency allele (as rare alleles are more easily lost during the shrinkage phase), while population expansion may result in a higher proportion of low frequency variants. Taking these considerations into account is extremely important when human diversity data are being analysed, as different human populations are known to have experienced diverse demographic scenarios. Therefore, calculation of the statistical significance of neutrality tests is not trivial and two main

approaches are currently applied. The first is based on the empirical comparison with diversity data calculated for a large number of genomic regions analysed in the same population samples as the locus being analysed. The basis of this approach is that demography affects all loci equally, while selection is a locus-specific force. Therefore, the test statistic (e.g. Tajima' s D) is calculated for the gene region of interest and for a set of reference regions and significance is calculated by attributing a percentile rank in the distribution of reference loci (i.e. significant results are obtained for D values below the 5th or above the 95th percentile in the distribution). The second approach is based on coalescent simulations. The coalescent is a process in which, looking backward in time, the genealogies of alleles merge at a common ancestor. Therefore, a large number of coalescent trees can be simulated (e.g. 10000) and the statistic of interest (D) calculated for each of the 10,000 simulations. The proportion of simulations with D < (or >) of the observed value is the *p*-value with precision to 1/(# simulations). Recently, dedicated programs have been developed that allow the incorporation of demographic scenarios in coalescent simulations so that the resulting p values are corrected against demographic events. One such program is *cosi* [23] that incorporates a demographic model (Fig. 5) for major human populations.

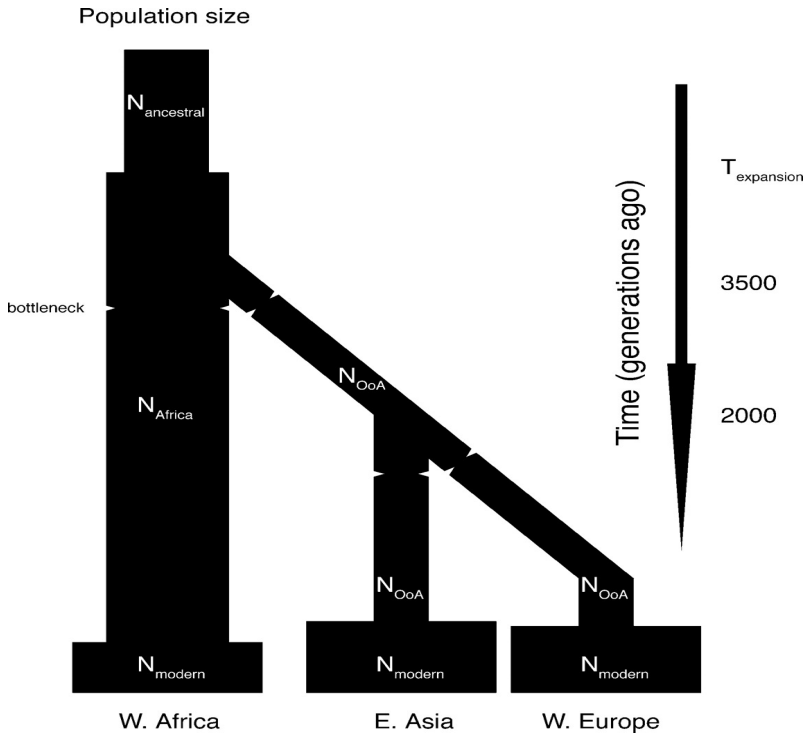


Figure 5. Demographic model for major human populations obtained by fitting empirical data. $N_{ancestral}$: ancestral population size; N_{Africa}: African population size; N_{OoA}: non-African population size; $T_{expansion}$: Time of ancestral population expansion. Bottlenecks are indicated by constrictions. Figure taken from [23]

1.2.2 Tests that confront different classes of changes.

Under neutrality, the level of intra-specific variation and that of inter-specific diversity are expected to be proportional to the neutral mutation rate. Therefore, one possibility to detect natural selection is to test whether this expectation is verified. In the HKA (Hudson-Kreitman-Aguadè) test [24], the assumption is made that under neutrality the ratio of polymorphism to divergence will be the same for two or more genes. In a classic HKA test, a gene of interest is compared to a putatively neutral locus, and differences in the ratio of polymorphism to divergence between these is taken as evidence of selection in the gene of interest. More recently, statistical frameworks have been developed that exploit the availability of genetic divergence and variability data on several genes to apply multi-locus HKA tests [25]. In these cases one gene of interest is compared to several genes, so that the results are more robust to chance variation.

1.2.3 Tests based on population subdivision

Another approach for detecting natural selection is to study between-population differences in allele frequencies. Under neutrality population genetic difference is mainly driven by demographic processes and genetic drift. Positive selection may increase levels of genetic differentiation among populations for two reasons. First, selection might act locally and be related to adaptations to the local environment (reviewed in [16]). Second, selective sweeps acting on mutations that arise in specific geographical regions might cause increased levels of population subdivision during the period of time in which the mutation is still increasing in frequency. Even if the mutation is beneficial in all environments, the fact that the mutation arose in a particular geographical location might temporarily increase levels of population differentiation in the genomic region that is affected by selection [16].

Conversely, balancing selection may result in lower levels of population differentiation, especially when the selected allele(s) is old (long-standing balancing selection). This is due to the fact the selected allele(s) and all linked variants tend to be maintained at appreciable levels in different populations. Yet, cases of balancing selection restricted to one specific geographic location may result in an effect similar to that of positive selection and therefore increases population genetic differentiation.

The most commonly used statistical measure of population differentiation was devised by Wright [26] and is known as the fixation index (F_{ST}). The main difficulty in using F_{ST} for the detection of natural selection is that, as mentioned above, differentiation among populations is sensitive to a variety of demographic factors (including the rate of drift and the extent of gene flow among them), making it difficult to rule out demographic scenarios that could account for the observed high or low F_{ST} values. Again, one possibility is to take advantage from the large number of genetic loci for which we have information of population differentiation to create an empirical F_{ST} distribution. Thus, rather than statistically testing specific loci, we can use their position relative to this distribution to gain insights about their possible selective histories. Another approach is to use computer simulations under realistic demographic scenarios (inferred from multilocus studies) to obtain the distribution of F_{ST} values under neutrality.

1.2.4 Tests based on linkage disequilibrium

An influential approach for detecting recent and strong natural selection is the extended haplotype test and its derivatives [10, 27]. The extended haplotype test relies on the linkage-disequilibrium structure of local regions of the genome and is based on the observation that when a mutation is under positive selection, it will rise in frequency quickly. If the rise in frequency of the favoured mutation occurs comparatively faster than the rate of recombination, an extended region around the selected site (including all DNA variants that may be present within this region) will also rise in frequency, creating 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 method identifies tracts of homozygosity within a 'core' haplotype, using the 'extended haplotype homozygosity' (EHH) as a statistic. A relative EHH (rEHH) is calculated by comparing the EHH of the core haplotype to the EHH of all other haplotypes in the region. In the version by Voight et al. [10], the EHH is summed over all sites away from a core SNP, and compared between the haplotypes that carry the ancestral and the derived allele in the SNP. The statistic iHS (integrated haplotype score) is then normalized to have a mean of 0 and variance of 1.

1.2.5 Tests based on geographic-explicit models

Another possibility to identify genes subjected to natural selection is to search for variants that display strongly differentiated allele frequencies among populations that live in different environments (i.e. allele frequencies that correlate with a specific environmental variable). Such correlations can arise when selective pressures exerted by the environmental variable are sufficiently divergent between geographic locations, such that differences in allele frequency can be maintained irrespective of gene flow. Some examples in humans include loci involved in adaptation to climate [28, 29], diet/subsistence [30], and pathogens [31]. The advent of genome-wide data sets with individuals from many populations, across a wide geographic range [3], allows investigators to search for correlations at the genome-wide and worldwide level**.**

A clear caveat in these studies is that, as mentioned above, a major contribution of genetic diversity and variation across geographic locations is accounted for by demography rather than selection. Therefore, it is important to disentangle neutral from selective effect when selecting candidate loci. Another possible problem with this approach is that environmental factors may change over evolutionary times, and in most cases we have a measure that reflects the contemporary status of the variable. Therefore, the implicit assumption is that environmental variables (i.e. selective pressures) have remained relatively constant along human history.

1.3 Evolutionary frameworks for common diseases

These models are substantially based on the observation whereby the environment in which we now live is radically different from the one that ancestral human populations adapted to. A number of transitions may have have resulted in shifts in the selective pressures acting on the biological processes that, in contemporary populations, underlie major classes of common diseases. Crucial turning points

include the migration of modern humans out of Africa, the shift from a hunting and gathering life-style to a subsistence based on agriculture and pastoralism, and the creation of ever larger, connected communities that allowed the spread of several pathogen species unknown to small, nomadic populations.

The impact of these selective shifts on susceptibility genes for many common diseases has been formalized in a series of hypotheses, which are based on epidemiological trends and the disease pathophysiology. Perhaps the most popular is the “hygiene hypothesis”, which is described below. Another framework for common diseases is the ‘thrifty genotype’ hypothesis first proposed in 1962 to explain the high prevalence of type 2 diabetes and obesity [32]. This hypothesis posits that since ancestral hunter-gatherer populations underwent seasonal cycles of feast and famine, they would have benefited from being ‘thrifty’ (i.e. having extremely efficient fat and carbohydrate storage). With changes in subsistence strategies, this ‘thriftiness’ became detrimental as food availability in general increased and the diet shifted to products with higher starch an sugar contents. As an example, Vander Molen et al. [33] recently described the occurrence of balancing selection around a diabetes susceptibility SNP in *CAPN10* intron 13.

A related scenario, referred to as the ‘sodium retention’ hypothesis, was proposed to explain the interethnic differences in the prevalence of hypertension, and, in general, the high incidence of hypertension in modern human societies. Under this model, ancestral populations living in equatorial Africa adapted to hot and humid climates by increasing the rate of sodium and water retention. Also, in those environmental conditions, alleles that increased arterial tone and cardiac contraction force were likely favoured. As populations moved out of Africa into cooler, drier environments with higher salt availability, this high level of sodium retention probably became deleterious and ancestral alleles adapted to the African environment are responsible for the high incidence of hypertension in modern societies. This hypothesis has found some support in the analysis of the frequency distribution of a few hypertension associated alleles in worldwide populations [29].

1.4 The hygiene hypothesis: an evolutionary perspective

INSTITUT PASTEUR

Microbes and Infection 12 (2010) 421–427

www.elsevier.com/locate/micinf

Review

The hygiene hypothesis: an evolutionary perspective

Manuela Sironi [a,*], Mario Clerici [b]

[a] Scientific Institute IRCCS E. Medea, Bioinformatics, Via don L. Monza 20, 23842 Bosisio Parini (LC), Italy
[b] Immunology, University of Milano, Milano and Fondazione Don C. Gnocchi, IRCCS, Milano, Italy

Received 2 February 2010; accepted 10 February 2010
Available online 21 February 2010

Abstract

The hygiene hypothesis relies on the assumption that humans have adapted to a pathogen-rich environment that no longer exists in industrialized societies. Recent advances in molecular immunology and population genetics allow deeper insight into the evolution and co-evolution of host–pathogen interactions and, therefore, into the foundations of the hygiene hypothesis.

Keywords: Natural selection; Population genetics; Host–pathogen interactions

1. Introduction

The incidence of most chronic inflammatory disease has steadily increased in the industrialized world. In the attempt to justify this curious phenomenon, the so called "hygiene hypothesis" was formulated [1]. This highly fascinating hypothesis captured the imagination of scientists worldwide, and has been the focus of both vehement attacks and strong praises. The hypothesis states that the profound changes in the environmental conditions and in the health care system associated with life in the industrialized world have resulted in a relative sterilization of the world that surrounds us. This in turn has drastically reduced the exposure of the immune system to antigens, leading to an imbalance in immune responses that favours the development of chronic inflammatory conditions. In brief: Too much cleanliness prevents the development of a well-balanced immune response; a touch of microbes is good for your health.

At the beginning, epidemiology drove the generation of the hygiene hypothesis. Thus, the observations that atopic diseases are less frequent in: 1) families characterized by a big number of siblings [1,2]; 2) children who attended day care centers early in life [3–5]; and 3) individuals who grew up in the countryside compared to people who were born in cities [6,7], led to the hypothesis that "dirtier" environments have a protective effect toward the development of such diseases.

Once the hygiene hypothesis was formulated, as stated mostly using epidemiological criteria, the efforts to explain it on immunological, genetic, and biochemical terms began. The first line of thought was based on the Th1/Th2 paradigm. In the 80s, this paradigm had a deflagrating impact on the scientific community, leading to a profound rethinking of the pathogenesis of most human diseases, including cancer, autoimmune conditions, and HIV infection (reviewed in [8]). It seemed just natural to apply such parading to explain the hygiene hypothesis. The Th1/Th2 theory states that Th1 lymphocytes are associated with the production of pro-inflammatory cytokines and the development of cell-mediated (CMI) autoimmune conditions, whereas anti-inflammatory cytokines are generated by Th2 lymphocytes; the hyper-activation of these cells also results in humoral immunity-mediated diseases, including allergies. It was suggested that a reduced exposure to Th1-stimulating antigens would have created a niche in which Th2 cells could expand (reviewed in [9–11]). Although highly suggestive, this idea was subsequently weakened by the observations that: 1) high concentration of the Th1 cytokines IL-12 and IFNg are seen in conditions such as atopic dermatitis and asthma [12], and 2) atopic conditions are not more frequent in

* Corresponding author. Tel.: +39 031877915; fax: +39 031877499.
E-mail address: manuela.sironi@bp.lnf.it (M. Sironi).

1286-4579/$ - see front matter © 2010 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.micinf.2010.02.002

genetic abnormalities associated with alterations of Th1 cytokines [13]. The additional realization that some of the chronic inflammatory disease whose incidence is growing in industrialized countries are classically CMI (i.e. Th1)-mediated clarified that the Th1/Th2 paradigm cannot justify by itself the hygiene hypothesis.

The "next big thing" in immunology was the identification of T regulatory (T_{reg}) cells. In the yin and yang complex network of factors that regulate the homeostasis of the immune system, allowing the immune responses to destroy exogenous invaders, but preventing the immune-mediated attack of the host, these cells play a fundamental role. T_{reg} cells have convincingly been shown to modulate immune reactivity and inflammation, and quantitative/qualitative alterations of such cells result in the honing of inflammation, favouring autoimmune processes. Thus, T_{reg} cells have the fundamental role of dampening immune responses (for a recent review see [14]). A recent hypothesis has suggested that some pathogens, including intestinal saprophytes and some helminths are part of the mammalian evolutionary history ("old friends"), and the immune system has grown accustomed to them [11,15]. In immunological terms, this means that the immune response is tolerized toward these "old friends". To simplify, pattern recognition receptors, including toll like receptors, expressed on immature dendritic cells (DC) are stimulated by antigens on "old friends", this interaction results in the maturation of DC that elicit T_{reg} cells-mediated responses to these organisms. This provokes a continuous activation of the immune regulatory mechanisms that modulate the homeostasis of the immune response. The necessary corollary to this is that the "excessive" hygiene associated with life in industrialized countries leads to a reduction of "old friends", a diminished stimulation of T_{reg} cells, and an impairment in the homeostasis of the immune response, with a consequent resulting increased incidence of disease stemming from alterations in immune regulation.

The road ahead is the clarification of the relationships between the environment, the immune system, and the host's genetic background. Because: 1) the quantity and the quality of immune responses is under tight genetic control, and 2) the evolutionary pressure exerted over million of years by the environment has shaped our genetic patrimony, it appears that population genetics data are an essential tool to shed further light on the hygiene hypothesis.

2. Guilty fellow travellers

Similarly to all living organisms, humans have co-evolved with a wide range of organisms, both pathogenic and harmless. Historically, much more effort has been devoted to the understanding of the genetic basis of human–pathogen interactions, while the relevance of commensal and pseudo-commensal organisms in human health and disease (and in human evolution) has only received attention in recent years.

Since their emergence as a species humans are thought to have gone through different phases characterized by extremely different pathogen loads. Early human societies with their small population size probably supported a relatively limited pathogen fauna mainly consisting of organisms with high transmission rates and conferring little immunity [16]. The advent of agriculture around 10,000 years ago, allowed the development of much larger communities and, with them, the spread of several new pathogens (e.g. measles and pertussis) that require high population densities to maintain themselves and become endemic [16,17]. Similarly, pastoralism exposed humans to a wide range of zoonoses, further contributing to the burden of infectious diseases [16,17].

Overall, it has been estimated that the average life expectancy since the Paleolithic through to the mid 19th century was 25 years, with most casualties resulting from infections [18]. While in the first world life expectancy has greatly increased in the last 150 years, in many developing countries the situation has not changed, as infectious diseases still account for about 48% of worldwide deaths among people younger than 45 years [19].

Quite obviously, the improved life expectancy in industrialized societies is not the result of a genetic adaptation to pathogens but derives from the development of vaccines and antibiotics, as well as from improved hygiene conditions. Thus, technological and cultural adaptation has greatly outpaced genetic change and the outcome of this subversion of cause–consequence relationships is a central conundrum in the hygiene hypothesis.

In the late 1940s Haldane [20] first suggested that the selective pressure imposed by infectious agents might be responsible for the maintenance of deleterious alleles and, consequently phenotypes, in human populations; his observation referred to the prevalence of thalassemia in Mediterranean regions, suggested to result from the selective pressure imposed by malaria. This hypothesis was subsequently confirmed and today we know that *Plasmodium* has shaped the evolutionary fate of thousands of human genes [21]. Following Haldane's insight, we may now wonder how many deleterious alleles that contribute to chronic inflammatory conditions are segregating in human populations as a result of pathogen-driven selection, and how many of these alleles are unfit (or useless) to the hygienic environment of industrialized societies. While *Plasmodium* might be unique in certain respects due to its targeting red blood cells and to its long-lasting interactions with humans, other pathogens might have left selection signatures which are more difficult to identify. Hence, many different organisms might target the same host molecule and impinge on the same cellular pathways by exerting different and possibly contrasting pressures.

Evolutionary immunobiology moved its steps after Haldane's observations but it was not until the development of new technologies that scientists were allowed to gain a general overview of the evolutionary processes in the immune system. In the last decade, comparative genomic studies indicated that both at the protein coding and at the non-coding regulatory level, immune response genes are generally less evolutionary constrained and more frequently targeted by positive selection than those involved in other biological processes [22–24]. This observation confirms the dynamic nature of the immune

system, a system that has been constantly adapting to a rapidly changing and highly diverse pathogen fauna, and indicates the general tendency for evolution to favour novelty as a response to biotic threats.

More recently the increasing availability of human intraspecific genetic data and the development of novel population genetics methods, allowed extensive analysis of the evolutionary processes that shaped the immune system in our recent evolutionary history, and that underlie phenotypic variability in humans.

Genome-wide population genetics approaches aiming at identifying genes subjected to positive selection (a selection regime favouring the spread of a selected allele in populations) did not identify an excess of immune response genes among selection targets, with the exception of studies focusing on very recent (less than 30,000 years) selective events [25,26]. Nevertheless, as the authors of these latter studies also note, several of the identified genes might represent balancing selection allele systems. Balancing selection is a situation whereby genetic variability is maintained as a result of selection. The best known example in humans involves MHC genes, genes that are characterized by extreme polymorphism levels partially maintained by pathogen-driven selective pressure [27]. Recent evidences indicated that balancing selection, which is considered to be rare in humans, has acted on several immune-related genes [28–32]. This observation has been regarded both as a confirmation that genetic diversity allows increased flexibility in immune responses (e.g. though heterozygote advantage or by maintaining rare alleles) and as an indication that immune responses need to find a "balance" between efficacy in fighting invaders and tolerance to self and innocuous antigens.

Recent studies have also addressed the role of specific pathogen classes on the evolution of human genes. Viruses were shown to act as a strong selective force on *TLR3*, *TLR7*, *TLR8* and *TLR9* [33], while a genome-wide search identified several additional genes subjected to virus-driven selective pressure [34]. Interestingly, some of these loci (e.g. *NR4A2* and *PPP3CA*) have been involved in the pathogenesis of multiple sclerosis (MS), and *IFIH1* (also known as *MDA-5*), a susceptibility gene for type 1 diabetes (T1D) and a sensor of double strand RNA, has evolved in response to viral threats. Similarly, a genome-wide search for signatures of protozoa-driven selective pressure identified a large number of variants and loci, both involved in red cell homeostasis and in immune-related processes [21].

Among the large diversity of organisms that can cause disease in humans, helminths have attracted considerable attraction in relation to the hygiene hypothesis. The ability of parasitic worms to elicit immunoregulatory circuits and to modify the prevalence or severity of different immune-related conditions in both humans and experimental mouse models has been the topic of several reviews [35,36]. From an evolutionary standpoint, it is worth mentioning that although helminthic infections are rarely fatal, the highest parasite burdens are usually observed during childhood, often resulting in anemia, malnourishment and growth stunting (reviewed in [37]). During pregnancy helminth infection is a risk factor for preterm delivery, reduced birth weight and maternal mortality [37]. Moreover, by chronically infecting their host, parasitic worms increase the susceptibility to other pathogens such as viruses, bacteria and protozoa [37]. Therefore, these organisms may have acted as a powerful selective force throughout human history.

Helminths are thought to have appeared more or less at the same time as the adaptive immune system developed, indicating that vertebrates and parasites are likely to have co-evolved [38]. This implies that our immune system has had a long training in fighting helminths and has probably grown prepared to meet these organisms and their antigens.

In line with these considerations, we have recently shown that parasitic worms have acted as a stronger selective pressure, compared to viral/microbial pathogens, on a set of human genes encoding interleukins and interleukin receptors [29]. It should be kept in mind that our ability to detect selection signatures depends not only on the selection coefficient (i.e. strength of selective pressure) but also on additional factors such as the stability and duration of the selective pressure over time (e.g. pathogens such as HIV have appeared too recently to leave detectable signatures despite their clear role as selective agents) and space (different populations respond to pathogens transmitted in the environment they live in). Whatever the reasons, parasitic worms have left marks on our genes that possibly involve thousands of human loci (MS and MC unpublished results).

Therefore, all these studies have provided new evidence for an old idea: that host–pathogen interactions represent one of the major drivers of molecular evolution.

3. Adaptation to pathogens and disease

The identification of several immune response genes as targets of natural selection maybe necessary to support the hygiene hypothesis, but is not sufficient. In order to go a step forward we have to draw a direct link among genetic adaptation to invading organisms (either pathogenic or harmless), infection, and susceptibility to chronic inflammatory diseases.

An advance in this direction came from analysis of the *HAVCR1* gene. *HAVCR1* encodes a membrane protein expressed by different T cell subtypes (including Th2 and T_{reg}) and functions as the receptor for hepatitis A virus (HAV). HAV infection has been shown to be protective against atopy in subjects carrying a polymorphic 6 aminoacid insertion in the *HAVCR1* mucin-like domain [39]. Although the mechanisms responsible for HAV-induced atopy protection are unknown, these observations provide proof of principle that infectious agents can modulate the predisposition to atopic disorders and that this effect depends on the host's genetic background. Moreover, in industrialized countries the seroprevalence of antibodies against HAV has dropped from 100% to 25–30% in the last 40 years [40], suggesting that this virus might contribute to several epidemiological observations supporting the hygiene hypothesis. *HAVCR1* has been subjected to positive selection during primate evolution [24] and nucleotide

diversity in its mucin-like domain (containing the 6 aminoacid insertion/deletion polymorphism) has been maintained by balancing selection in human populations [41]. The selective pressure underlying the evolutionary history of *HAVCR1* is likely to be pathogen-driven, suggesting that the identification of genes and alleles that have been selected by human pathogens might help to address whether the foundations of the hygiene hypothesis hold. In this direction, we have recently shown that several interleukin/interleukin receptor genes involved in the pathogenesis of inflammatory bowel disease (IBD) and celiac disease (CeD) have been subjected to pathogen-driven selective pressure. Specifically, IBD/CeD risk alleles in *IL18RAP* (Fig. 1), *IL18R1*, *IL23*, *IL18R1* and in the intergenic region between *IL2* and *IL21* display higher frequencies in populations exposed to high microbial/viral loads, suggesting that these variants play a role in the response to these organisms. The reaction to microbial infections is mainly Th1 mediated (as opposed to the Th2-mediated response elicited by parasitic worms), as is the pathogenesis of IBD; these data therefore offer further support to the Th1/Th2 paradigm for the development of chronic inflammatory conditions, and suggest that helminth infections in populations carrying risk alleles for IBD might be partially responsible for the low incidence of this disease in developing countries.

More recently, Barreiro and Quintana-Murci [42] have analysed a set of single nucleotide polymorphisms (SNPs) associated with immunity-related phenotypic traits. Interestingly, they observed that several alleles or haplotypes associated with augmented susceptibility to autoimmune disorders including T1D, CeD and MS have been the target of recent positive selection, suggesting that some unknown selective pressure, possibly pathogen-driven, has resulted in an increased frequency of risk alleles/haplotypes in human populations.

Therefore, these observations support the notion whereby adaptation to pathogen exposure may result in the selection for alleles that confer increased protection against infections but predispose to autoimmunity.

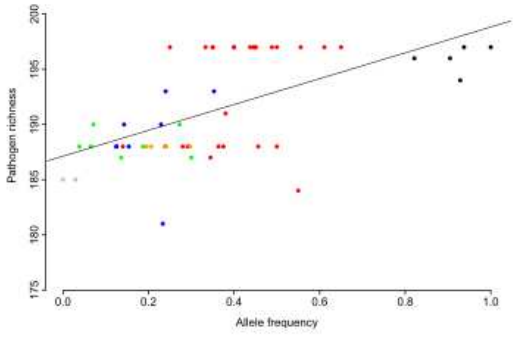


Fig. 1. Correlation between pathogen richness (a measure of pathogen-driven selection) and allele frequency in 52 human populations distributed worldwide for rs917997 in *IL18RAP*. Populations from different broad geographic areas are coded by different colors: Sub-Saharan Africa (green), Central/South America (black), Asia (red), Europe (blue), Middle East (orange), and Oceania (gray) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

Although these data seem to support, from an evolutionary perspective, the hygiene hypothesis the situation is likely to be more complex. First, although the theory of imbalanced Th1/Th2 responses might represent an oversimplification, it is likely that helminths and microbial/viral agents have exerted a selective pressure both on different gene sets and on a portion of common genes so that the resulting phenotypes, once infections are removed, are not easily predictable. These considerations might also help to reconcile the hygiene hypothesis with the observation that specific infections can trigger or exacerbate some autoimmune diseases (reviewed in [43]). Second, some observations on the role of known disease risk alleles pose unanswered questions. As an example, an IBD risk allele in *IL18RAP* was shown to decrease gene expression level [44] and we found this same allele to be subjected to pathogen-driven selection (Fig. 1) [29]. The reason why decreased *IL18RAP* expression is a risk factor for IBD, which is characterized by increased inflammation, and is selected by pathogens remains unknown. Third, growing evidences indicate that, while risk alleles are often shared among different autoimmune diseases, alleles also exist, both in HLA and non-HLA genes, that predispose to one condition but protect from another [45]. As an example of opposite risk profiles, the A allele of rs10484565 in *TAP2* is protective against MS and autoimmune thyroid disease but predisposes to T1D, rheumatoid arthritis and ankylosing spondylitis [45]. These observations imply that the idea that alleles protecting from infections are maintained in populations due to selection, but these same alleles predispose to autoimmunity once the environment is hygienised is an oversimplification. At present, no study has specifically addressed the evolutionary history of *TAP2*, nor the selective regime underlying the maintenance of other alleles with opposite risk profiles. Further studies will be required to address this point but we do not necessarily need to invoke selection at all costs and at all loci. Indeed, from an evolutionary perspective, it is worth wondering whether predisposition to autoimmunity or atopy/allergy is likely to have caused a significant fitness reduction in populations so as to result in the counter-selection of risk alleles. For example, T1D, a juvenile-onset disease, was lethal until the discovery of insulin in 1922; yet, genetic variants that predispose to T1D are segregating at high frequency in human populations. Why has evolution failed to eliminate them? One possibility is that the majority of these potentially lethal alleles have been maintained by their conferring some other advantage (or by being in linkage disequilibrium with some favourable allele) therefore resulting in a balancing selection regime. An alternate possibility is that a portion of these alleles has behaved nearly neutrally during most of human history because the environmental conditions did not predispose to T1D. In a situation where abundant pathogens, harmless microbes (see below) and "old friends" provided the development of immunoregulatory circuits (e.g. via T_{regs}), allelic variation at a number of genes might have resulted in little phenotypic effect. In brief, the removal of several pathogenic and harmless

organisms from our environment might have unveiled the presence in our genome of many alleles that were once "balanced" or nearly neutral but are harmful today. If this were the case, we might expect (and strive to) identify selection signatures in a probably limited number of "master regulatory genes".

4. Blameless guests

As reported above, humans have not only evolved to fight their pathogens, but also to live with and take advantage from a large number of symbionts. Specific microbial communities thrive in distinct sites of the human body such as the skin and mucosal surfaces, and the gut (reviewed in [46]). Although extensive inter-individual differences exist, the human microbial community is distinct from that of other mammals, indicating that humans have co-evolved with their microbial partners (reviewed in [46,47]). Co-evolution is based on the vertical transmission of microbial communities from parent to off-spring and on the ability of the symbiont community as a whole to confer a fitness advantage to the host [48]. The selective benefits originating from the maintenance of a diverse and complex microbial community were hypothesized to be responsible for the evolution of the adaptive immune system in vertebrates (invertebrates, relying on the innate immune system alone, display a much simpler microbiota) [49]. This idea is based on the concept whereby adaptive immunity maybe essential to the tolerance of our microbial partners. Indeed, recent evidences indicated that tolerance to the gut microbiota is partially mediated by IgA production in mice [50,51]. Furthermore, genetic diversity at mouse MHC loci seems to be involved in shaping the intestinal microbiota composition [52] and in determining microbial-mediated scent production [53]. Yet, the relative contribution of genetic and environmental factors in determining the individual's microbiota composition is still poorly understood. Studies in mice indicated that genotype at the leptin gene influences the relative abundance of different microbial species, while the microbiota composition is mainly determined by kinship [48]. Another study reported that the knocking out of *Myd88*, encoding an adaptor molecule for different TLRs, changes the distal gut microbiota composition resulting in protection from T1D in NOD (non-obese diabetic) mice [54].

In humans, the contribution of the host genotype in shaping the composition of the symbiont community is even less clear and analyses in monozygotic and dizygotic twins [55–57] yielded different results, highlighting the need for larger and possibly time-course studies.

The role of early life environmental factors (e.g. diet, microbial exposure, delivery method) in driving human gut microbiota composition is beginning to be investigated [58], although the extreme inter-personal diversity is likely to pose several hurdles to the clarification of these issues.

In parallel, increasing evidence suggests that our microbial partners participate in immune system development and in predisposition/protection from immune-related diseases. A polysaccharide produced by *Bacteroides fragilis*, a ubiquitous commensal bacterium, promotes immune system maturation in mice and mediates the establishment of a proper Th1/Th2 balance [59]. Additionally, exposure to enteric bacterial antigens is crucial for the generation and expansion of T_{reg} cells within peripheral tissues [60], and the composition of the intestinal microbiota regulates the balance between Th17 and T_{reg} cells in the lamina propria [61]. These data clearly imply that changes of the gut microbial composition have the potential to affect immune system homeostasis and, consequently, predisposition to disease. As mentioned above, $Myd88^{-/-}$ NOD mice are protected from T1D secondary to alteration of gut microbial composition [54] and IBD patients display a reduced diversity of gut microbiota (especially Firmicutes and Bacteroidetes) [62]. Similarly, alteration of gut microbial composition has been observed in infants and adults suffering from eczema, allergy and atopic dermatitis [63–65].

These studies therefore establish a link between alteration of our symbiont community and loss of immune homeostasis. Our knowledge of the human microbiome composition may still be too limited to allow extensive analyses on how modern sanitation and, in general, life conditions in industrialized countries, affect our microbial partners. The widespread use of antibiotics during infancy, for example might drastically deplete symbiont communities at a critical stage of immune system development and result in immune imbalances. In line with this view, antibiotic use in early childhood has been associated with an increase risk of developing asthma [66,67]. Therefore, the use of new compounds, and the overall ecological changes in industrialized countries, might have partially disrupted a co-evolved mutualism between humans and microbes.

5. Conclusions

"We should think of each host and its parasites as a superorganism with the respective genomes yoked into a chimera of sorts" Lederberg noticed a few years ago [68]; this statement applies very well to the communion of humans and their symbionts. Are our genes or theirs to blame for the rise in chronic inflammatory diseases we have witnessed in the last decades? As noted above, an interesting possibility is that the change or depletion of microbiota composition and diversity (together with the disappearance of several pathogens) has created the environmental conditions that favour the development of chronic inflammatory disorders in genetically susceptible individuals. Molecular evolutionary biology and population genetics are relatively new sciences that offer extremely potent tools helping to clarify how humans deal with a hostile environment. An interesting implication of results obtained by the application of these tools is that adaptation to the environment is a double-edged sword: what protects us from pathogens could favour non-transmissible degenerative diseases. To quote old wisdom: you cannot have your cake and eat it too. The data summarized herein offer further support to the hygiene hypothesis, a hypothesis that has so far mostly relied on epidemiologic and immunologic results.

In recent years, association studies have provided a wealth of information on the genetic basis of chronic inflammatory diseases. Yet, the large majority of genetic variants that have been identified in these studies are likely to be genetic markers rather than causal polymorphisms. Similarly, most population genetics studies have been successful in the identification of genes or gene regions subjected to natural selection but the real selection targets and selective pressures have rarely been determined. Further insight into the complex relationship among genetic background, immune imbalances and infections (or lack thereof) will require extensive efforts aimed at describing genotype/phenotype relationships and host–microbial interactions at the molecular level.

Up to now association studies, population genetics analyses and epidemiological surveys have largely proceeded on separate routes. Closer integration of these research fields might help to clarify how humans adapt to the environment and how environmental change affects immune responses and immunopathology.

Acknowledgments

MC is supported by grants from Istituto Superiore di Sanita' "Programma Nazionale di Ricerca sull' AIDS", the EMPRO and AVIP EC WP6 Projects, the nGIN EC WP7 Project, the Japan Health Science Foundation, 2008 Ricerca Finalizzata [Italian Ministry of Health], 2008 Ricerca Corrente [Italian Ministry of Health], Progetto FIRB RETI: Rete Italiana Chimica Farmaceutica CHEM-PROFARMA-NET [RBPR05NWWC], and Fondazione CARIPLO. MS is a member of the Doctorate School in Molecular Medicine, University of Milan.

References

[1] D.P. Strachan, Hay fever, hygiene, and household size. BMJ 299 (1989) 1259–1260.
[2] D.P. Strachan, Allergy and family size: a riddle worth solving. Clin. Exp. Allergy 27 (1997) 235–236.
[3] R.F. Lemanske Jr., The childhood origins of asthma (COAST) study. Pediatr. Allergy Immunol. 13 (Suppl. 15) (2002) 38–43.
[4] J.C. Celedon, R.J. Wright, A.A. Litonjua, D. Sredl, L. Ryan, S.T. Weiss, D.R. Gold, Day care attendance in early life, maternal history of asthma, and asthma at the age of 6 years. Am. J. Respir. Crit. Care Med. 167 (2003) 1239–1243.
[5] S. Hoffjan, J.T. Epplen, The genetics of atopic dermatitis: recent findings and future options. J. Mol. Med. 83 (2005) 682–692.
[6] J. Riedler, C. Braun-Fahrlander, W. Eder, M. Schreuer, M. Waser, S. Maisch, D. Carr, R. Schierl, D. Nowak, E. von Mutius, ALEX Study Team, Exposure to farming in early life and development of asthma and allergy: a cross-sectional survey. Lancet 358 (2001) 1129–1133.
[7] C. Braun-Fahrlander, J. Riedler, U. Herz, W. Eder, M. Waser, L. Grize, S. Maisch, D. Carr, F. Gerlach, A. Bufe, R.P. Lauener, R. Schierl, H. Renz, D. Nowak, E. von Mutius, Allergy and Endotoxin Study Team, Environmental exposure to endotoxin and its relation to asthma in school-age children. N. Engl. J. Med. 347 (2002) 869–877.
[8] D.R. Lucey, M. Clerici, G.M. Shearer, Type 1 and type 2 cytokine dysregulation in human infectious, neoplastic, and inflammatory diseases. Clin. Microbiol. Rev. 9 (1996) 532–562.
[9] D. Vercelli, Mechanisms of the hygiene hypothesis–molecular and otherwise. Curr. Opin. Immunol. 18 (2006) 733–737.
[10] H. Garn, H. Renz, Epidemiological and immunological evidence for the hygiene hypothesis. Immunobiology 212 (2007) 441–452.
[11] G.A. Rook, Review series on helminths, immune modulation and the hygiene hypothesis: the broader implications of the hygiene hypothesis. Immunology 126 (2009) 3–11.
[12] N. Krug, J. Madden, A.E. Redington, P. Lackie, R. Djukanovic, U. Schauer, S.T. Holgate, A.J. Frew, P.H. Howarth, T-cell cytokine profile evaluated at the single cell level in BAL and blood in allergic asthma. Am. J. Respir. Cell Mol. Biol. 14 (1996) 319–326.
[13] D.A. Lammas, J.L. Casanova, D.S. Kumararatne, Clinical consequences of defects in the IL-12-dependent interferon-gamma (IFN-gamma) pathway. Clin. Exp. Immunol. 121 (2000) 417–425.
[14] S. Sakaguchi, K. Wing, Y. Onishi, P. Prieto-Martin, T. Yamaguchi, Regulatory T cells: how do they suppress immune responses? Int. Immunol. 21 (2009) 1105–1111.
[15] G.A. Rook, C.A. Lowry, The hygiene hypothesis and psychiatric disorders. Trends Immunol. 29 (2008) 150–158.
[16] A. Dobson, in: S. Jones, R. Martin, D. Pilbeam (Eds.), The Cambridge Encyclopedia of Human Evolution, Cambridge University Press, Cambridge, 1992, pp. 411–420.
[17] N.D. Wolfe, C.P. Dunavan, J. Diamond, Origins of major human infectious diseases. Nature 447 (2007) 279–283.
[18] J.L. Casanova, L. Abel, Inborn errors of immunity to infection: the rule rather than the exception. J. Exp. Med. 202 (2005) 197–201.
[19] C. Kapp, WHO warns of microbial threat. Lancet 353 (1999) 2222.
[20] J.B.S. Haldane, Anonymous Selected Genetic Papers of J.B.S. Haldane. Garland Publ. Inc., New York and London, 1949, 325–334.
[21] U. Pozzoli, M. Fumagalli, R. Cagliani, G.P. Comi, N. Bresolin, M. Clerici, M. Sironi, The role of protozoa-driven selection in shaping human genetic variability. Trends Genet. (2010). doi: 10.1016/j.tig.2009.12.010.
[22] C.I. Castillo-Davis, F.A. Kondrashov, D.L. Hartl, R.J. Kulathinal, The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. Genome Res. 14 (2004) 802–811.
[23] M. Sironi, G. Menozzi, G.P. Comi, R. Cagliani, N. Bresolin, U. Pozzoli, Analysis of intronic conserved elements indicates that functional complexity might represent a major source of negative selection on non-coding sequences. Hum. Mol. Genet. 14 (2005) 2533–2546.
[24] C. Kosiol, T. Vinar, R.R. da Fonseca, M.J. Hubisz, C.D. Bustamante, R. Nielsen, A. Siepel, Patterns of positive selection in six mammalian genomes. PLoS Genet. 4 (2008) e1000144.
[25] B.F. Voight, S. Kudaravalli, X. Wen, J.K. Pritchard, A map of recent positive selection in the human genome. PLoS Biol. 4 (2006) e72.
[26] E.T. Wang, G. Kodama, P. Baldi, R.K. Moyzis, Global landscape of recent inferred Darwinian selection for *Homo sapiens*. Proc. Natl. Acad. Sci. U.S.A. 103 (2006) 135–140.
[27] F. Prugnolle, A. Manica, M. Charpentier, J.F. Guegan, V. Guernier, F. Balloux, Pathogen-driven selection and worldwide HLA class I diversity. Curr. Biol. 15 (2005) 1022–1027.
[28] M. Fumagalli, R. Cagliani, U. Pozzoli, S. Riva, G.P. Comi, G. Menozzi, N. Bresolin, M. Sironi, Widespread balancing selection and pathogen-driven selection at blood group antigen genes. Genome Res. 19 (2009) 199–212.
[29] M. Fumagalli, U. Pozzoli, R. Cagliani, G.P. Comi, S. Riva, M. Clerici, N. Bresolin, M. Sironi, Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions. J. Exp. Med. 206 (2009) 1395–1408.
[30] R. Cagliani, M. Fumagalli, S. Riva, U. Pozzoli, G.P. Comi, G. Menozzi, N. Bresolin, M. Sironi, The signature of long-standing balancing selection at the human defensin beta-1 promoter. Genome Biol. 9 (2008) R143.
[31] A.M. Andres, M.J. Hubisz, A. Indap, D.G. Torgerson, J.D. Degenhardt, A.R. Boyko, R.N. Gutenkunst, T.J. White, E.D. Green, C.D. Bustamante, A.G. Clark, R. Nielsen, Targets of balancing selection in the human genome. Mol. Biol. Evol. 26 (2009) 2755–2764.
[32] A. Ferrer-Admetlla, E. Bosch, M. Sikora, T. Marques-Bonet, A. Ramirez-Soriano, A. Muntasell, A. Navarro, R. Lazarus, F. Calafell, J. Bertranpetit, F. Casals, Balancing selection is the main force shaping the evolution of innate immunity genes. J. Immunol. 181 (2008) 1315–1322.

[33] L.B. Barreiro, M. Ben-Ali, H. Quach, G. Laval, E. Patin, J.K. Pickrell, C. Bouchier, M. Tichit, O. Neyrolles, B. Gicquel, J.R. Kidd, K.K. Kidd, A. Alcais, J. Ragimbeau, S. Pellegrini, L. Abel, J.L. Casanova, L. Quintana-Murci, Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. PLoS Genet. 5 (2009) e1000562.
[34] M. Fumagalli, U. Pozzoli, R. Cagliani, G.P. Comi, N. Bresolin, M. Clerici, M. Sironi, Genome-wide identification of susceptibility alleles for viral infections through a population genetics approach. PLoS Genet. 6 (2010) e1000849.
[35] D.W. Dunne, A. Cooke, A worm's eye view of the immune system: consequences for evolution of human autoimmune disease. Nat. Rev. Immunol. 5 (2005) 420–426.
[36] R.M. Maizels, A. Balic, N. Gomez-Escobar, M. Nair, M.D. Taylor, J.E. Allen, Helminth parasites–masters of regulation. Immunol. Rev. 201 (2004) 89–116.
[37] P.J. Hotez, P.J. Brindley, J.M. Bethony, C.H. King, E.J. Pearce, J. Jacobson, Helminth infections: the great neglected tropical diseases. J. Clin. Invest. 118 (2008) 1311–1321.
[38] J.A. Jackson, I.M. Friberg, S. Little, J.E. Bradley, Review series on helminths, immune modulation and the hygiene hypothesis: immunity against helminths and immunological phenomena in modern human populations: coevolutionary legacies? Immunology 126 (2009) 18–27.
[39] J.J. McIntire, S.E. Umetsu, C. Macaubas, E.G. Hoyte, C. Cinnioglu, L.L. Cavalli-Sforza, G.S. Barsh, J.F. Hallmayer, P.A. Underhill, N.J. Risch, G. J. Freeman, R.H. DeKruyff, D.T. Umetsu, Immunology: hepatitis A virus link to atopic disease. Nature 425 (2003) 576.
[40] J.F. Bach, The effect of infections on susceptibility to autoimmune and allergic diseases. N. Engl. J. Med. 347 (2002) 911–920.
[41] T. Nakajima, S. Wooding, Y. Satta, N. Jinnai, S. Goto, I. Hayasaka, N. Saitou, J. Guan-Jun, K. Tokunaga, L.B. Jorde, M. Emi, I. Inoue, Evidence for natural selection in the HAVCR1 gene: high degree of amino-acid variability in the mucin domain of human HAVCR1 protein. Genes Immun. 6 (2005) 398–406.
[42] L.B. Barreiro, L. Quintana-Murci, From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat. Rev. Genet. 11 (2010) 17–30.
[43] S. Kivity, N. Agmon-Levin, M. Blank, Y. Shoenfeld, Infections and autoimmunity–friends or foes? Trends Immunol. 30 (2009) 409–414.
[44] K.A. Hunt, A. Zhernakova, G. Turner, G.A. Heap, L. Franke, M. Bruinenberg, J. Romanos, L.C. Dinesen, A.W. Ryan, D. Panesar, R. Gwilliam, F. Takeuchi, W.M. McLaren, G.K. Holmes, P.D. Howdle, J.R. Walters, D. S. Sanders, R.J. Playford, G. Trynka, C.J. Mulder, M.L. Mearin, W.H. Verbeek, V. Trimble, F.M. Stevens, C. O'Morain, N.P. Kennedy, D. Kelleher, D.J. Pennington, D.P. Strachan, W.L. McArdle, C.A. Mein, M.C. Wapenaar, P. Deloukas, R. McGinnis, R. McManus, C. Wijmenga, D.A. van Heel, Newly identified genetic risk variants for celiac disease related to the immune response. Nat. Genet. 40 (2008) 395–402.
[45] M. Sirota, M.A. Schaub, S. Batzoglou, W.H. Robinson, A.J. Butte, Autoimmune disease classification by inverse association with SNP alleles. PLoS Genet. 5 (2009) e1000792.
[46] L. Dethlefsen, M. McFall-Ngai, D.A. Relman, An ecological and evolutionary perspective on human-microbe mutualism and disease. Nature 449 (2007) 811–818.
[47] R.E. Ley, C.A. Lozupone, M. Hamady, R. Knight, J.I. Gordon, Worlds within worlds: evolution of the vertebrate gut microbiota. Nat. Rev. Microbiol. 6 (2008) 776–788.
[48] R.E. Ley, D.A. Peterson, J.I. Gordon, Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell 124 (2006) 837–848.
[49] M. McFall-Ngai, Adaptive immunity: care for the community. Nature 445 (2007) 153.
[50] D.A. Peterson, N.P. McNulty, J.L. Guruge, J.I. Gordon, IgA response to symbiotic bacteria as a mediator of gut homeostasis. Cell. Host Microbe 2 (2007) 328–339.
[51] S. Fagarasan, M. Muramatsu, K. Suzuki, H. Nagaoka, H. Hiai, T. Honjo, Critical roles of activation-induced cytidine deaminase in the homeostasis of gut flora. Science 298 (2002) 1424–1427.
[52] P. Toivanen, J. Vaahtovuo, E. Eerola, Influence of major histocompatibility complex on bacterial composition of fecal flora. Infect. Immun. 69 (2001) 2372–2377.
[53] C.V. Lanyon, S.P. Rushton, A.G. O'donnell, M. Goodfellow, A.C. Ward, M. Petrie, S.P. Jensen, L. Morris Gosling, D.J. Penn, Murine scent mark microbial communities are genetically determined. FEMS Microbiol. Ecol. 59 (2007) 576–583.
[54] L. Wen, R.E. Ley, P.Y. Volchkov, P.B. Stranges, L. Avanesyan, A.C. Stonebraker, C. Hu, F.S. Wong, G.L. Szot, J.A. Bluestone, J.I. Gordon, A. V. Chervonsky, Innate immunity and intestinal microbiota in the development of Type 1 diabetes. Nature 455 (2008) 1109–1113.
[55] J.A. Stewart, V.S. Chadwick, A. Murray, Investigations into the influence of host genetics on the predominant eubacteria in the faecal microflora of children. J. Med. Microbiol. 54 (2005) 1239–1242.
[56] E.G. Zoetendal, A.D.L. Akkermans, W.M. Akkermans-van Vliet, J.A.G.M. de Visser, W.M. de Vos, The host genotype affects the bacterial community in the human gastronintestinal tract. Microb. Ecol. Health Dis. 13 (2001) 129–134.
[57] P.J. Turnbaugh, M. Hamady, T. Yatsunenko, B.L. Cantarel, A. Duncan, R. E. Ley, M.L. Sogin, W.J. Jones, B.A. Roe, J.P. Affourtit, M. Egholm, B. Henrissat, A.C. Heath, R. Knight, J.I. Gordon, A core gut microbiome in obese and lean twins. Nature 457 (2009) 480–484.
[58] J. Penders, C. Thijs, C. Vink, F.F. Stelma, B. Snijders, I. Kummeling, P. A. van den Brandt, E.E. Stobberingh, Factors influencing the composition of the intestinal microbiota in early infancy. Pediatrics 118 (2006) 511–521.
[59] S.K. Mazmanian, C.H. Liu, A.O. Tzianabos, D.L. Kasper, An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. Cell 122 (2005) 107–118.
[60] U.G. Strauch, F. Obermeier, N. Grunwald, S. Gurster, N. Dunger, M. Schultz, D.P. Griese, M. Mahler, J. Scholmerich, H.C. Rath, Influence of intestinal bacteria on induction of regulatory T cells: lessons from a transfer model of colitis. Gut 54 (2005) 1546–1552.
[61] I.I. Ivanov, L. Frutos Rde, N. Manel, K. Yoshinaga, D.B. Rifkin, R.B. Sartor, B.B. Finlay, D.R. Littman, Specific microbiota direct the differentiation of IL-17-producing T-helper cells in the mucosa of the small intestine. Cell. Host Microbe 4 (2008) 337–349.
[62] D.N. Frank, A.L. St Amand, R.A. Feldman, E.C. Boedeker, N. Harpaz, N.R. Pace, Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proc. Natl. Acad. Sci. U.S.A. 104 (2007) 13780–13785.
[63] Y.M. Sjogren, M.C. Jenmalm, M.F. Bottcher, B. Bjorksten, E. Sverremark-Ekstrom, Altered early infant gut microbiota in children developing allergy up to 5 years of age. Clin. Exp. Allergy 39 (2009) 518–526.
[64] K.W. Mah, B. Bjorksten, B.W. Lee, H.P. van Bever, L.P. Shek, T.N. Tan, Y.K. Lee, K.Y. Chua, Distinct pattern of commensal gut microbiota in toddlers with eczema. Int. Arch. Allergy Immunol. 140 (2006) 157–163.
[65] S. Watanabe, Y. Narisawa, S. Arase, H. Okamatsu, T. Ikenaga, Y. Tajiri, M. Kumemura, Differences in fecal microflora between patients with atopic dermatitis and healthy control subjects. J. Allergy Clin. Immunol. 111 (2003) 587–591.
[66] F. Marra, C.A. Marra, K. Richardson, L.D. Lynd, A. Kozyrskyj, D.M. Patrick, W.R. Bowie, J.M. Fitzgerald, Antibiotic use in children is associated with increased risk of asthma. Pediatrics 123 (2009) 1003–1010.
[67] A.L. Kozyrskyj, P. Ernst, A.B. Becker, Increased risk of childhood asthma from antibiotic use in early life. Chest 131 (2007) 1753–1759.
[68] J. Lederberg, Infectious history. Science 288 (2000) 287–293.

2. RESULTS AND DISCUSSION

2.1 Widespread balancing selection and pathogen-driven selection at blood group antigen genes

Letter

Widespread balancing selection and pathogen-driven selection at blood group antigen genes

Matteo Fumagalli,[1,2] Rachele Cagliani,[1] Uberto Pozzoli,[1] Stefania Riva,[1] Giacomo P. Comi,[3] Giorgia Menozzi,[1] Nereo Bresolin,[1,3] and Manuela Sironi[1,4]

[1]*Scientific Institute IRCCS E. Medea, Bioinformatic Lab, 23842 Bosisio Parini (LC), Italy;* [2]*Bioengineering Department, Politecnico di Milano, 20133 Milan, Italy;* [3]*Dino Ferrari Centre, Department of Neurological Sciences, University of Milan, IRCCS Ospedale Maggiore Policlinico, Mangiagalli and Regina Elena Foundation, 20100 Milan, Italy*

Historically, allelic variations in blood group antigen (BGA) genes have been regarded as possible susceptibility factors for infectious diseases. Since host–pathogen interactions are major determinants in evolution, BGAs can be thought of as selection targets. In order to verify this hypothesis, we obtained an estimate of pathogen richness for geographic locations corresponding to 52 populations distributed worldwide; after correction for multiple tests and for variables different from selective forces, significant correlations with pathogen richness were obtained for multiple variants at 11 BGA loci out of 26. In line with this finding, we demonstrate that three BGA genes, namely *CD55*, *CD151*, and *SLC14A1*, have been subjected to balancing selection, a process, rare outside MHC genes, which maintains variability at a locus. Moreover, we identified a gene region immediately upstream the transcription start site of *FUT2* which has undergone non-neutral evolution independently from the coding region. Finally, in the case of *BSG*, we describe the presence of a highly divergent haplotype clade and the possible reasons for its maintenance, including frequency-dependent balancing selection, are discussed. These data indicate that BGAs have been playing a central role in the host–pathogen arms race during human evolutionary history and no other gene category shows similar levels of widespread selection, with the only exception of loci involved in antigen recognition.

[Supplemental material is available online at www.genome.org.]

Since the discovery of the ABO blood group in 1900 by Karl Landsteiner, as many as 29 blood group (BG) systems have been identified in humans (Blood Group Antigen Gene Mutation Database, BGMUT [Blumenfeld and Patnaik 2004]). Each system is specified by a blood group antigen (BGA) constituted by a protein or carbohydrate molecule which is expressed on the erythrocyte membrane and is polymorphic in human populations.

The molecular basis of all blood group systems (except for P1) has been clarified, with one or more polymorphic loci accounting for BG phenotypes. BGA genes belong to different functional categories, including receptors, transporters, channels, adhesion molecules, and enzymes; among the latter, the great majority of loci code for glycotransferases. While a few BGAs are confined to the erythrocyte membrane, others are expressed at the surface of different cell types or secreted in body fluids (Reid and Lomas-Frances 1997).

The number of different alleles is highly variable among BGA genes and ranges from two to >100 (Blumenfeld and Patnaik 2004) with the most common form of variation being accounted for by missense or nonsense single nucleotide polymorphisms (SNPs). BGA polymorphisms have attracted considerable attention over recent years not only with respect to erythrocyte physiology per se, but also due to the possibility that variations in BGAs might underlie different susceptibility to diseases. In particular, the association between infections and BGA polymorphisms has been extensively investigated, although conclusive results have been obtained in a minority of cases. For example, specific BGA alleles have been shown to alter susceptibility to malaria (Moulds and Moulds 2000), while *FUT2* variants (Lewis system) influence the predisposition to Norwalk virus (Lindesmith et al. 2003) and *Campylobacter* (Ruiz-Palacios et al. 2003) infection, as well as to vulvovaginal candidisis (Hurd and Domino 2004) and urinary tract infections (Schaeffer et al. 2001).

Such findings are in line with the vision whereby different BGAs serve as "incidental receptors for viruses and bacteria" (Moulds et al. 1996), but also function as modulators of innate immune response (Ruiz-Palacios et al. 2003; Linden et al. 2008) and possibly as "decoy-sink" molecules targeting pathogens to macrophages (Gagneux and Varki 1999).

Given this premise and the conundrum whereby host–pathogen interactions are major determinants in evolution, BGAs can be thought of as possible targets of diverse selective pressures. This view is in agreement with the geographic differentiation pattern observed for BGAs and with previous reports of non-neutral evolution at the *ABO*, *DARC*, *GYPA*, and *FUT2* loci (Saitou and Yamamoto 1997; Koda et al. 2001; Baum et al. 2002; Hamblin et al. 2002; Calafell et al. 2008).

Here we exploited the availability of extensive resequencing data, as well as of SNP genotyping in world-wide populations, to investigate the evolutionary forces underlying the evolution of BGA genes: Our data provide evidence that balancing and pathogen-driven selections have acted at multiple BGA loci.

Results and Discussion

Pathogen richness and BGA gene polymorphisms

As a first step we wished to verify whether allele frequencies of SNPs in BGA genes varied with pathogen richness, in terms of

[4]**Corresponding author.**
E-mail manuela.sironi@bp.lnf.it; fax 39-031-877499.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.082768.108.

different species in a geographic location. Similar approaches have been applied to test this same hypothesis for HLA genes (Prugnolle et al. 2005) and for other gene–environment interactions (Thompson et al. 2004; Young et al. 2005; Hancock et al. 2008). To this aim we exploited the fact that a set of over 650,000 tag SNPs has been typed in 52 populations (HGDP-CEPH panel) distributed world wide (Li et al. 2008). As for pathogen richness, we gathered information concerning the number of different micropathogen species from the Gideon database; pathogen richness was calculated on a country basis by pooling together viruses, bacteria, fungi and protozoa (see Methods for further details). A total of 262 SNPs in BGA genes had been typed in the HGDP-CEPH panel allowing analysis of the following loci: *RHCE, ERMAP, DARC, CD55, CR1, GYPC, GYPA, GCNT2, RHAG, C1GALT1, AQP1, KEL, AQP3, ABO, CD44, ART4* (also known as *DO*), *SEMA7A, SLC4A1, SLC14A1, FUT3, BCAM* (also known as *LU*), *FUT2, FUT1, A4GALT, XG*, and *XK*.

For all 262 BGA SNPs in the data set we calculated Kendall's rank correlation coefficient (τ) between pathogen richness and allele frequencies in HGDP-CEPH populations; a normal approximation with continuity correction to account for ties was used for *P*-value calculations (Kendall 1976). We verified that, after Bonferroni correction for multiple tests, 26 BGA gene SNPs were significantly associated with pathogen richness (Table 1). Since variables different from selective forces (e.g., colonization routes; Handley et al. 2007) are expected to affect allele frequency spectra across populations, we compared the strength of BGA gene SNP correlations to control sets of SNPs extracted from the data set. In particular, for each BGA SNP in Table 1 we extracted from the full data set all SNPs having an overall minor allele frequency (averaged over all populations) differing less than 0.01 from its frequency; for all SNPs in the 26 frequency-matched groups we calculated Kendall's τ between pathogen richness and allele frequencies. Next, we calculated the percentile rank of BGA gene SNPs in the distribution of Kendall's τ obtained for the control sets and in the distribution of all SNPs in the data set. Data are reported in Table 1 and indicate that all SNPs ranked above the 90th percentile of τ-values, with 19 of them ranking above the 95th (data for all 262 SNPs are available in Supplemental File 1). By performing 30,000 simulations using samples of 262 SNPs we verified that the probability of obtaining 19 SNPs with a correlation value above the 95th percentile amounted to 0.045; a similar result is obtained by considering the probability to obtain *n* SNPs with a τ higher than the 95th percentile in sample of 262 to be Poisson-distributed ($P = 0.043$). These data therefore indicate that the fraction of BGA SNPs that correlate with pathogen richness is higher than expected; yet, these calculations also suggest that a portion of SNPs in Table 1 might represent false positive associations in that the retrieval of 13 variants with a percentile rank above the 95th would be expected by chance. An estimation of the magnitude of selective effects exerted by pathogens on human genes would be required to accurately estimate the expected fraction of truly correlated SNPs.

Table 1. BGA gene SNPs significantly associated with pathogen richness

SNP	Gene	τ[a]	*P* (Bonferroni)	Rank[b] (all)	Rank[c] (matched)
rs11210729	*ERMAP*	0.446	0.002	0.945	0.937
rs6700168	*CD55*	0.427	0.005	0.923	0.927
rs4143022	*GYPC*	−0.440	0.003	0.938	0.930
rs7589096	*GYPC*	0.548	0.000	0.997	0.997
rs4663038	*GYPC*	−0.460	0.001	0.958	0.961
rs17741574	*GYPC*	0.459	0.002	0.957	0.953
rs13034269	*GYPC*	0.493	0.001	0.981	0.978
rs6568	*GYPC*	0.417	0.009	0.910	0.911
rs10487590	*C1GALT1*	0.549	0.000	0.997	0.997
rs9466910	*GCNT2*	−0.491	0.001	0.979	0.975
rs9466912	*GCNT2*	−0.491	0.001	0.979	0.975
rs17576994	*GCNT2*	0.476	0.001	0.970	0.972
rs2228332	*AQP3*	−0.459	0.001	0.957	0.957
rs2073824	*ABO*	−0.506	0.000	0.987	0.988
rs2421826	*CD44*	−0.436	0.004	0.933	0.929
rs1547059	*CD44*	0.439	0.006	0.937	0.933
rs2072081	*SLC4A1*	0.487	0.000	0.978	0.979
rs2074108	*SLC4A1*	0.473	0.001	0.968	0.970
rs692899	*SLC14A1*	0.425	0.007	0.920	0.916
rs10853535	*SLC14A1*	−0.509	0.000	0.988	0.985
rs566309	*SLC14A1*	0.461	0.005	0.958	0.951
rs900971	*SLC14A1*	−0.611	0.000	1.000	1.000
rs6507641	*SLC14A1*	−0.466	0.001	0.962	0.963
rs602662	*FUT2*	−0.499	0.000	0.984	0.980
rs485186	*FUT2*	−0.513	0.000	0.989	0.987
rs504963	*FUT2*	0.472	0.001	0.967	0.965

[a]Kendall's correlation coefficient.
[b]Percentile rank relative to the distribution of all SNPs.
[c]Percentile rank relative to the distribution of allele frequency-matched SNPs.

The strongest correlation between BGA SNP allele frequency and pathogen richness was obtained for rs900971 in *SLC14A1* (Fig. 1; similar representations for all SNPs in Table 1 are available as Supplemental Fig. 1). In order to verify that environmental variables correlating with pathogen richness (Guernier et al. 2004) did not determine the association signal with BGA genes, we calculated the mean temperature and maximum precipitation rate for geographic locations corresponding to HGDP-CEPH populations; none of the SNPs reported in Table 1 significantly correlated with either variable (data not shown).

The identification of correlations between specific environmental variables and allele frequencies has been regarded as a strategy complementary to common population genetic approaches for the detection of selection signatures (Hancock et al. 2008). All such analyses rely on the assumption that the environmental variable we measure nowadays has changed little over human history and that gene flow due to recent admixture has had a minor impact on human genetic diversity. In this case, we implicitly assumed that the number of different pathogen species per country has been maintained proportionally unchanged along human evolutionary history. Although an oversimplification, this might not be so different from the reality, given that climatic variables have been shown to be of primary importance in driving the distribution of human pathogens (Guernier et al. 2004). As for gene flow, the influence of recent admixture in most populations is considered to be modest (Li et al. 2008), as also demonstrated by the good relationship between population genetic diversity and distance from Africa (Handley et al. 2007).

Our data therefore indicate that the allele frequencies of a subset of BGA genes vary with pathogen richness, supporting the vision whereby these loci affect the susceptibility to infectious diseases. This hypothesis had previously been formulated for *ABO* and *FUT2* (Greenwell, 1997; Hill, 2006; Casanova and Abel 2007), while in the case of *GYPC, DARC*, and *SLC4A1* the ability of specific alleles to modulate infection susceptibility has been demonstrated for malaria (Moulds and Moulds 2000). Also, in the case of *AQP3*, modulation of malaria severity can be hypothesized since *AQP3* represents the major channel for glycerol transport in

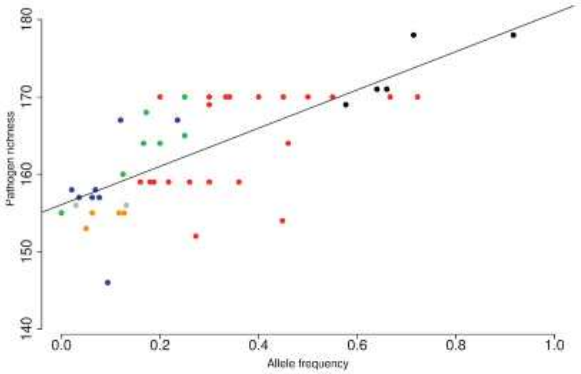


Figure 1. Correlation between pathogen richness and allele frequency for rs900971 in *SLC14A1*. Populations from different broad geographic areas are coded by different colors: (Green) Sub-Saharan Africa, (black) America, (red) Asia, (blue) Europe, (orange) Middle East, and (gray) Oceania.

human erythrocytes (Roudier et al. 2002) and mice knockout for *Aqp9*, a related glycerol transporting aquaporin, display increased survival to *P. berghei* (Liu et al. 2007). One possibility to explain the observed associations is that pathogen richness has co-varied with malaria prevalence and that nucleotide variations, even different from those previously reported to confer resistance in *SLC4A1* and *GYPC*, affect *Plasmodium* entry, spread, or rosetting. Conversely, no association with infectious disease predisposition has ever been reported for *SLC14A1*, *ERMAP*, *C1GALT1*, and *GCNT2*; yet these genes encode either glycotransferases or surface glycosylated proteins, suggesting that carbohydrate determinants might affect pathogen attachment and entry.

Another possibility involving *SLC14A1* variants is that the association with pathogen richness reflects some important aspect of urea metabolism during infection; indeed the gene codes for an urea transporter and the intracellular availability of urea has been shown to be a limiting factor for the ability of *Mycobacterim bovis* to attenuate expression of MHC class II molecules during macrophage infection through urease-induced alkalinization of intracellular compartments (Sendide et al. 2004).

As for *CD44*, it has been shown to act as a receptor for group A *Streptococcus* (Cywes et al. 2000), *Mycobacterium tuberculosis* (Leemans et al. 2003), and *Escherichia coli* in urinary tract infections (Rouschop et al. 2006).

Population genetics analysis of BGA genes

Given the results obtained above and the premise whereby BGA genes might represent selective targets, we wished to verify whether selection signatures could be identified at BGA genes. To this aim we exploited the fact that 22 out of 38 loci involved in BGA specification have been included in the SeattleSNPs program so that resequencing data (although with some gaps) in at least two populations are available; in particular, all data refer to one population with European ancestry (EA) and one with African ancestry (either Yorubans [YRI] or African American [AA]). From the SeattleSNP gene list we excluded *ABO*, which has been previously studied (Saitou and Yamamoto 1997; Calafell et al. 2008), *A4GALT*, *XK*, and *ART4* due to poor resequencing coverage, and *KEL*, because of its being located in a region subjected to a selective sweep possibly driven by the nearby *TRPV6* locus (Akey et al. 2006). The following genes were left for analysis: *AQP1*, *AQP3*, *ACHE*, *BSG*, *B3GALNT1* (previously *B3GALT3*), *CD55* (previously *DAF*), *CD151*, *SLC4A1*, *ICAM4*, *FUT3*, *FUT2*, *FUT1*, *BCAM*, *ERMAP*, *GYPC*, *SEMA7A*, and *SLC14A1*.

With the aim of identifying loci that have been subjected to natural selection, and following the conundrum whereby selection signatures might extend over relatively short gene regions (due to the action of mutation and recombination; Wiuf et al. 2004; Bubb et al. 2006), we applied a sliding window approach to all BGA genes (except for *ACHE*, due to its small size and *FUT2*, as detailed below) and calculated population genetic differentiation, measured as F_{ST}. Under the assumption of neutrality, F_{ST} is determined by demographic history (i.e., genetic drift and gene flow), which affects all loci similarly. We therefore calculated the 2.5th and 97.5th percentiles in the distribution of F_{ST}-values obtained for sliding windows across SeattleSNPs genes (see Methods for details) and searched for BGA gene regions that display unusually high or low population differentiation. Overall, 8.3% of sliding windows deriving from the 17 BGA genes displayed exceedingly high or low F_{ST}-values; estimation of an empirical probability (see Methods) to obtain an equal or higher fraction of outliers in windows deriving from SeattleSNP genes yielded a *P*-value of 0.19. These data indicate that an excess of unusual F_{ST}-values can be observed for BGA genes, with the failure to reach statistical significance being likely due to the presence of other non-neutrally evolving genes in the SeattleSNP data set (which mainly gathers genes involved in inflammatory processes).

BGA gene regions displaying unusual F_{ST}-values were further studied by application of population genetics statistics. In particular, widely used test include Tajima's *D* (Tajima 1989) and Fu and Li's *D** and *F** (Fu and Li 1993). Tajima's *D* (D_T) tests the departure from neutrality by comparing two nucleotide diversity indexes: θ_W (Watterson 1975), an estimate of the expected per site heterozygosity, and π (Nei and Li 1979), the average number of pairwise sequence nucleotide differences. Positive values of D_T indicate an excess of intermediate frequency variants and are a hallmark of balancing selection; negative D_T-values indicate either purifying selection or a high representation of rare variants as a result of a selective sweep. Fu and Li's *F** and *D** are also based on SNP frequency spectra and differ from D_T in that they also take into account whether mutations occur in external or internal branches of a genealogy. Since population history, in addition to selective processes, is known to affect frequency spectra and all related statistics; we performed coalescent simulations using a calibrated population genetics model that incorporates demographic scenarios (Schaffner et al. 2005). Also, in order to disentangle the effects of selection and population history, we exploited the conundrum whereby selection acts on a single locus while demography affects the whole genome: as a control data set we therefore calculated diversity parameters and test statistics for 5 kb windows deriving from 238 genes resequenced by the NIEHS program (see Methods for details). A similar comparison with

SeattleSNPs gene data is provided in Supplementary File 2 (Supplemental Table 2). Sliding window analyses identified gene regions showing unusual F_{ST}-values (Supplemental Fig. 3), which were selected for further study, as reported in the following paragraphs. In addition, for the remaining BGA loci we calculated summary statistics (D_T, D^*, and F^*) for the entire gene region and unusual values were found for *BSG* (detailed below) and *FUT1*. The latter was not further analyzed as high homology with other gene family members and a pseudogene suggested that gene conversion events might affect the result (this was not the case for *FUT2* since we focused on the promoter region, as detailed below).

CD55 (Cromer system)

A sliding-window analysis along the *CD55* gene (OMIM no. +125240; referred to as *DAF* in the SeattleSNPs database) revealed the presence of a region encompassing nucleotides ~9000–19,000 showing exceedingly low F_{ST}-values (Supplemental Fig. 3). Both nucleotide diversity estimates and test statistics (Table 2) revealed no significant departure from neutrality for both, AA and EA, yet D_T and Fu and Li's D^* and F^* ranked relatively high in the distribution of 5 kb windows from NIEHS genes (see Supplemental Table 2 for a comparison with SeattleSNPs genes).

Under neutral evolution, the amount of within-species diversity is predicted to correlate with levels of between-species divergence, since both depend on the neutral mutation rate (Kimura 1983). The HKA test (Hudson et al. 1987) is commonly used to verify whether this expectation is verified. Here we performed a maximum likelihood HKA test (MLHKA) by comparing the *CD55* region to 16 neutrally evolving genes (see Methods for details): a significant result was obtained for both AA and EA (Table 3).

Therefore, we wished to study the genealogy of *CD55* haplotypes in the region and to this aim a neighbor-joining network was constructed. Two major clades separated by long branch lengths are evident (Fig. 2), each containing common haplotypes. In order to estimate the TMRCA (time to the most recent common ancestor) of the two haplotype clades, we applied a phylogeny-based method (Bandelt et al. 1999) based on the measure ρ, the average pairwise difference between the two haplotype clusters. ρ resulted in a value equal to 13.28, so that, using a mutation rate based on 50 fixed differences between chimpanzee and humans, and a separation time of 6 million years (Myr) (Glazko and Nei 2003), we estimated a TMRCA of 3.19 Myr (SD = 673 Kyr). Given the low recombination rate in the region, we wished to verify this result using GENETREE, which is based on a maximum-likelihood coalescent analysis (Griffiths and Tavare 1994, 1995). The method assumes an infinite-site model without recombination and, therefore, haplotypes and sites that violate these assumptions need to be removed: in this case, only three single segregating sites had to be removed. The resulting gene tree, rooted using the chimpanzee sequence, is partitioned into two deep branches (Supplemental Fig. 4). A maximum-likelihood estimate of θ (θ_{ML}) of 9.2 was obtained, resulting in an estimated effective population size (N_e) of 22,000, a value comparable to most figures reported in the literature (Tishkoff and Verrelli 2003). Using this method, the TMRCA of the *CD55* haplotype lineages amounted to 2.61 Myr (SD = 552 Kyr). Such deep coalescent time is unusual, as estimates for neutrally evolving autosomal loci range between 0.8 Myr and 1.5 Myr (Tishkoff and Verrelli 2003).

Overall these data strongly support the idea that the *CD55* region we analyzed has evolved under long-standing balancing selection. This gene portion covers roughly 10 kb surrounding exon 6–7 and contains four DNase I hypersensitive sites in $CD4^+$ T cells (Boyle et al. 2008); five intermediate frequency SNPs (rs6700079, rs2184476, rs1507760, rs10746462, and rs10746463) located along the branch separating the two haplogroups lie within DNase

Table 2. Summary statistics for selected BGA regions

							D			*D**			*F**		
Gene	L[a]	P[b]	N[c]	S[d]	θ^e	π^f		P[g]	Rank[h]		P[g]	Rank[h]		P[g]	Rank[h]
CD55	9.5	AA	48	48	11.39	12.36	0.30	0.11	0.88	−0.52	0.46	0.45	−0.27	0.29	0.56
		EA	46	34	8.14	10.63	1.04	0.14	0.82	0.68	0.16	0.77	0.96	0.11	0.81
CD151	1.6	YRI	48	15	21.12	30.70	1.40	0.022	0.99	0.66	0.15	0.88	1.08	0.048	0.96
		AA	46	14	19.91	35.94	2.48	0.0014	>0.99	0.59	0.18	0.87	1.44	0.015	>0.99
		EA	46	13	18.49	26.01	1.24	0.11	0.85	0.51	0.31	0.71	0.89	0.19	0.78
		AS	40	12	17.63	10.74	−1.20	0.095	0.15	−3.11	0.0067	0.029	−2.94	0.013	0.043
FUT2 pm[i]	5.8	YRI	48	40	15.58	30.06	3.18	<0.0001	>0.99	1.29	0.0054	0.98	2.34	<0.0001	>0.99
		EA	46	8	3.15	1.63	−1.38	0.046	0.13	−0.80	0.27	0.27	−1.14	0.17	0.24
FUT2 cds[j]	3	YRI	48	26	19.52	29.94	1.76	0.0033	>0.99	−0.60	0.55	0.42	0.27	0.18	0.77
		EA	46	20	15.16	24.88	2.07	0.018	0.97	0.60	0.23	0.73	1.30	0.068	0.92
SLC14A1	9.9	YRI	48	86	19.57	18.52	−0.19	0.26	0.64	0.61	0.045	0.86	0.37	0.073	0.80
		AA	48	87	19.80	22.89	0.55	0.051	0.94	1.58	<0.0001	>0.99	1.42	0.0005	0.99
		EA	46	62	14.25	21.97	1.91	0.015	0.96	0.89	0.087	0.83	1.50	0.022	0.95
		AS	40	57	13.54	24.19	2.82	0.0007	>0.99	1.51	0.0074	>0.99	2.34	0.0006	>0.99
BSG	11.3	YRI	48	80	15.95	16.45	0.11	0.24	0.83	0.38	0.24	0.84	0.34	0.22	0.84
		EA	46	81	16.31	11.91	−0.96	0.18	0.18	−2.08	0.052	0.071	−2.00	0.054	0.076

[a]Length of analyzed resequenced region (kb).
[b]Population.
[c]Sample size.
[d]Number of segregating sites.
[e]θ_W estimation per site ($\times 10^{-4}$).
[f]π estimation per site ($\times 10^{-4}$).
[g]*P*-values obtained by applying a calibrated population genetics model, as described in the text.
[h]Percentile rank relative to the distribution of 5 kb deriving from 238 NIEHS genes.
[i]Promoter region.
[j]Coding region.

Table 3. MLHKA test results for BGA regions

Gene	Population	MLHKA	
		k	*P*-value
CD55	AA	3.56	0.0013
	EA	3.25	0.0035
CD151	YRI	3.65	0.0057
	AA	3.08	0.015
	EA	4.46	0.0031
	AS	4.30	0.0042
FUT2 pm	YRI	1.86	0.098
	EA	0.43	0.017
SLC14A1	YRI	2.68	0.0044
	AA	2.79	0.0018
	EA	2.51	0.0066
	AS	2.84	0.0059

Note: ^{k}=Selection parameter

I hypersensitive sites. Since DNase I hyperaccessibilty is thought to be a hallmark of active *cis*-regulatory regions (Gross and Garrard 1988; Felsenfeld and Groudine 2003), these variants might represent good candidates as functional SNPs with a role in transcriptional regulation of *CD55*. Importantly, another variant (rs6700168) located in this genomic portion was found to correlate with pathogen richness (Table 1) and it lies along the branch separating the two haplotype clusters (Fig. 2). In order to verify whether heterozygote advantage might underlie the action of balancing selection we calculated the observed over expected heterozygosity for rs6700168 and verified whether this ratio varied with pathogen richness. Since this was not the case, we suggest that the maintenance of the two haplotype lineages is not due to overdominance but possibly to antagonistic selection (see below).

CD55 (also known as *DAF*, decay-accelerating factor) is a complement-regulatory protein expressed by most cell types, which protects host tissues from damage by the autologous complement system (Nicholson-Weller and Wany 1994). Previous studies have indicated that the membrane-anchored form of *CD55* serves as a receptor for very common human pathogens, such as Dr^{+} *E. coli* (Nowicki et al. 1993), coxsackieviruses B1, B3, and B5 (Shafren et al. 1995), and echovirus 7 (Clarkson et al. 1995), suggesting that decreased or abolished *DAF* expression might confer decrease susceptibility to these infectious agents. Total absence of *CD55* (Inab phenotype) is very rare in humans (Blumenfeld and Patnaik 2004) and associates with no overt phenotype. Yet, other observations point to a possible role of the gene in fertility and pregnancy: *CD55* is dynamically regulated during the menstrual cycle (Young et al. 2002) and it is highly expressed at the feto–maternal interface (Sood et al. 2006); moreover, reduced *DAF* expression has been associated with luteal phase defect of the endometrium associated with infertility or pregnancy loss. Also, mice lacking *DAF* are more susceptible to autoimmune manifestations (Kaul et al. 1995).

These evidences might therefore suggest that regulation of *CD55* expression levels, either in a cell-type- or stage-dependent fashion might affect vital processes, such as reproduction and immunity. Also, the lack of evidence supporting heterozygote advantage and the phenotype of $cd55^{-/-}$ mice possibly suggest that balancing selection ensues from antagonist selection trading-off resistance to infection with autoimmune phenomena. Obviously, other hypotheses are possible (e.g., adaptation to variable environmental conditions with special reference to different environmental pathogens) and further studies on the biological function of *CD55* will be instrumental in addressing this issue.

CD151 (RAPH system)

A sliding window analysis along *CD151* (OMIM no. *602243) indicated the 3′ gene region displays reduced population differentiation and exceedingly low F_{ST}-values are observed in a region roughly corresponding to the terminal region extending from exon 6 to the 3′ UTR (Supplemental Fig. 3). Nucleotide diversity (both θ_W and π), in this restricted region, ranked above the 97.5th percentile in the distribution of 5 kb windows deriving from NIEHS genes (Supplemental Table 1) for both EA and YRI. Summary statistics revealed significantly positive values for D_l, D^*, and F^* in YRI, but not in EA. In order to further investigate the possible departure from neutrality in other human populations, the same region was resequenced in two additional samples: Asians (AS) and AA. As shown in Table 2, significantly positive test statistics were obtained for populations of African ancestry but not for Asians and Europeans. In the case of AS, negative values of summary statistics are due to the presence of a single highly divergent haplotype (Fig. 3).

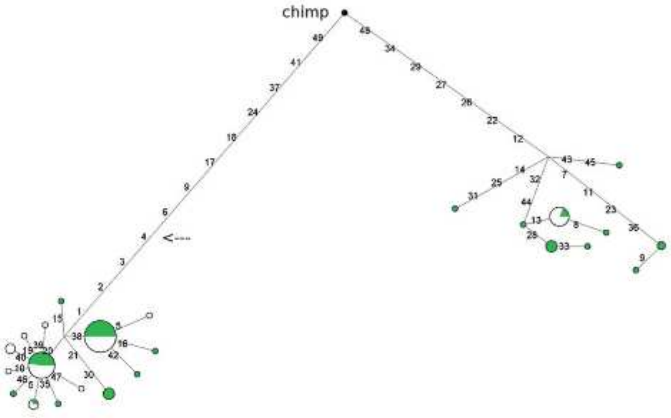


Figure 2. Genealogy of *CD55* haplotypes reconstructed through a median-joining network. The analysis corresponds to the gene region spanning nucleotides ~9500–18,300 (as described in the text). Each node represents a different haplotype, with the size of the circle proportional to the haplotype frequency. Nucleotide differences between haplotypes are indicated on the branches of the network. Circles are color-coded according to population (green, AA; white, EA). The chimpanzee sequence is also shown. The arrow shows the position of rs6700168 (Table 1). Note that the relative position of mutations along a branch is arbitrary.

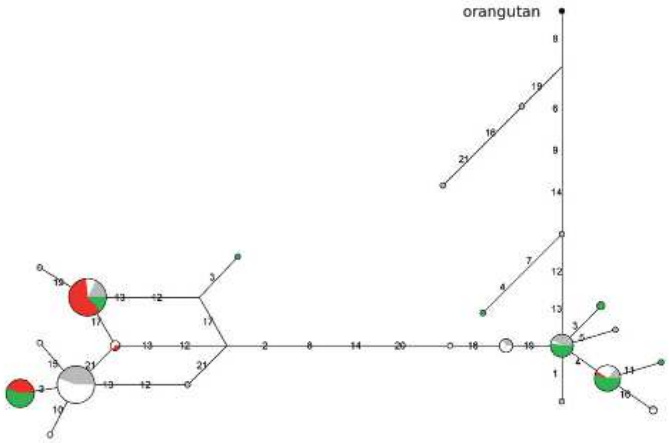


Figure 3. Genealogy of *CD151* haplotypes reconstructed through a median-joining network. The analysis corresponds to the gene region spanning nucleotides 9100–10,400. Population color codes are as follows: (green) AA; (white) EA; (red) AS; (gray) YRI.

Application of the MLHKA test, using 16 neutrally evolving genes (see Methods), rejected the hypothesis of neutrality for all populations (Table 3).

Next, we wished to examine haplotype genealogy for the terminal *CD151* gene region and, to this aim, a median-joining network was constructed (Fig. 3). The topology of this network was relatively unambiguous showing two major clades, each containing common haplotypes, separated by long branch lengths. Calculation of the TMRCA ($\rho = 8.55$; fixed differences = 58 using *Pongo pygmaeus*) yielded an estimate of 3.83 Myr (SD = 1.06 Myr), again a deep coalescent time compared to neutral loci. As reported above for *CD55*, this result was verified using GENETREE and an estimated TMRCA of 2.14 Myr was obtained (SD = 643 Kyr, $\theta_{ML} = 1.9$, $N_e = 13{,}722$; Supplemental Fig. 5). In analogy to *CD55*, these features suggest the action of long-standing balancing selection in African populations.

The analyzed *CD151* region covers the last four coding exons and the 3′ UTR; most variants are located in noncoding regions, with the majority of intermediate frequency SNPs falling within the UTR. Analysis of known functional elements in the 3′ UTR was performed using UTRscan and no SNPs were found to affect predicted motifs. Conversely, a search for microRNA target sites (miRBase) indicated that one variant, namely rs1130698, falls within highest scoring predicted target site. In particular the T allele changes a G–C pairing between the *CD151* UTR sequence and hsa-miR-940 to a G–U wobble; unfortunately, little is known about the expression pattern of hsa-miR-940, except for the fact that it was cloned from cervical cell lines (Lui et al. 2007).

CD151, a member of the tetraspanin protein family involved in cell adhesion and motility, is expressed in most human tissues (Fitter et al. 1995). Mutations of *CD151* in humans result in nephropathy with epidermolysis bullosa and deafness (Karamatic Crew et al. 2004), while different phenotypes have been reported for *cd151*$^{-/-}$ mice, including abnormal hemostasis (Wright et al. 2004), defective wound healing (Cowin et al. 2006), and renal defects (Sachs et al. 2006). Members of the tetraspanin family have been implicated in virus infection in animals and humans; in particular, different tetraspanins have been shown to act as receptors for HCV (Pileri et al. 1998), HIV (von Lindern et al. 2003), canine distemper virus (Loffler et al. 1997), feline leukemia virus (Willett et al. 1994), and porcine reproductive and respiratory syndrome virus (Shanmukhappa et al. 2007); yet a recent report has also shown that members of the tetraspanin family, including *CD151*, protect human macrophages from HIV-1 and vescicular stomatitis virus infection, possibly by blocking virion binding/uptake (Ho et al. 2006).

Whether the maintenance of balancing selection at the *CD151* locus is pathogen-driven remains to be elucidated, and unfortunately no SNP mapping to the gene has been typed in the HGDP-CEPH panel; it is worth noting that besides its possible direct role in predisposing to infections (by acting as a viral receptor/binding factor), its function in wound healing (Cowin et al. 2006) might also be regarded as linked to pathogens and their prevalence, in that the risk of wound infection likely depends on how long the healing process takes to completion.

FUT2 (Lewis system)

In humans two alpha (1,2)-fucosyltransferases, encoded by the paralogous *FUT1* and *FUT2* genes, determine expression of the human H antigen, a precursor of blood group A and B antigens.

The two genes differ in substrate specificities and tissue expression (Costache et al. 1997): *FUT1* (H enzyme, H/h system) is responsible for the expression of H antigen in red cells and vascular endothelia, whereas the *Se* enzyme (encoded by *FUT2*, Lewis system, OMIM no. +182100) is responsible for the synthesis of the same antigen in secretory glands and the intestinal mucosa; individuals referred to as "secretors" (*Se*) have at least one functional *FUT2* allele.

Common *FUT2* null alleles are present in many populations; in particular a frequent null allele (se^{428}) is responsible for most nonsecretor phenotypes in Europe and Africa, while a missense mutation (se^{385}) is widespread in East Asians (Kelly et al. 1995; Koda et al. 1996; Liu et al. 1998). Interestingly, the coding region of *FUT2* has previously been hypothesized to be subjected to balancing selection, possibly under an overdominance regime (Koda et al. 2001).

In the case of *FUT2* we did not perform a sliding window analysis as described for the above genes due to extensive resequencing gaps. Rather, we divided the gene in three major portions: coding exon, intron, and 5′ upstream region (10 kb upstream the transcription start site, thereafter referred to as putative promoter). In line with previous findings, the coding exon displayed high nucleotide diversity and positive statistics (Table 2), while we verified that low levels of nucleotide variation characterize the only intron (not shown). Interestingly, an unusual

pattern was observed at the putative promoter region: as shown in Table 2, YRI displayed high values of θ_W, while EA presented low nucleotide diversity (percentile rank of $\theta_W = 0.18$). Calculation of F_{ST} yielded a high value of 0.45, corresponding to a percentile rank of 0.977 in the distribution of SeattleSNPs gene windows and being 20-fold greater than population differentiation calculated for the coding region ($F_{ST} = 0.022$).

Summary statistics for the putative promoter revealed deviation from neutrality in YRI, since all tests yielded significantly positive values (Table 2); conversely statistics for EA resulted in negative values, although significance was only obtained for D_T. Human/chimpanzee divergence in this gene region amounted to 1.35%, a value higher than the genome average (average = 1.06%, SD = 0.25%; Chimpanzee Sequencing and Analysis Consortium 2005) and greater than that of control loci used in the MLHKA test; the latter gave no significant results for YRI, while a reduction of polymorphism compared to intraspecific divergence was evidenced for EA (Table 3). The greater than average divergence and high polymorphism level observed for YRI might be consistent with the region having low sequence constraints, resulting in an increase of both divergence and diversity; yet, this hypothesis does not fit the EA data whereby low diversity is observed; moreover, the high population differentiation we observed can hardly be reconciled with a neutral pattern of evolution.

Low diversity values and negative statistics are consistent with both purifying and directional selection; Fay and Wu's H (Fay and Wu 2000) is usually applied to distinguish between these possibilities, since significantly negative H-values indicate an excess of high-frequency derived alleles, consistent with directional selection. H equaled −4.66 in EA with a borderline P-value of 0.062 (calculated using the calibrated model). It should be noted that the interpretation of H can be complicated by the fact that the power of this statistic to detect selection is poor when the sweep is relatively old (Przeworski 2002) and population structure can result in significantly negative H statistics (Przeworski 2002).

Reconstruction of haplotype genealogy for the *FUT2* putative promoter using yielded a topology with two major clades separated by long branch lengths (Fig. 4); consistent with the high degree of geographic structure, all European haplotypes cluster with the same haplogroup while African chromosomes are divided in the two clades. Calculation of the TMRCA ($\rho = 13.10$; fixed differences = 79, using chimpanzee) yielded an estimate of 1.99 Myr (SD = 410 Kyr). A similar TMRCA was estimated with the use of GENETREE (TMRCA = 1.70 Myr, SD = 375 Kyr, $\theta_{ML} = 3.4$, $N_e = 8681$; Supplemental Fig. 6). Construction of a haplotype genealogy for the coding region (data not shown) resulted in a TMRCA of 3 Myr, in agreement with previous findings (Koda et al. 2001).

Overall, the data presented above are consistent with the presence of a selected variant/haplotype in the promoter region of *FUT2*; this is in line with a recent report indicating that distinct promoter haplotypes have an effect on the gene transcription levels (Soejima and Koda 2008). In the case of EA, the statistics we performed did not allow a firm rejection of the neutral model; in part this might be due to the small number of SNPs in the region (only eight) which reduces the power of all tests; also, failure to reject neutrality might be accounted for by the pattern being a relic of older selective events.

In the case of YRI, we consider that our observations might be consistent with the presence of a balanced polymorphism. This raises the possibility that the signatures we obtained at the promoter region are due to hitchhiking and linkage disequilibrium (LD) with the coding exon. Nonetheless, different observations suggest that this is not the case. First, calculation of D' between the se^{428} variant and common SNPs in the putative promoter revealed a maximum value of 0.27, indicating low LD, in agreement with a previous report (Soejima and Koda 2008). Second, summary statistics yielded stronger results for the putative promoter compared to the coding exon. Third, although hitchhiking has the potential to affect large genomic regions, the signatures of balancing selection are predicted to extend over relatively short distances (Wiuf et al. 2004; Bubb et al. 2006); as an example, the high nucleotide diversity that characterizes the second exon of MHC loci decays rapidly in flanking intronic sequences (Cereb et al. 1997; Fu et al. 2003) and neighboring exons (Takahata and Sata 1998). This suggests that the departure from neutrality and the high level of nucleotide diversity we observe in the *FUT2* putative promoter region is not merely a result of hitchhiking with the coding exon, given the 7 kb separating the transcription start site from the second exon.

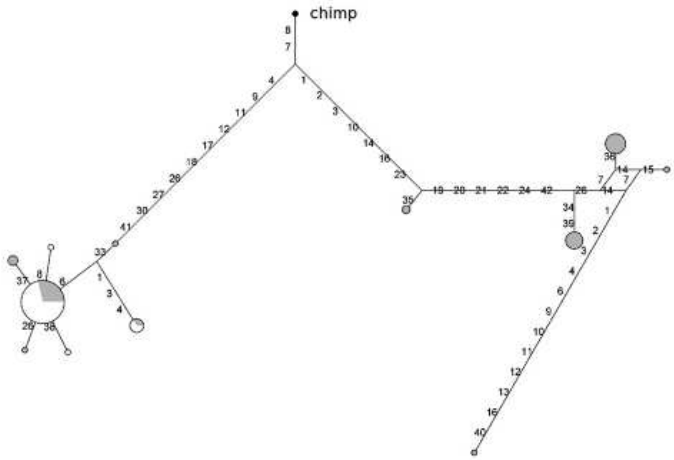


Figure 4. Median-joining network for the putative promoter region of *FUT2*. The analysis corresponds to the gene region spanning nucleotides 7400–17,800. Population color codes are as follows: (white) EA; (gray) YRI.

As reported above, a recent study (Soejima and Koda 2008) of the *FUT2* proximal promoter region indicated that nucleotide diversity patterns differ between African and non-African populations and the authors identified two common haplotypes with different cell-type specific activities. These observations raise the interesting possibility that balancing selection at the *FUT2* promoter region might result from overdominance

due to differential activity of the two promoter haplotypes in different tissues. The "secretor" status has been associated with increased susceptibility to infection by caliciviruses (Lindesmith et al. 2003), HIV (Ali et al. 2000), and respiratory viruses (Raza et al. 1991); yet secretor subjects also display advantages compared to nonsecretors, such as lower susceptibility to urinary tract and *Candida* infections, and increased protection against *Neisseria meningitis* and *Streptococcus* (Haverkorn and Goslings 1969; Blackwell et al. 1990). Also, situations exist where the secretor status might underly a double-faceted situation. One example involves *Campylobacter jejuni* infection (the most common cause of bacterial diarrhea): the pathogen exploits H antigens for tethering to the intestinal mucosa, but at the same time alpha (1,2)-linked fucosyloligosaccharides in human milk inhibit *Campylobacter* infection by competing with intestinal cell surface receptors (Ruiz-Palacios et al. 2003). As a result, a breast-fed infant is expected to be at variable risk of infectious diarrhea depending on his/her intestinal expression of H antigens and his/her mother secretion of the same molecule in milk; in this scenario, maximization of *FUT2* expression in lactating epithelia might be extremely important in providing immunization to newborns. Indeed, different oligosaccharide species in human milk form part of the innate immune system with activity against different pathogens (Newburg et al. 2005) and fucosyloligosaccharides containing alpha (1,2)-linked fucose are prevalent (Chaturvedi et al. 2001). Women who are nonsecretors do not express measurable 2-linked fucosyloligosaccharides and the amount of milk fucosyloligosaccharides varies even among secretors (Chaturvedi et al. 2001), possibly suggesting the presence of genotype differences responsible for such variation (Chaturvedi et al. 2001). Since diarrhea represents a very common cause of mortality in newborns throughout the world, the adaptive significance of decreasing the chance of infection in breast-fed infants is evident. Therefore, maintenance of the advantages conferred by the secretor status, while modulating the levels of glycotransferase activity in a cell-type-dependent fashion, might represent a beneficial strategy in specific circumstances. Obviously, other explanations for the maintenance of different *FUT2* promoter haplotypes are possible, and further studies will be required to analyze the activity of *FUT2* promoter haplotypes. Unfortunately, no SNP located in the putative promoter region of *FUT2* was available to test association with pathogen richness or verify whether a heterozygote excess could be observed in specific geographic locations. Conversely, significantly associated SNPs are located in the coding exon (rs602662 and rs485186) or 3′ UTR (rs504963) and are in full LD with the null se^{428} allele in both EA and AA (in all cases, $D' = 1$, $P < 0.001$). No correlation was observed between the observed/expected heterozygosity ratio for these SNPs and pathogen richness, suggesting that, although pathogens have exerted a selective pressure on the gene and balancing selection has been operating, the underlying explanation is not accounted for by overdominance. This might be expected if the null allele is thought of as the selected variant: heterozygotes are secretors and they are not expected to have an advantage compared to subjects carrying two active alleles. One possibility is that secretors and nonsecretors experience advantages or disadvantages depending on variable environmental conditions in terms of pathogen prevalence, since they display different susceptibility to diverse pathogen types. Alternatively, as suggested above, more complex scenarios can be envisaged that also take into account promoter variants.

SLC14A1 (Kidd system)

Sliding window analysis of *SLC14A1* (OMIM no. *111000) revealed an extended region of about 6.3 kb showing high levels of population differentiation (Supplemental Fig. 3). The single variant (Asp280Asn) responsible for the common *JK*A/JK*B* antigens (Blumenfeld and Patnaik 2004) is located ~6 kb downstream and, with the aim of analyzing the evolutionary history of the gene, we decided to resequence the entire region in YRI and AS with the exception of a small central gap of 2 kb (Supplemental Fig. 3). Two novel nonsynonymous variants were identified, Val10Met and Val76Ile, and both were present in the same three AA subjects.

Summary statistics and diversity parameters for the four populations (Table 2) revealed high levels of polymorphism and allowed rejection of neutrality for AA, AS, and EA, while borderline values were obtained for YRI.

Application of the MLHKA test, as described above, rejected the hypothesis of neutrality for all populations (Table 3). Construction of the median-joining network is recommended when regions displaying low recombination are being analyzed; in the

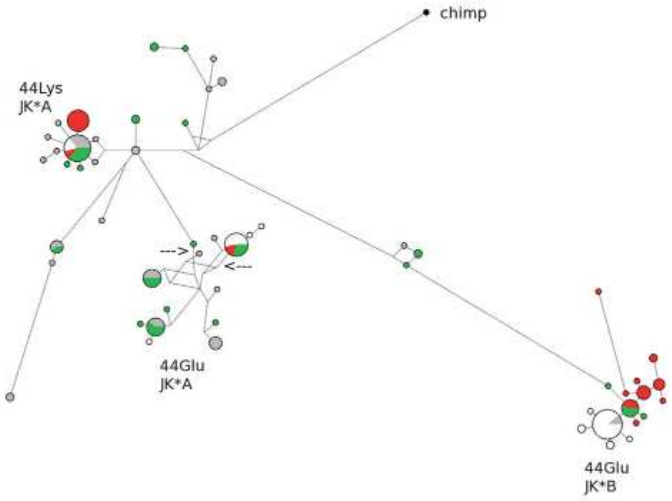


Figure 5. Genealogy of *SLC14A1* haplotypes reconstructed through a median-joining network. The analysis corresponds to the gene region spanning nucleotides 4887–17,350. Population color codes are as follows: (green) AA; (white) EA; (red) AS; (gray) YRI. The allelic status at amino acid position 44 and 280 (*JK*A/JK*B*) is reported for the three major clusters. The two arrows denote the position of rs10853535 and rs692899, which correlate with pathogen richness (Table 1).

case of *SLC14A1*, the gene region carrying the Asp280Asn polymorphism displays low LD with the more 5′ region (Supplemental Fig. 7); yet, we decided to calculate the network over the entire region so that the relative distribution of chromosomes carrying the *JK*A/JK*B* variants could be visualized (Fig. 5); conversely, TMRCA estimate was performed using GENETREE and for this analysis only variants in linkage disequilibrium were included (Supplemental Fig. 7). In both cases, three major haplotype clades are evident and TMRCA estimated equaled 2.28 Myr (SD = 283 Kyr, $\theta_{MI.}$ = 12, N_e = 28,800). The median joining network shows that a long branch separates haplotypes carrying the *JK*B* allele from *JK*A*, while a nonsynonymous Glu44Lys SNP might be regarded as the selected variant maintaining the two closer clusters carrying *JK*A* (Fig. 5). It is interesting to notice that two variants located in this gene region (rs10853535 and rs692899) correlate with pathogen richness (Table 1); one of them lies on the branch leading to the haplotype cluster carrying the *JK*A* and 44Glu alleles, while the second is internal to this same cluster and defines a smaller haplotype group (Fig. 5).

Interestingly, for both these variants the observed over expected heterozygosity ratio significantly correlated with pathogen richness (rs692899, τ = 0.327, P = 0.0012; rs10853535, τ = 0.311, P = 0.0019, Supplemental Fig. 2), possibly suggesting that the two subclades carrying the *JK*A* allele are maintained by overdominance. Conversely, we found no heterozygote excess for rs900971, which showed the strongest correlation with pathogens among all BGA SNPs (Table 1). This variant is located further downstream the Asp280Asn SNP and displays low linkage disequilibrium with the balancing selection region. It is therefore tempting to speculate that different variants in *SLC14A1* have been subjected to pathogen-driven selection under different regimes that might include heterozygote advantage and, possibly, directional selection.

Altogether, the data reported above concur with the idea that multiallelic balancing selection has shaped the evolutionary history of *SLC14A1*, although several issues remain to be clarified. In particular, the possible role of urea metabolism in relation to pathogen resistance has been briefly mentioned above as a possible explanation for selection at this locus, but current knowledge on this issue is too limited to warrant extensive speculation. Moreover, consistent with the biological function of *SLC14A1*, Kidd-null subjects and knockout mice display mild urinary concentrating defects and greater urine output (Sands et al. 1992; Yang et al. 2002). This observation raises the possibility that, together with pathogen-driven selection, the transporter might also have adapted to climatic variables, possibly driven, for example, by the necessity to spare water in hot dry climates. In fact, we did not find any SNP in *SLC14A1* to correlate with climatic variables, such as mean temperature and maximum precipitation rate. Yet, the effect might be confounded by pathogen-driven selection or the power to detect a correlation might vary depending on the environmental variable, as previously suggested (Hancock et al. 2008).

BSG (OK system)

Calculation of nucleotide diversity parameters and summary statistics for the whole *BSG* gene (OMIM no. *109480) revealed an unusual pattern in EA. In both this population and in YRI we observed a θ_W of 16×10^{-4}, a value higher than the 97.5th percentile in EA (Supplemental Table 1). Yet, while in YRI relatively high values for Fu and Li's D^* and F^* were obtained, all statistics were negative in EA with borderline significance (Table 2). Closer examination indicated that the negative statistics in Europeans are due to the presence of a single highly divergent haplotype carrying 24 singletons. We therefore verified whether this haplotype was present in the African sample and identified six additional chromosomes carrying closely related haplotypes. We next constructed a median-joining network of a 2-kb gene region showing low recombination (Supplemental Fig. 8): the topology indicated the presence of two distantly related haplotype clusters (Fig. 6) with an estimated TMRCA of 1.76 Myr (SD = 576 Kyr, ρ = 4.99; fixed differences with chimpanzee = 34). Calculation of the TMRCA using GENETREE resulted in a comparable estimate (TMRCA = 1.53 Myr, SD = 443 Kyr, $\theta_{MI.}$ = 2.5, N_e = 10,714; Supplemental Fig. 8). Such divergent haplotype clades can be expected under two different circumstances, namely balancing selection and ancient population structure. Yet, some difference exists in that symmetric balancing selection is expected to elongate the entire neutral genealogy, while the effects of ancient population structure are reflected in an increase in the genealogical time occupied by two single lineages (Takahata 1990;

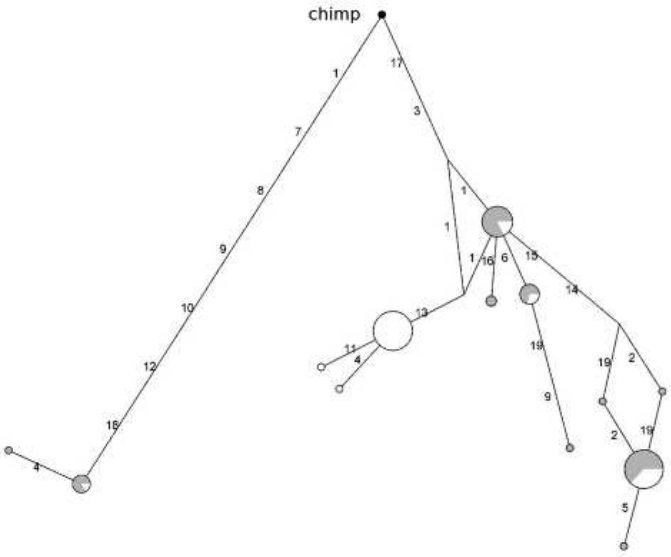


Figure 6. Genealogy of *BSG* haplotypes reconstructed through a median-joining network. The analysis corresponds to the gene region spanning nucleotides 8500–10,300. Population color codes are as follows: (white) EA; (gray) YRI.

Wall 2000). A possibility to discriminate between these scenarios is to calculate the percentage of congruent mutations, meaning those that occur on the basal branches of a genealogy (Wall 2000). When we applied this approach to the two major *BSG* clades, a percentage of congruent mutations equal to 34% was obtained; this is lower than previous estimates under a model of ancient population structure, which ranged from 42% to 45% (Barreiro et al. 2005; Garrigan et al. 2005); also, the TMRCA we estimated for the *BSG* gene is not unusual (Tishkoff and Verrelli 2003; Garrigan and Hammer 2006), while deep coalescent times are expected when ancient population subdivision is involved. The asymmetric structure of the haplotype genealogy whereby most chromosomes cluster in one clade with a relatively deep coalescent, while a minor branch is accounted for by a small number of less diverged chromosomes is difficult to interpret within a theoretical framework. Different explanations might account for the *BSG* genealogy, one appealing possibility being frequency-dependent balancing selection, accounting for the maintenance of a distantly related haplotype with a low frequency in the population. Another possibility is that different selective events have being acting on the *BSG* locus or, else, that complex demographic scenarios account for the pattern we observe nowadays.

Basigin (also known as CD147) has been involved in different biologic and pathologic processes, such as amyloid-beta production, thymocyte maturation, cellular invasion, and rheumatoid arthritis (Iacono et al. 2007). Moreover, functioning as a receptor for cyclophilin A makes CD147 a facilitator of HIV-1 infection (Pushkarsky et al. 2001). This property derives from the ability of HIV-1 to incorporate cyclophilin A into virions, a feature which is common to other viruses (Castro et al. 2003; Lin and Emerman 2006). Additional studies aimed at clarifying the evolutionary history of BSG and its role in infections might benefit from this initial description.

Conclusions

Haldane's hypothesis as formulated in 1932 posits that infectious diseases have been a major threat to human populations and have therefore exerted strong selective pressures throughout human history (Haldane 1932). A few years later he also presciently proposed that antigens constituted of protein-carbohydrates molecules account for "surprising biochemical diversity by serological tests" and possibly play a role in resistance/predisposition to pathogen infection (Haldane 1949). These lines seem to perfectly fit BGA genes, as demonstrated by both this study and previous descriptions (Saitou and Yamamoto 1997; Koda et al. 2001; Baum et al. 2002; Hamblin et al. 2002).

On the one hand, despite medical advances in treatment and prevention, infectious diseases represent a major selective pressure in humans and account for about 48% of deaths in people younger than 45 yr worldwide (Kapp 1999). On the other hand, different BGAs have been shown to act as receptors for one or more pathogens and differential disease susceptibility has been substantiated in some cases depending on BG phenotype. In this scenario, it is not surprising that BGA genes have been the target of selective pressures and associations between pathogen richness and BGA alleles can be identified.

Indeed, here we show that four BGA genes have been subjected to balancing selection (the underlying selective pressure possibly being an infectious agent) and that pathogen richness has shaped allele frequencies in 11 genes. These data, together with previous description of non neutral evolution for *ABO* (Saitou and Yamamoto 1997; Calafell et al. 2008), *FUT2* (Koda et al. 2001), *GYPA* (Baum et al. 2002), and *DARC* (Hamblin et al. 2002), indicate that BGAs played a central role in the host–pathogen arms race during human evolutionary history.

Methods

DNA samples and sequencing

Human genomic DNA was obtained from the Coriell Institute for Medical Research. All analyzed regions were PCR amplified and directly sequenced; primer sequences are available upon request. PCR products were treated with ExoSAP-IT (USB Corporation), directly sequenced on both strands with a Big Dye Terminator sequencing Kit (v3.1 Applied Biosystem) and run on an Applied Biosystems ABI 3130 XL Genetic Analyzer (Applied Biosystem). Sequences were assembled using AutoAssembler version 1.4.0 (Applied Biosystems), inspected manually by two distinct operator and singletons were re-amplified and resequenced.

Data retrieval and haplotype construction

Genotype data for two populations, one of African ancestry and one of Caucasian ancestry, were retrieved from the SeattleSNPs website (http://pga.mbt.washington.edu). Nucleotide positions for all analyzed genes correspond to those of SeattleSNPs, which in turn are derived from the following GenBank accession nos.: AY942196 (*BSG*), AY851161 (*CD55*), DQ074789 (*CD151*), AY937240 (*FUT2*), and AY942197 (*SLC14A1*).

Genotype data for 238 resequenced human genes were derived from the NIEHS SNPs Program website (http://egp.gs.washington.edu). In particular we selected genes that had been resequenced in populations of defined ethnicity including African American (AA), Caucasians (European ancestry, EA), Yoruba (YRI), and Asians (AS) (NIEHS panel 2). Similarly, genotype data from 304 resequenced genes were derived from the SeattleSNPs Web site. In particular, 201 and 103 genes have been resequenced across panels 1 and 2, respectively, the former containing African American and European American, the latter Yoruban and European subjects.

Haplotypes were inferred using PHASE version 2.1 (Stephens et al. 2001; Stephens and Scheet 2005), a program for reconstructing haplotypes from unrelated genotype data through a Bayesian statistical method. Haplotypes for individuals resequenced in this study are available as supplementary material (Supplemental File 3).

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

Data concerning HGDP-CEPH SNPs derive from a previous work (Li et al. 2008). A SNP was ascribed to a specific gene if it was located within the transcribed region or no more than 700 bp upstream the transcription start site.!

Statistical analysis

The correlation between pathogen richness and BGA allele frequencies was assessed by Kendall's rank correlation coefficient (τ), a non-parametric statistic used to measure the degree of correspondence between two rankings. The reason for using this test is that even in the presence of ties, the sampling distribution of τ satisfactorily converges to a normal distribution for values of n larger than 10 (Salkind 2007).

In order to evaluate the probability of obtaining 19 SNPs out of 262 with a τ higher than the 95th percentile, we performed 30,000 simulations. In particular, samples of 262 SNPs were extracted from the full data set by searching for each BGA SNP, one with an allele frequency matched at the 0.001 level. For each sample, the number of SNPs with a percentile rank higher than the 95th percentile (calculated over all SNPs) was counted. By this procedure, the empirical probability of obtaining 19 or more SNPs was estimated equal to 0.045. A theoretic approach can also be applied by considering that the probability to obtain n SNPs with a τ higher than the 95th percentile in sample of 262 is Poisson-distributed with lambda = 13 (5% of 262). For such a distribution the probability of obtaining 19 or more SNPs equals 0.043.

The F_{ST} statistic (Wright 1950) estimates genetic differentiation among populations and was calculated as previously proposed (Hudson et al. 1992). In order to identify gene regions showing extreme F_{ST}-values, sliding windows of 5 kb moving along BGA genes with a step of 150 bp were used; the same procedure was applied to all genes resequenced by the SeattleSNPs program. Values deriving from sliding windows obtained from all genes resequenced in panels 1 and 2 were used to identify the 2.5th and 97.5th percentiles that represented the threshold to define unusually high or low F_{ST}-values in BGA genes. It is worth noting that negative F_{ST} should be interpreted as 0 and the 2.5th percentile value of F_{ST} from SeattleSNPs gene sliding windows resulted extremely close to 0. Therefore, BGA windows displaying an F_{ST}-value negative or equal to 0 were considered to display exceedingly low population differentiation. In order to evaluate the probability of obtaining 8.3% of windows showing F_{ST}-values either below the 2.5th or above the 97.5th percentiles, we used a simulation-based approach. In particular, 17 genes were randomly selected from the SeattleSNPs database and for each group the fraction of sliding windows showing exceedingly low or high values was counted. Ten thousand simulations were performed and the probability of obtaining a fraction of outliers equal to or higher than 8.3% was calculated.

Tajima's D (Tajima 1989), Fu and Li's D^* and F^* (Fu and Li 1993) statistics, as well as diversity parameters θ_W (Watterson 1975) and π (Nei and Li 1979) and Fay and Wu's H (Fay and Wu 2000) were calculated using *libsequence* (Thornton 2003), a C++ class library providing an object-oriented framework for the analysis of molecular population genetic data. Calibrated coalescent simulations were performed using the *cosi* package (Schaffner et al. 2005) and its best-fit parameters for YRI, AA, EA, and AS populations with 10,000 iterations. As a further control, summary statistics were calculated for 5 kb windows deriving from NIEHS genes and the values obtained for BGA gene regions compared to their distribution. In particular, for each gene a 5 kb region was randomly selected; the only requirement was that it did not contain any long (>500 bp) resequencing gap; if the gene did not fulfill this requirement it was discarded, as were 5 kb regions displaying less than five SNPs. The numbers of analyzed windows for AA, YRI, EA, and AS were 209, 203, 177, and 172, respectively. The same procedure was applied to SeattleSNPs genes and a total of 103, 201, and 298 windows were obtained for YRI, AA, and subjects with European ancestry, respectively.

The maximum-likelihood-ratio HKA test was performed using the MLHKA software (Wright and Charlesworth 2004) using multilocus data of 16 genes and *Pan troglodytes* (NCBI panTro2) as an outgroup. The 16 reference genes were randomly selected among NIEHS loci shorter than 20 kb that have been resequenced in the four populations (YRI, AA, EA, and AS; panel 2); the only criterion was that Tajima's D did not suggest the action of natural selection (i.e., D_T is higher than the 2.5th and lower than the 97.5th percentiles in the distribution of NIEHS genes; see Supplemental Table 3). The reference set was accounted for by the following genes: *VNN3, PLA2G2D, MB, MAD2L2, HRAS, CYP17A1, ATOX1, BNIP3, CDC20, NGB, TUBA1, MT3, NUDT1, PRDX5, RETN,* and *JUND*.

We evaluated the likelihood of the model under two different assumptions: that all loci evolved neutrally and that only the region under analysis was subjected to natural selection; statistical significance was assessed by a likelihood ratio test. We used a chain length (the number of cycles of the Markov chain) of 2×10^5 and, as suggested by Wright and Charlesworth (2004), we ran the program several times with different seeds to ensure stability of results.

Median-joining networks to infer haplotype genealogy was constructed using NETWORK 4.5 (Bandelt et al. 1999). Estimate of the time to the most common ancestor (TMRCA) was obtained using a phylogeny based approach implemented in NETWORK using a mutation rate based on the number of fixed differences between human and chimpanzee or orangutan and assuming a separation time from humans of 6 Myr and 13 Myr ago, respectively (Glazko and Nei 2003). In all cases, a second TMRCA estimate derived from application of a maximum-likelihood coalescent method implemented in GENETREE (Griffiths and Tavare 1994, 1995). Again, the mutation rate μ was obtained on the basis of the divergence between human and a primate, assuming a generation time of 25 yr. In using this μ and the estimated maximum likelihood θ (θ_{ML}), we estimated the effective population size parameter (N_e). With these assumptions, the coalescence time, scaled in $2N_e$ units, was converted into years. For the coalescence process, 10^6 simulations were performed. All calculations were performed in the R environment (www.r-project.org).

Environmental variables

Pathogen absence/presence matrices for the 21 countries where HGDP-CEPH populations are located were derived from the Gideon database (http://www.gideononline.com) following previous indications (Prugnolle et al. 2005). Briefly, only species that are transmitted in the countries were included, meaning that cases of transmission due to tourism and immigration were not taken into account; also, species that have recently been eradicated as a result, for example, of vaccination campaigns, were recorded as present in the matrix. It should be noted that the final number of different pathogen species per country differ from those calculated by Prugnolle et al. (2005), since these authors only took into account intracellular disease agents. Precipitation rate and mean temperature were derived for the geographic coordinates corresponding to HGDP-CEPH populations from the NCEP/NCAR database (Kistler et al. 2001).

Sequence annotation

Data concerning DNase I hypersensitive sites in CD4+ T cells derive from a previous work (Boyle et al. 2008) and were retrieved from the UCSC annotation tables (http://genome.ucsc.edu, Duke DNase I HS track). MicroRNA binding sites were identified through the dedicated utility at miRBase, which relies on the miRanda algorithm (John et al. 2004) and requires a target site to be conserved in at least two species. Functional elements in 3′UTR were searched for using UTRscan (Pesole and Liuni 1999).

Acknowledgments

We thank Dr. Roberto Giorda for helpful comments and discussion about the manuscript. M.S. and R.C. are part of the Doctorate School of Molecular Medicine, University of Milan.

References

Akey, J.M., Swanson, W.J., Madeoy, J., Eberle, M., and Shriver, M.D. 2006. TRPV6 exhibits unusual patterns of polymorphism and divergence in worldwide populations. *Hum. Mol. Genet.***15:** *2106–2113.*

Ali, S., Niang, M.A., N'doye, I., Critchlow, C.W., Hawes, S.E., Hill, A.V., and Kiviat, N.B. 2000. Secretor polymorphism and human immunodeficiency virus infection in Senegalese women. *J. Infect. Dis.***181:** *737–739.*

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

Barreiro, L.B., Patin, E., Neyrolles, O., Cann, H.M., Gicquel, B., and Quintana-Murci, L. 2005. The heritage of pathogen pressures and ancient demography in the human innate-immunity CD209/CD209L region. *Am. J. Hum. Genet.***77:** *869–886.*

Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. 2005. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics***21:** *263–265.*

Baum, J., Ward, R.H., and Conway, D.J. 2002. Natural selection on the erythrocyte surface. *Mol. Biol. Evol.***19:** *223–229.*

Blackwell, C.C., Weir, D.M., James, V.S., Todd, W.T., Banatvala, N., Chaudhuri, A.K., Gray, H.G., Thomson, E.J., and Fallon, R.J. 1990. Secretor status, smoking and carriage of *Neisseria meningitidis*. *Epidemiol. Infect.***104:** *203–209.*

Blumenfeld, O.O. and Patnaik, S.K. 2004. Allelic genes of blood group antigens: A source of human mutations and cSNPs documented in the Blood Group Antigen Gene Mutation Database. *Hum. Mutat.***23:** *8–16.*

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., and Crawford, G.E. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell***132:** *311–322.*

Bubb, K.L., Bovee, D., Buckley, D., Haugen, E., Kibukawa, M., Paddock, M., Palmieri, A., Subramanian, S., Zhou, Y., Kaul, R., et al. 2006. Scan of human genome reveals no new loci under ancient balancing selection. *Genetics***173:** *2165–2177.*

Calafell, F., Roubinet, F., Ramírez-Soriano, A., Saitou, N., Bertranpetit, J., and Blancher, A. 2008. Evolutionary dynamics of the human ABO gene. *Hum. Genet.***124:** *123–135.*

Casanova, J.L. and Abel, L. 2007. Human genetics of infectious diseases: A unified theory. *EMBO J.***26:** *915–922.*

Castro, A.P., Carvalho, T.M., Moussatche, N., and Damaso, C.R. 2003. Redistribution of cyclophilin A to viral factories during vaccinia virus infection and its incorporation into mature particles. *J. Virol.***77:** *9052–9068.*

Cereb, N., Hughes, A.L., and Yang, S.Y. 1997. Locus-specific conservation of the HLA class I introns by intra-locus homogenization. *Immunogenetics***47:** *30–36.*

Chaturvedi, P., Warren, C.D., Altaye, M., Morrow, A.L., Ruiz-Palacios, G., Pickering, L.K., and Newburg, D.S. 2001. Fucosylated human milk oligosaccharides vary between individuals and over the course of lactation. *Glycobiology***11:** *365–372.*

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature***437:** *69–87.*

Clarkson, N.A., Kaufman, R., Lublin, D.M., Ward, T., Pipkin, P.A., Minor, P.D., Evans, D.J., and Almond, J.W. 1995. Characterization of the echovirus 7 receptor: Domains of CD55 critical for virus binding. *J. Virol.***69:** *5497–5501.*

Costache, M., Apoil, P.A., Cailleau, A., Elmgren, A., Larson, G., Henry, S., Blancher, A., Iordachescu, D., Oriol, R., and Mollicone, R. 1997. Evolution of fucosyltransferase genes in vertebrates. *J. Biol. Chem.***272:** *29721–29728.*

Cowin, A.J., Adams, D., Geary, S.M., Wright, M.D., Jones, J.C., and Ashman, L.K. 2006. Wound healing is defective in mice lacking tetraspanin CD151. *J. Invest. Dermatol.***126:** *680–689.*

Cywes, C., Stamenkovic, I., and Wessels, M.R. 2000. CD44 as a receptor for colonization of the pharynx by group A Streptococcus. *J. Clin. Invest.***106:** *995–1002.*

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

Felsenfeld, G. and Groudine, M. 2003. Controlling the double helix. *Nature***421:** *448–453.*

Fitter, S., Tetaz, T.J., Berndt, M.C., and Ashman, L.K. 1995. Molecular cloning of cDNA encoding a novel platelet-endothelial cell tetra-span antigen, PETA-3. *Blood***86:** *1348–1355.*

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

Fu, Y., Liu, Z., Lin, J., Chen, W., Jia, Z., Pan, D., and Xu, A. 2003. Extensive polymorphism and different evolutionary patterns of intron 2 were identified in the *HLA-DQB1* gene. *Immunogenetics***54:** *761–766.*

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. 2002. The structure of haplotype blocks in the human genome. *Science***296:** *2225–2229.*

Gagneux, P. and Varki, A. 1999. Evolutionary considerations in relating oligosaccharide diversity to biological function. *Glycobiology***9:** *747–755.*

Garrigan, D. and Hammer, M.F. 2006. Reconstructing human origins in the genomic era. *Nat. Rev. Genet.***7:** *669–680.*

Garrigan, D., Mobasher, Z., Kingan, S.B., Wilder, J.A., and Hammer, M.F. 2005. Deep haplotype divergence and long-range linkage disequilibrium at xp21.1 provide evidence that humans descend from a structured ancestral population. *Genetics***170:** *1849–1856.*

Glazko, G.V. and Nei, M. 2003. Estimation of divergence times for major lineages of primate species. *Mol. Biol. Evol.***20:** *424–434.*

Greenwell, P. 1997. Blood group antigens: Molecules seeking a function? *Glycoconj. J.***14:** *159–173.*

Griffiths, R.C. and Tavare, S. 1994. Sampling theory for neutral alleles in a varying environment. *Philos. Trans. R. Soc. Lond. B Biol. Sci.***344:** *403–410.*

Griffiths, R.C. and Tavare, S. 1995. Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Math. Biosci.***127:** *77–98.*

Gross, D.S. and Garrard, W.T. 1988. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.***57:** *159–197.*

Guernier, V., Hochberg, M.E., and Guegan, J.F. 2004. Ecology drives the worldwide distribution of human diseases. *PLoS Biol.***2:** *e141. doi: 10.1371/journal.pbio.0020141.*

Haldane, J.B.S. 1932. *The causes of evolution*. Longmans, Green & Co., London, UK.

Haldane, J.B.S. 1949. Disease and evolution. Symposium sui fattori ecologici e genetici della speciazione negli animali. In *Selected genetic papers of J.B.S. Haldane* (Anonymous), pp. 325–334. Garland Publishing Inc., New York/London.

Hamblin, M.T., Thompson, E.E., and Di Rienzo, A. 2002. Complex signatures of natural selection at the Duffy blood group locus. *Am. J. Hum. Genet.***70:** *369–383.*

Hancock, A.M., Witonsky, D.B., Gordon, A.S., Eshel, G., Pritchard, J.K., Coop, G., and Di Rienzo, A. 2008. Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genet.***4:** *e32. doi: 10.1371/journal.pgen.0040032.*

Handley, L.J., Manica, A., Goudet, J., and Balloux, F. 2007. Going the distance: Human population genetics in a clinal world. *Trends Genet.***23:** *432–439.*

Haverkorn, M.J. and Goslings, W.R. 1969. Streptococci, ABO blood groups, and secretor status. *Am. J. Hum. Genet.***21:** *360–375.*

Hill, A.V. 2006. Aspects of genetic susceptibility to human infectious diseases. *Annu. Rev. Genet.***40:** *469–486.*

Ho, S.H., Martin, F., Higginbottom, A., Partridge, L.J., Parthasarathy, V., Moseley, G.W., Lopez, P., Cheng-Mayer, C., and Monk, P.N. 2006. Recombinant extracellular domains of tetraspanin proteins are potent inhibitors of the infection of macrophages by human immunodeficiency virus type 1. *J. Virol.***80:** *6487–6496.*

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

Hudson, R.R., Slatkin, M., and Maddison, W.P. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics***132:** *583–589.*

Hurd, E.A. and Domino, S.E. 2004. Increased susceptibility of secretor factor gene *Fut2*-null mice to experimental vaginal candidiasis. *Infect. Immun.***72:** *4279–4281.*

Iacono, K.T., Brown, A.L., Greene, M.I., and Saouaf, S.J. 2007. CD147 immunoglobulin superfamily receptor function and role in pathology. *Exp. Mol. Pathol.***83:** *283–295.*

John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C., and Marks, D.S. 2004. Human microRNA targets. *PLoS Biol.***2:** *e363. doi: 10.1371/journal.pbio.0020363.*

Kapp, C. 1999. WHO warns of microbial threat. *Lancet***353:** *2222. doi: 1016/S0140-6736(05)76281-4.*

Karamatic Crew, V., Burton, N., Kagan, A., Green, C.A., Levene, C., Flinter, F., Brady, R.L., Daniels, G., and Anstee, D.J. 2004. CD151, the first member of the tetraspanin (TM4) superfamily detected on erythrocytes, is essential for the correct assembly of human basement membranes in kidney and skin. *Blood***104:** *2217–2223.*

Kaul, A., Nagamani, M., and Nowicki, B. 1995. Decreased expression of endometrial decay accelerating factor (DAF), a complement regulatory protein, in patients with luteal phase defect. *Am. J. Reprod. Immunol.***34:** *236–240.*

Kelly, R.J., Rouquier, S., Giorgi, D., Lennon, G.G., and Lowe, J.B. 1995. Sequence and expression of a candidate for the human secretor blood group alpha(1,2)fucosyltransferase gene (FUT2). Homozygosity for an enzyme-inactivating nonsense mutation commonly correlates with the non-secretor phenotype. *J. Biol. Chem.***270:** *4640–4649.*

Kendall, M.G. 1976. *Rank correlation methods*. Griffin, London.

Kimura, M. 1983. *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge.

Kistler, R., Kalnay, E., Collins, W., Saha, S., White, G., Woollen, J., Chelliah, M., Ebisuzaki, W., Kanamitsu, M., Kousky, V., et al. 2001. The NCEP-NCAR 50-year reanalysis: Monthly means CD-ROM and documentation. *Bull. Am. Meteorol. Soc.***82:** *247–268.*

Koda, Y., Soejima, M., Liu, Y., and Kimura, H. 1996. Molecular basis for secretor type alpha(1,2)-fucosyltransferase gene deficiency in a Japanese population: A fusion gene generated by unequal crossover responsible for the enzyme deficiency. *Am. J. Hum. Genet.***59:** *343–350.*

Koda, Y., Tachida, H., Pang, H., Liu, Y., Soejima, M., Ghaderi, A.A., Takenaka, O., and Kimura, H. 2001. Contrasting patterns of polymorphisms at the ABO-secretor gene (FUT2) and plasma alpha(1,3)fucosyltransferase gene (FUT6) in human populations. *Genetics***158:** *747–756.*

Leemans, J.C., Florquin, S., Heikens, M., Pals, S.T., van der Neut, R., and Van Der Poll, T. 2003. CD44 is a macrophage binding site for Mycobacterium tuberculosis that mediates macrophage recruitment and protective immunity against tuberculosis. *J. Clin. Invest.***111:** *681–689.*

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. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science***319:** *1100–1104.*

Lin, T.Y. and Emerman, M. 2006. Cyclophilin A interacts with diverse lentiviral capsids. *Retrovirology***3:** *70. doi: 10.1186/1742-4690-3-70.*

Linden, S., Mahdavi, J., Semino-Mora, C., Olsen, C., Carlstedt, I., Boren, T., and Dubois, A. 2008. Role of ABO secretor status in mucosal innate immunity and *H. pylori* infection. *PLoS Pathog.***4:** *e2. doi: 10.1371/journal.ppat.0040002.*

Lindesmith, L., Moe, C., Marionneau, S., Ruvoen, N., Jiang, X., Lindblad, L., Stewart, P., LePendu, J., and Baric, R. 2003. Human susceptibility and resistance to Norwalk virus infection. *Nat. Med.***9:** *548–553.*

Liu, Y., Koda, Y., Soejima, M., Pang, H., Schlaphoff, T., du Toit, E.D., and Kimura, H. 1998. Extensive polymorphism of the *FUT2* gene in an African (Xhosa) population of South Africa. *Hum. Genet.***103:** *204–210.*

Liu, Y., Promeneur, D., Rojek, A., Kumar, N., Frokiaer, J., Nielsen, S., King, L.S., Agre, P., and Carbrey, J.M. 2007. Aquaporin 9 is the major pathway for glycerol uptake by mouse erythrocytes, with implications for malarial virulence. *Proc. Natl. Acad. Sci.***104:** *12560–12564.*

Loffler, S., Lottspeich, E., Lanza, F., Azorsa, D.O., ter Meulen, V., and Schneider-Schaulies, J. 1997. CD9, a tetraspan transmembrane protein, renders cells susceptible to canine distemper virus. *J. Virol.***71:** *42–49.*

Lui, W.O., Pourmand, N., Patterson, B.K., and Fire, A. 2007. Patterns of known and novel small RNAs in human cervical cancer. *Cancer Res.***67:** *6031–6043.*

Moulds, J.M. and Moulds, J.J. 2000. Blood group associations with parasites, bacteria, and viruses. *Transfus. Med. Rev.***14:** *302–311.*

Moulds, J.M., Nowicki, S., Moulds, J.J., and Nowicki, B.J. 1996. Human blood groups: Incidental receptors for viruses and bacteria. *Transfusion***36:** *362–374.*

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

Newburg, D.S., Ruiz-Palacios, G.M., and Morrow, A.L. 2005. Human milk glycans protect infants against enteric pathogens. *Annu. Rev. Nutr.***25:** *37–58.*

Nicholson-Weller, A. and Wang, C.E. 1994. Structure and function of decay accelerating factor CD55. *J. Lab. Clin. Med.***123:** *485–491.*

Nowicki, B., Hart, A., Coyne, K.E., Lublin, D.M., and Nowicki, S. 1993. Short consensus repeat-3 domain of recombinant decay-accelerating factor is recognized by *Escherichia coli* recombinant Dr adhesin in a model of a cell-cell interaction. *J. Exp. Med.***178:** *2115–2121.*

Pesole, G. and Liuni, S. 1999. Internet resources for the functional analysis of 5′ and 3′ untranslated regions of eukaryotic mRNAs. *Trends Genet.***15:** *378. doi: 10.1016/S0168-9525(99)01795-3.*

Pileri, P., Uematsu, Y., Campagnoli, S., Galli, G., Falugi, F., Petracca, R., Weiner, A.J., Houghton, M., Rosa, D., Grandi, G., et al. 1998. Binding of hepatitis C virus to CD81. *Science***282:** *938–941.*

Prugnolle, F., Manica, A., Charpentier, M., Guegan, J.F., Guernier, V., and Balloux, F. 2005. Pathogen-driven selection and worldwide HLA class I diversity. *Curr. Biol.***15:** *1022–1027.*

Przeworski, M. 2002. The signature of positive selection at randomly chosen loci. *Genetics***160:** *1179–1189.*

Pushkarsky, T., Zybarth, G., Dubrovsky, L., Yurchenko, V., Tang, H., Guo, H., Toole, B., Sherry, B., and Bukrinsky, M. 2001. CD147 facilitates HIV-1 infection by interacting with virus-associated cyclophilin A. *Proc. Natl. Acad. Sci.***98:** *6360–6365.*

Raza, M.W., Blackwell, C.C., Molyneaux, P., James, V.S., Ogilvie, M.M., Inglis, J.M., and Weir, D.M. 1991. Association between secretor status and respiratory viral illness. *BMJ***303:** *815–818.*

Reid, M.E. and Lomas-Frances, C. 1997. *The blood group antigen facts book*. Academic, San Diego.

Roudier, N., Bailly, P., Gane, P., Lucien, N., Gobin, R., Cartron, J.P., and Ripoche, P. 2002. Erythroid expression and oligomeric state of the AQP3 protein. *J. Biol. Chem.***277:** *7664–7669.*

Rouschop, K.M., Sylva, M., Teske, G.J., Hoedemaeker, I., Pals, S.T., Weening, J.J., van der Poll, T., and Florquin, S. 2006. Urothelial CD44 facilitates *Escherichia coli* infection of the murine urinary tract. *J. Immunol.***177:** *7225–7232.*

Ruiz-Palacios, G.M., Cervantes, L.E., Ramos, P., Chavez-Munguia, B., and Newburg, D.S. 2003. *Campylobacter jejuni* binds intestinal H(O) antigen (Fuca1, 2Galβ1, 4GlcNAc), and fucosyloligosaccharides of human milk inhibit its binding and infection. *J. Biol. Chem.***278:** *14112–14120.*

Sachs, N., Kreft, M., van den Bergh Weerman, M.A., Beynon, A.J., Peters, T.A., Weening, J.J., and Sonnenberg, A. 2006. Kidney failure in mice lacking the tetraspanin CD151. *J. Cell Biol.***175:** *33–39.*

Saitou, N. and Yamamoto, F. 1997. Evolution of primate ABO blood group genes and their homologous genes. *Mol. Biol. Evol.***14:** *399–411.*

Salkind, N.J. 2007. *Encyclopedia of measurement and statistics*. Sage Publications, Thousand Oaks, CA.

Sands, J.M., Gargus, J.J., Frohlich, O., Gunn, R.B., and Kokko, J.P. 1992. Urinary concentrating ability in patients with Jk(a-b-) blood type who lack carrier-mediated urea transport. *J. Am. Soc. Nephrol.***2:** *1689–1696.*

Schaeffer, A.J., Rajan, N., Cao, Q., Anderson, B.E., Pruden, D.L., Sensibar, J., and Duncan, J.L. 2001. Host pathogenesis in urinary tract infections. *Int. J. Antimicrob. Agents***17:** *245–251.*

Schaffner, S.F., Foo, C., Gabriel, S., Reich, D., Daly, M.J., and Altshuler, D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.***15:** *1576–1583.*

Sendide, K., Deghmane, A.E., Reyrat, J.M., Talal, A., and Hmama, Z. 2004. *Mycobacterium bovis* BCG urease attenuates major histocompatibility complex class II trafficking to the macrophage cell surface. *Infect. Immun.***72:** *4200–4209.*

Shafren, D.R., Bates, R.C., Agrez, M.V., Herd, R.L., Burns, G.F., and Barry, R.D. 1995. Coxsackieviruses B1, B3, and B5 use decay accelerating factor as a receptor for cell attachment. *J. Virol.***69:** *3873–3877.*

Shanmukhappa, K., Kim, J.K., and Kapil, S. 2007. Role of CD151, A tetraspanin, in porcine reproductive and respiratory syndrome virus infection. *Virol. J.***4:** *62. doi: 10.1186/1743-422X-4-62.*

Soejima, M. and Koda, Y. 2008. Distinct single nucleotide polymorphism pattern at the *FUT2* promoter among human populations. *Ann. Hematol.***87:** *19–25.*

Sood, R., Zehnder, J.L., Druzin, M.L., and Brown, P.O. 2006. Gene expression patterns in human placenta. *Proc. Natl. Acad. Sci.***103:** *5478–5483.*

Stephens, M. and 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, N.J., and 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.*

Takahata, N. 1990. A simple genealogical structure of strongly balanced allelic lines and *trans*-species evolution of polymorphism. *Proc. Natl. Acad. Sci.***87:** *2419–2423.*

Takahata, N. and Satta, Y. 1998. Footprints of intragenic recombination at HLA loci. *Immunogenetics***47:** *430–441.*

Thompson, E.E., Kuttab-Boulos, H., Witonsky, D., Yang, L., Roe, B.A., and Di Rienzo, A. 2004. CYP3A variation and the evolution of salt-sensitivity variants. *Am. J. Hum. Genet.***75:** *1059–1069.*

Thornton, K. 2003. libsequence: A C++ class library for evolutionary genetic analysis. *Bioinformatics***19:** *2325–2327.*

Tishkoff, S.A. and Verrelli, B.C. 2003. Patterns of human genetic diversity: Implications for human evolutionary history and disease. *Annu. Rev. Genomics Hum. Genet.***4:** *293–340.*

von Lindern, J.J., Rojo, D., Grovit-Ferbas, K., Yeramian, C., Deng, C., Herbein, G., Ferguson, M.R., Pappas, T.C., Decker, J.M., Singh, A., et al. 2003. Potential role for CD63 in CCR5-mediated human immunodeficiency virus type 1 infection of macrophages. *J. Virol.***77:** *3624–3633.*

Wall, J.D. 2000. Detecting ancient admixture in humans using sequence polymorphism data. *Genetics***154:** *1271–1279.*

Watterson, G.A. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7:** 256–276.
Willett, B.J., Hosie, M.J., Jarrett, O., and Neil, J.C. 1994. Identification of a putative cellular receptor for feline immunodeficiency virus as the feline homologue of CD9. *Immunology* **81:** 228–233.
Wiuf, C., Zhao, K., Innan, H., and Nordborg, M. 2004. The probability and chromosomal extent of *trans*-specific polymorphism. *Genetics* **168:** 2363–2372.
Wright, S. 1950. Genetical structure of populations. *Nature* **166:** 247–249.
Wright, S.I. and Charlesworth, B. 2004. The HKA test revisited: A maximum-likelihood-ratio test of the standard neutral model. *Genetics* **168:** 1071–1076.
Wright, M.D., Geary, S.M., Fitter, S., Moseley, G.W., Lau, L.M., Sheng, K.C., Apostolopoulos, V., Stanley, E.G., Jackson, D.E., and Ashman, L.K. 2004. Characterization of mice lacking the tetraspanin superfamily member CD151. *Mol. Cell. Biol.* **24:** 5978–5988.
Yang, B., Bankir, L., Gillespie, A., Epstein, C.J., and Verkman, A.S. 2002. Urea-selective concentrating defect in transgenic mice lacking urea transporter UT-B. *J. Biol. Chem.* **277:** 10633–10637.
Young, S.L., Lessey, B.A., Fritz, M.A., Meyer, W.R., Murray, M.J., Speckman, P.L., and Nowicki, B.J. 2002. In vivo and in vitro evidence suggest that HB-EGF regulates endometrial expression of human decay-accelerating factor. *J. Clin. Endocrinol. Metab.* **87:** 1368–1375.
Young, J.H., Chang, Y.P., Kim, J.D., Chretien, J.P., Klag, M.J., Levine, M.A., Ruff, C.B., Wang, N.Y., and Chakravarti, A. 2005. Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. *PLoS Genet.* **1:** e82–. doi: 10.1371/journal.pgen.0010082.

Received June 30, 2008; accepted in revised form November 4, 2008.

2.2 Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions

Published May 25, 2009

JEM

ARTICLE

Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions

Matteo Fumagalli,[1,2] Uberto Pozzoli,[1] Rachele Cagliani,[1] Giacomo P. Comi,[3] Stefania Riva,[1] Mario Clerici,[4,5] Nereo Bresolin,[1,3] and Manuela Sironi[1]

[1]Scientific Institute IRCCS E. Medea, Bioinformatic Laboratory, 23842 Bosisio Parini, Italy
[2]Bioengineering Department, Politecnico di Milano, 20133 Milan, Italy
[3]Dino Ferrari Centre, Department of Neurological Sciences, University of Milan, IRCCS Ospedale Maggiore Policlinico, Mangiagalli and Regina Elena Foundation, 20100 Milan, Italy
[4]Department of Biomedical Sciences and Technologies LITA Segrate, University of Milan, 20090 Milan, Italy
[5]Laboratory of Molecular Medicine and Biotechnology, Don C. Gnocchi ONLUS Foundation IRCCS, 20148 Milan, Italy

Many human genes have adapted to the constant threat of exposure to infectious agents; according to the "hygiene hypothesis," lack of exposure to parasites in modern settings results in immune imbalances, augmenting susceptibility to the development of autoimmune and allergic conditions. Here, by estimating the number of pathogen species/genera in a specific geographic location (pathogen richness) for 52 human populations and analyzing 91 interleukin (IL)/IL receptor genes (IL genes), we show that helminths have been a major selective force on a subset of these genes. A population genetics analysis revealed that five IL genes, including *IL7R* and *IL18RAP*, have been a target of balancing selection, a selection process that maintains genetic variability within a population. Previous identification of polymorphisms in some of these loci, and their association with autoimmune conditions, prompted us to investigate the relationship between adaptation and disease. By searching for variants in IL genes identified in genome-wide association studies, we verified that six risk alleles for inflammatory bowel (IBD) or celiac disease are significantly correlated with micropathogen richness. These data support the hygiene hypothesis for IBD and provide a large set of putative targets for susceptibility to helminth infections.

CORRESPONDENCE
Manuela Sironi:
manuela.sironi@BP.LNF.it

Abbreviations used: AA, African American; CD, Crohn's disease; CeD, celiac disease; CNV, copy number variant; D_T, Tajima's D; EU, European; HGDP-CEPH, Human Genome Diversity Project-Centre d'Etude du Polymorphisme Humain; HKA, Hudson, Kreitman, and Aguade; IBD, inflammatory bowel disease; LD, linkage disequilibrium; MLHKA, maximum likelihood HKA; MY, million years; NIEHS, National Institute of Environmental Health Science; SNP, single-nucleotide polymorphism; TMRCA, time to the most recent common ancestor; YRI, Yoruba.

It is commonly believed that infectious diseases have represented one of the major threats to human populations and have therefore acted as a powerful selective force. Even today, despite the advances in treatment and prevention, infectious diseases account for ~48% of worldwide deaths among people >45 yr of age (1). These figures do not include the heavy burden imposed by helminth infections, which have recently been designated as the "great neglected tropical diseases" (2). With an estimated 2 billion individuals infected worldwide (3), helminths represent the prevalent chronic infectious diseases of humans. Although parasitic worms determine severe clinical symptoms in a minority of heavily infected individuals, even apparently subclinical parasite burdens can result in impaired nutritional status and growth retardation (4, 5). It is therefore conceivable that several human genes have evolved in response to both microbial/viral infectious agents and parasitic worms; indeed, it has been suggested (6, 7) that human populations may have adapted to parasites to such a degree that the lower exposure to infectious agents in modern developed societies results in immune imbalances, with autoimmune and allergic conditions being the outcome.

Genes involved in immunity and inflammation are known to be frequent targets of natural selection; balancing selection, which is thought to be a relatively rare phenomenon in humans, has particularly shaped the evolutionary fate of

M. Fumagalli and U. Pozzoli contributed equally to this paper.

The Rockefeller University Press $30.00
J. Exp. Med. Vol. 206 No. 6 1395–1408
www.jem.org/cgi/doi/10.1084/jem.20082779

Supplemental Material can be found at:
http://jem.rupress.org/cgi/content/full/jem.20082779/DC1

Published May 25, 2009
JEM

genes involved in immune responses (8, 9). Balancing selection is the situation whereby genetic variability is maintained in a population via selection. The best known example in the human genome affects the MHC genes, which are characterized by extreme polymorphism levels. Recently, Prugnolle et al. (10) demonstrated that populations from areas with high pathogen diversity display increased MHC genetic variability, indicating the action of pathogen-driven balancing selection.

Quite obviously, the presence of a functional variant is a prerequisite for selection to act, and the identification of nonneutrally evolving genes has been regarded as a strategy complementary to classical clinical and epidemiological studies to provide insight into the mechanisms of host defense (11). Similarly, analysis of the evolutionary history of genes involved in immune defense might provide novel insights into the delicate balance between efficient response to pathogens and autoimmune/allergic manifestations.

In this study, we focused our attention on a large gene family that includes ILs and their receptors (hereafter referred to as IL genes). ILs are small secreted molecules that regulate most aspects of immune and inflammatory responses and exert their effects through binding to specific receptors expressed on target cells. Various IL genes have been associated with differential susceptibility to specific infections (12, 13), and with an augmented likelihood to develop autoimmune or allergic/atopic diseases (for review see reference [14]). Finally, whereas previous reports have demonstrated nonneutral evolution at single IL genes (15–17), no comprehensive analysis has been performed and no attempt to take into account pathogen richness across different human populations has ever been described.

RESULTS

Pathogen-driven selection acts on IL genes

Pathogen-driven selection is a situation whereby the genetic diversity at a specific locus is influenced by pathogens; this is expected to occur because one or more alleles are associated with the modulation of susceptibility to infectious agents. One way to identify loci or variants subjected to pathogen-driven selection is to search for correlations between genetic variability and pathogen richness (10, 18). The latter is a measure of pathogen diversity, and is basically calculated as the number of different pathogen species/genera in a specific geographic location (see Materials and methods for further details). The choice to use pathogen richness rather than more conventional epidemiological parameters such as prevalence/burden stems from several considerations, as follows: (a) comprehensive data on prevalence are impossible to retrieve for many infections; (b) even when prevalence data are available, they may vary considerably within the same country depending on the surveyed regions, the survey period (e.g., before or after eradication campaigns), the population surveyed (e.g., city dwellers rather than farmers/bushmen/nomads or children rather than adults); (c) the prevalence of specific infections might have changed greatly over recent years as a result of eradication campaigns, and historical prevalence data are rarely available; (d) we were not interested in a single species/genus. As a consequence, the prevalence of all (or at least the most common) pathogen species would be required. Still, prevalence data are difficult to combine; e.g., in endemic regions, individuals can be infected with multiple parasite species (2), and these subjects tend to harbor the most intense infections, possibly because of an additive and/or multiplicative impact on nutrition and organ pathology (19).

To verify whether pathogen-driven selection has been acting on IL genes, we exploited the fact that a set of >650,000 single-nucleotide polymorphisms (SNPs) has been genotyped in 52 populations (Human Genome Diversity Project-Centre d'Etude du Polymorphisme Humain [HGDP-CEPH] panel: http://www.cephb.fr/en/hgdp/) distributed worldwide (Fig. S1) (20). Pathogen richness was evaluated by gathering information on the number of different pathogen species/genera present in different geographical areas of the world from the Gideon database. Specifically, pathogens were divided into two major groups: micro- and macropathogens. Micropathogens include viruses, bacteria, fungi, and protozoa. Macropathogens include insects, arthropods, and helminths; in this group parasitic worms were by far the most abundant class (90% of species/genera), so when we refer to macropathogens we basically mean helminths. Analyses were also separately performed for viruses, bacteria, protozoa, and fungi; data are available as supplemental material (Table S1). After data organization in Gideon, we calculated both micro- and macropathogen richness on a country by country basis (i.e., they represent the number of different micro- and macropathogen species/genera per country; Fig. S1). As expected, we observed that micro- and macropathogen richness strongly correlated across geographic locations (Kendall's rank correlation coefficient = 0.67; $P < 2^{-10}$); this is likely caused by the major impact of climatic factors on the spatial distribution of both pathogen classes (21).

IL genes were retrieved from the HGNC Gene Families/Grouping Nomenclature web site (http://www.genenames.org/genefamily.html). From the resulting 99 genes, *IL3RA* and *IL9R* were removed because they are located on the pseudoautosomal regions of sexual chromosomes; *IL15RB* and *IL11RB* were not analyzed because their sequence and chromosomal locations are not present in public databases. Finally, four *IL9R* pseudogenes were discarded. The remaining 91 genes (Table S2) included most known ILs, their receptors, and receptor-accessory proteins.

A total of 1,052 SNPs in IL genes had been typed in the HGDP-CEPH panel, allowing analysis of all genes except for *IL2*, *IL6RL1*, *IL6STP*, *IL8RBP*, *IL17C*, *IL23A*, *IL28A*, *IL28B*, *IL31*, and *IL34*.

For all 1,052 IL gene SNPs in the dataset, we calculated Kendall's rank correlation coefficient (τ) between allele frequencies in HGDP-CEPH populations and micro- or macropathogen richness; a normal approximation with continuity correction to account for ties was used for p-value calculations. After Bonferroni correction for multiple tests, we observed that 48 and 94 SNPs significantly correlate with micro- and macropathogen richness, respectively, with 32 SNPs correlating with

Published May 25, 2009

Table I. Correlations with micro- and macropathogen richness

SNP	Gene	Micropathogens			Macropathogens		
		τ	P-value (corrected)[a]	Rank[b]	τ	P-value (corrected)[a]	Rank[b]
rs17561	*IL1A*	0.38	0.151	0.854	0.47	0.006	0.937
rs1143634	*IL1B*	0.55	<0.0001	0.996	0.58	<0.0001	0.996
rs6761276	*IL1F10*	−0.31	2.011	0.740	−0.51	0.001	0.980
rs10496447	*IL1F8*	0.54	<0.0001	0.995	0.57	<0.0001	0.994
rs3917304	*IL1R1*	−0.53	<0.0001	0.994	−0.51	<0.0001	0.979
rs6444444	*IL1RAP*	−0.55	<0.0001	0.998	−0.50	0.003	0.969
rs6800609	*IL1RAP*	0.39	0.092	0.865	0.47	0.005	0.930
rs2885373	*IL1RAP*	0.55	<0.0001	0.998	0.50	0.003	0.969
rs17196143	*IL1RAP*	−0.49	0.003	0.974	−0.51	0.002	0.966
rs6444435	*IL1RAP*	−0.47	0.003	0.972	−0.54	<0.0001	0.989
rs6630730	*IL1RAPL1*	−0.41	0.079	0.883	−0.53	<0.0001	0.980
rs7052954	*IL1RAPL1*	−0.48	0.009	0.970	−0.52	0.002	0.976
rs6526833	*IL1RAPL1*	0.43	0.020	0.923	0.52	<0.0001	0.982
rs7890572	*IL1RAPL1*	0.44	0.054	0.917	0.50	0.007	0.951
rs10521948	*IL1RAPL1*	0.49	0.003	0.974	0.54	<0.0001	0.988
rs7056388	*IL1RAPL1*	0.41	0.131	0.875	0.51	0.003	0.964
rs196990	*IL1RAPL1*	0.41	0.057	0.899	0.56	<0.0001	0.994
rs7881819	*IL1RAPL1*	0.38	0.406	0.830	0.51	0.003	0.966
rs10521946	*IL1RAPL1*	−0.49	0.003	0.974	−0.62	<0.0001	0.999
rs1318832	*IL1RAPL1*	−0.51	<0.0001	0.987	−0.55	<0.0001	0.995
rs5943559	*IL1RAPL1*	−0.45	0.007	0.955	−0.56	<0.0001	0.995
rs12387961	*IL1RAPL1*	0.37	0.261	0.832	0.46	0.009	0.928
rs721953	*IL1RAPL2*	0.46	0.004	0.967	0.44	0.017	0.918
rs1384360	*IL1RAPL2*	−0.45	0.009	0.952	−0.44	0.018	0.908
rs6621992	*IL1RAPL2*	−0.45	0.010	0.955	−0.46	0.008	0.940
rs7583215	*IL1RL2*	−0.55	<0.0001	0.996	−0.54	<0.0001	0.991
rs2287041	*IL1RL2*	0.51	<0.0001	0.985	0.47	0.004	0.945
rs3771188	*IL1RL2*	−0.54	<0.0001	0.993	−0.52	<0.0001	0.976
rs315931	*IL1RN*	−0.41	0.040	0.916	−0.46	0.007	0.941
rs2637988	*IL1RN*	−0.44	0.011	0.953	−0.47	0.004	0.954
rs2029582	*IL1RN*	−0.30	2.565	0.728	−0.46	0.008	0.936
rs3087266	*IL1RN*	−0.34	0.728	0.780	−0.46	0.008	0.928
rs2386841	*IL2RA*	0.48	0.002	0.970	0.47	0.004	0.942
rs11256497	*IL2RA*	−0.49	0.002	0.973	−0.49	0.002	0.953
rs2284034	*IL2RB*	0.37	0.244	0.834	0.48	0.003	0.950
rs1003694	*IL2RB*	0.32	1.570	0.734	0.46	0.006	0.936
rs228966	*IL2RB*	−0.32	1.705	0.757	−0.48	0.003	0.963
rs2284033	*IL2RB*	0.37	0.245	0.853	0.56	<0.0001	0.996
rs228973	*IL2RB*	0.44	0.012	0.943	0.52	<0.0001	0.980
rs228975	*IL2RB*	−0.49	0.001	0.980	−0.59	<0.0001	0.999
rs2235330	*IL2RB*	0.41	0.059	0.889	0.54	<0.0001	0.987
rs2243268	*IL4*	−0.51	<0.0001	0.987	−0.58	<0.0001	0.997
rs2243288	*IL4*	−0.43	0.021	0.932	−0.47	0.004	0.948
rs2243290	*IL4*	0.49	0.001	0.978	0.54	<0.0001	0.991
rs2070874	*IL4*	−0.50	0.001	0.984	−0.55	<0.0001	0.993
rs3024672	*IL4R*	−0.40	0.189	0.866	−0.53	0.001	0.975
rs3024607	*IL4R*	−0.48	0.006	0.974	−0.53	0.001	0.987
rs17026370	*IL5RA*	0.47	0.011	0.951	0.55	<0.0001	0.990
rs2066992	*IL6*	−0.41	0.049	0.899	−0.46	0.006	0.933

Published May 25, 2009

JEM

Table I. Correlations with micro- and macropathogen richness *(Continued)*

SNP	Gene	Micropathogens			Macropathogens		
rs2069835	*IL6*	−0.46	0.021	0.949	−0.53	0.002	0.982
rs17505589	*IL7*	0.50	0.006	0.986	0.43	0.115	0.851
rs11567697	*IL7R*	0.43	0.086	0.904	0.51	0.005	0.972
rs1554286	*IL10*	−0.45	0.006	0.956	−0.56	<0.0001	0.994
rs3024490	*IL10*	−0.39	0.103	0.880	−0.52	<0.0001	0.988
rs2512144	*IL10RA*	−0.38	0.343	0.832	−0.47	0.009	0.938
rs999261	*IL10RB*	0.39	0.140	0.854	0.49	0.002	0.958
rs2243115	*IL12A*	−0.46	0.008	0.952	−0.43	0.045	0.843
rs17129789	*IL12RB2*	−0.43	0.026	0.910	−0.51	0.001	0.971
rs10521698	*IL13RA2*	−0.40	0.164	0.860	−0.50	0.003	0.956
rs1589241	*IL15*	0.48	0.005	0.970	0.48	0.006	0.950
rs2322262	*IL15*	0.47	0.008	0.962	0.49	0.005	0.953
rs13106911	*IL15*	−0.36	0.361	0.811	−0.46	0.007	0.925
rs8177636	*IL15RA*	0.39	0.179	0.858	0.47	0.006	0.942
rs8177685	*IL15RA*	−0.41	0.057	0.881	−0.47	0.004	0.931
rs3136614	*IL15RA*	−0.47	0.005	0.957	−0.40	0.125	0.798
rs2296139	*IL15RA*	0.44	0.011	0.933	0.53	<0.0001	0.985
rs12437819	*IL16*	−0.50	0.001	0.981	−0.41	0.071	0.846
rs12438640	*IL16*	0.50	0.001	0.982	0.41	0.060	0.854
rs10484879	*IL17A*	−0.49	0.008	0.970	−0.39	0.425	0.770
rs6518661	*IL17RA*	−0.45	0.009	0.944	−0.43	0.026	0.896
rs879576	*IL17RA*	0.45	0.025	0.931	0.50	0.004	0.953
rs999514	*IL17RB*	0.47	0.004	0.964	0.45	0.015	0.916
rs708567	*IL17RC*	−0.40	0.066	0.890	−0.51	0.001	0.976
rs6445854	*IL17RD*	−0.43	0.030	0.912	−0.52	<0.0001	0.982
rs4535195	*IL17RD*	−0.44	0.013	0.942	−0.48	0.003	0.957
rs12487790	*IL17RD*	0.43	0.018	0.935	0.52	<0.0001	0.983
rs17216900	*IL17RD*	0.45	0.009	0.946	0.36	0.368	0.779
rs12496746	*IL17RD*	−0.46	0.006	0.951	−0.42	0.048	0.866
rs455863	*IL17RE*	−0.46	0.005	0.957	−0.57	<0.0001	0.996
rs279581	*IL17RE*	0.39	0.116	0.878	0.50	0.001	0.974
rs279572	*IL17RE*	0.37	0.208	0.854	0.48	0.002	0.960
rs172155	*IL17RE*	−0.37	0.244	0.847	−0.48	0.003	0.953
rs2272128	*IL18RAP*	−0.47	0.004	0.965	-0.39	0.125	0.829
rs2243193	*IL19*	0.42	0.031	0.919	0.50	0.001	0.973
rs12044804	*IL19*	−0.48	0.002	0.973	−0.62	<0.0001	1.000
rs4845143	*IL19*	0.41	0.042	0.910	0.49	0.001	0.969
rs12409415	*IL19*	−0.45	0.012	0.928	−0.49	0.002	0.945
rs12046559	*IL19*	0.43	0.021	0.923	0.54	<0.0001	0.989
rs12145973	*IL19*	0.59	<0.0001	1.000	0.42	0.106	0.820
rs2138992	*IL19*	0.46	0.006	0.959	0.52	<0.0001	0.984
rs2232360	*IL20*	−0.38	0.166	0.858	−0.46	0.007	0.934
rs1322393	*IL20RA*	0.39	0.098	0.876	0.49	0.002	0.963
rs1322394	*IL20RA*	−0.38	0.206	0.844	−0.48	0.003	0.948
rs75977	*IL20RB*	-0.43	0.019	0.934	−0.47	0.005	0.940
rs835634	*IL20RB*	0.43	0.020	0.932	0.46	0.008	0.929
rs747842	*IL20RB*	0.43	0.023	0.929	0.46	0.008	0.930
rs12934152	*IL21R*	−0.57	<0.0001	0.999	−0.50	0.005	0.952
rs10903022	*IL22RA1*	−0.42	0.037	0.920	−0.49	0.002	0.971
rs10751768	*IL22RA1*	0.42	0.030	0.926	0.50	0.001	0.974
rs3795302	*IL22RA1*	0.39	0.087	0.898	0.49	0.001	0.971

Published May 25, 2009

Table I. Correlations with micro- and macropathogen richness *(Continued)*

SNP	Gene	Micropathogens			Macropathogens		
rs4486393	*IL22RA1*	0.39	0.146	0.863	0.50	0.002	0.968
rs4292900	*IL22RA1*	0.40	0.109	0.882	0.47	0.005	0.947
rs16829209	*IL22RA1*	0.39	0.156	0.856	0.50	0.002	0.959
rs11570915	*IL26*	0.40	0.132	0.852	0.53	0.001	0.980
rs3814240	*IL26*	0.42	0.026	0.929	0.49	0.002	0.964
rs3814241	*IL26*	0.50	0.001	0.982	0.47	0.004	0.946
rs10878789	*IL26*	0.47	0.004	0.962	0.41	0.052	0.868
rs9632389	*IL31RA*	−0.48	0.005	0.960	−0.43	0.058	0.837
rs10055201	*IL31RA*	0.39	0.120	0.869	0.45	0.009	0.927
rs1554999	*IL32*	−0.48	0.002	0.965	−0.55	<0.0001	0.988

[a]Bonferroni-corrected p-value.
[b]Percentile rank relative to the distribution of SNP control sets matched for allele frequency.

both. These variants map to a total of 44 IL genes (Table I). We next verified whether the correlations between IL SNP frequencies and pathogens could be secondary to associations with other environmental variables (e.g., climatic factors). Hence, for each geographic location corresponding to HGDP-CEPH populations, the following parameters were obtained: average annual mean and maximum temperature, precipitation rate, and short-wave radiation flux. None of the SNPs reported in Table I significantly correlated with any of these variables (Table S3).

Allele frequency spectra in human populations are affected by selective and nonselective events; whereas selection acts on specific genes, nonselective forces (e.g., demography or distance from Africa [22]) are expected to affect all loci equally. We thus compared the strength of IL gene correlations to sets of control SNPs in the dataset. In particular, for each IL SNP in Table I, we extracted from the full HGDP-CEPH dataset all SNPs having an overall minor allele frequency (averaged over all populations) differing by <0.01 from its frequency; for all SNPs in the frequency-matched groups, we calculated Kendall's rank correlation coefficient (τ) between micro- and macropathogen richness and allele frequencies. We next calculated the percentile rank of IL gene SNPs in the distribution of Kendall's τ obtained for the control sets. In most cases (46 for micro- and 66 for macropathogens), percentile ranks higher than the 95th were obtained for IL SNPs. These data indicate that SNPs in IL genes are clearly more strongly influenced by pathogen richness compared with control SNPs, suggesting that selective forces (i.e., pathogen-driven selection), and not nonselective forces, are responsible for the observed associations. All data are gathered in Table I, which shows the compilation of all SNPs that significantly correlate with either micro- or macropathogen richness; the value of τ for both correlations, as well as Bonferroni-corrected P values and percentile ranks, are reported. It is worth noting that data in Table I have been organized to keep together all SNPs in the same gene and to group ligands and receptors; therefore, the order does not reflect greater association with micro- or macropathogens, which can instead be inferred from correlation values.

Another issue that deserves attention relates to the genomic organization of IL genes, because many of them are located in clusters. This raises the possibility that many of the observed allele associations are spurious and derive from linkage to a single selected allele. Yet analysis of linkage disequilibrium (LD; Fig. S2) indicates that linkage is not extensive across gene clusters, with the exception of *IL17RE* and *IL17RC*. Therefore, with the exception of these two genes, the remaining loci are independent targets of pathogen-driven selection.

Balancing selection acts on *IL1F5*, *IL1F7*, *IL1F10*, *IL7R*, and *IL18RAP*

Pathogen-driven variations in allele frequencies can occur under different selection scenarios, such as directional or balancing selection. Given the aforementioned results and the role of IL genes in regulating immune responses, we next verified whether selection signatures could be identified at IL genes by using classical population genetic analyses. To this aim, we exploited the observation that 68 out of 91 IL loci have been included in genetic variation projects (i.e., the SeattleSNPs program and the Innate Immunity in Heart, Lung and Blood Disease Programs for Genomic Applications) so that resequencing data (although with some gaps) are available in at least two populations: one with European ancestry (EU) and one with African ancestry (either Yorubans [YRI] or African Americans [AA]). Common population genetics tests include Tajima's D (D_T) (23) and Fu and Li's D* and F* (24). D_T tests the departure from neutrality by comparing two nucleotide diversity indexes: θ_W (25), which is an estimate of the expected per site heterozygosity, and π (26), which is the average number of pairwise sequence nucleotide differences. Positive values of D_T indicate an excess of intermediate frequency variants and are a hallmark of balancing selection; negative D_T values indicate either purifying selection or a high representation of rare variants as a result of a selective sweep. Fu and Li's F* and D* (24) are also based on SNP frequency spectra and differ from D_T in that they also take into account whether mutations occur in external or internal branches of a genealogy. Population history, in addition to selective processes, affects frequency spectra and all related statistics; for this reason, statistical significance was evaluated by performing coalescent simulations using a population genetics model that

Published May 25, 2009
JEM

incorporates demographic scenarios (27). Simulations were performed using the cosi package (27) and were used to derive a p-value that indicates whether or not the value obtained for a given IL locus is expected under a given demographic scenario; a significant P value indicates that the obtained value is unlikely under the specified conditions and, therefore, that neutrality can be rejected.

Another method of figuring out the effects of selection and population history again lies in the assumption that selection acts on a single locus, whereas demography affects the whole genome; therefore, we calculated test statistics for 238 genes resequenced by the National Institute of Environmental Health Science (NIEHS) program (Table S4). These empirical distributions were used to calculate the percentile rank of D_T, D*, and F* values for analyzed IL genes; this procedure provides a direct comparison of the locus under analysis with a sample of human genes and allows an estimation of how unusual the value obtained is (e.g., a percentile rank of 0.99 suggests that neutral evolution is unlikely).

5 of the 68 IL genes available for population genetic analysis gave significant results in at least one population; three of them (*IL1F10*, *IL7R*, and *IL18RAP*) also display at least one SNP correlating with pathogen richness (Table I). Data concerning θ_W and π, as well as D_T, D*, and F*, are reported in Table II. For *IL1F5*, *IL1F7*, and *IL1F10*, θ_W and π were close to or higher than the 95th percentile in the distribution of NIEHS genes in both populations, indicating an excess of polymorphisms at these loci; conversely, no exceptional θ_W and π were observed for *IL18RAP* in AA and for *IL7R* in either population. Calculation of D_T, D*, and F* for the five genes indicated that one or more statistics rejected neutrality in both Africans and Europeans for *IL1F5*, *IL1F10*, and *IL18RAP*; conversely, unusually high values were obtained only for YRI and EU in the case of *IL1F7* and *IL7R*, respectively. Overall, these data suggest the action of balancing selection. It should be noted that *IL7R* is encompassed by a copy number variant (CNV; http://projects.tcag.ca/variation/); yet the CNV only occurs in 1 out of 110 chromosomes, and thus should not affect our results.

Another commonly used test to verify departure from selective neutrality is the Hudson, Kreitman, and Aguade (HKA) test (28); it is based on the assumption that under neutral evolution, the amount of within-species diversity correlates with levels of between-species divergence because both depend on the neutral mutation rate. An excess of intraspecific diversity compared with divergence ($k > 1$) is considered a signature of balancing selection. Here, we performed a maximum likelihood HKA test (MLHKA) (29) by comparing each IL gene in Table II to 16 neutrally evolving genes resequenced in the same individuals (see Materials and methods for details). The MLHKA test rejected the neutral evolution model for *IL1F5* and *IL7R* (in both populations), as well as *IL1F10* (in YRI), but not for *IL1F7* and *IL18RAP*. Indeed, in the latter case, interspecific divergence higher than the genome average paralleled the high levels of intraspecific diversity (unpublished data).

Population genetic diversity, measured as F_{ST}, can also provide information on selective processes. Under selective neutrality, F_{ST} is determined by genetic drift, which is mainly accounted for by demographic history and similarly affects all genomic loci. Conversely, natural selection being a locus-specific force, it can affect F_{ST} values for specific genes. Balancing selection may lead to a decrease in F_{ST} compared with neutrally evolving loci (8); specifically, low F_{ST} values among continental populations strongly suggests the action of balancing selection worldwide (i.e., irrespective of local environmental

Table II. Summary statistics for five IL genes

Gene	L[a]	P[b]	N[c]	S[d]	θ^e	π^f	D_T			D*			F*		
							Value	P-value[g]	Rank[h]	Value	P-value[g]	Rank[h]	Value	P-value[g]	Rank[h]
IL1F5	6.4	YRI	48	50	17.61	21.73	0.81	0.033	0.97	0.43	0.11	0.85	0.68	0.045	0.93
		EU	46	39	13.87	25.33	2.84	0.0004	>0.99	1.68	0.0018	>0.99	2.49	0.0002	>0.99
IL1F7	6.1	YRI	48	45	16.57	22.07	1.14	0.016	>0.99	0.82	0.037	0.95	1.11	0.012	0.98
		EU	46	35	13.02	8.69	−1.13	0.10	0.15	1.18	0.046	0.96	0.43	0.28	0.69
IL1F10	4	YRI	48	34	19.36	18.98	−0.066	0.21	0.76	1.15	0.012	0.98	0.85	0.027	0.95
		EU	46	16	9.20	12.68	1.19	0.070	0.89	1.59	0.0041	>0.99	1.72	0.0090	0.99
IL18RAP	17.8	AA	48	97	12.30	14.24	0.56	0.039	0.95	0.69	0.013	0.94	0.77	0.011	0.96
		EU	46	88	11.27	15.56	1.36	0.057	0.91	1.26	0.0099	0.97	1.54	0.013	0.98
IL7R	21	AA	48	119	12.80	9.94	−0.80	0.42	0.39	0.26	0.10	0.82	−0.16	0.21	0.71
		EU	46	70	7.60	10.33	1.28	0.084	0.90	1.53	0.0029	>0.99	1.71	0.011	0.99

[a]Length of analyzed resequenced region (in kilobasepairs).
[b]Population.
[c]Sample size (chromosomes).
[d]Number of segregating sites.
[e]θ_w estimation per site ($\times 10^{-4}$).
[f]π estimation per site ($\times 10^{-4}$).
[g]P-values obtained by applying a calibrated population genetics model, as described in the text.
[h]Percentile rank relative to the distribution of values obtained for 238 NIEHS genes (i.e., comparison with an empirical distribution).

Published May 25, 2009

pressures), whereas reduced population differentiation within continents is consistent with a locally exerted selective pressure resulting in balancing selection. We calculated F_{ST} for the 5 IL genes and compared it to the distribution of this parameter among 238 NIEHS genes. Unusual F_{ST} values were only observed for *IL18RAP*, in this case, a negative F_{ST} was obtained, indicating that the real value is close to 0, therefore corresponding to a percentile rank lower than the 2.5th.

One other effect of balancing selection is the maintenance of divergent haplotype clades separated by deep coalescence times (30). We therefore studied haplotype genealogies by constructing median-joining networks. In particular, regions displaying low recombination rates were selected (see Materials and methods and Fig. S3); in the case of *IL18RAP* and *IL1F10*, network analysis was not performed because of the low LD throughout the gene region.

Haplotype genealogies for *IL1F5*, *IL1F7*, and *IL7R* revealed two major clades separated by long branches (Fig. 1), each containing common haplotypes. To estimate the time to the most recent common ancestor (TMRCA) of the haplotype clades, we applied a phylogeny-based method (31) based on the measure ρ, which is the average pairwise difference between haplotypes and a root. TMRCA estimates of 2.76 million years (MY; SD = 660 KY), 2.10 MY (SD = 404 KY), and 1.68 MY (SD = 408 KY) were obtained for *IL1F5*, *IL1F7*, and *IL7R*, respectively. In all cases, we verified these results using GENETREE, which is based on a maximum-likelihood coalescent analysis (32). Consistently, the resulting gene trees, rooted using the chimpanzee sequences, are partitioned into two deep branches (Fig. 2). Using this method, a second estimate of the TMRCAs for the three genes was obtained (Fig. 2 and Table S5). Estimates of coalescence times for neutrally evolving autosomal human loci range between 0.8 and 1.5 MY (33); the TMRCAs we estimated for *IL1F5*, *IL1F7*, and *IL7R*, although not exceptional, are deeper than for most neutrally evolving loci. Again, this finding suggests the action of balancing selection (30).

Heterozygote advantage (also known as overdominance) is one of the possible causes of balancing selection. To verify whether this is the case for the five selected IL genes, for each population we calculated the ratio of observed heterozygosity to expected gene diversity. This same ratio calculation has recently been applied to human HapMap SNPs, and threshold values for inference of overdominance have been set to 1.160 and 1.165 for YRI and EU, respectively (34). To obtain an additional estimate of this parameter distribution in the human genome, ratios were also calculated for NIEHS genes. Whereas all other genes showed nonexceptional values, the observed heterozygosity to expected gene diversity ratio for *IL1F5* amounted to 1.07 and 1.20 for YRI and EU, respectively. This value is higher than the previously set threshold for EU and falls above the 99th percentile value obtained from NIEHS genes in this same population.

Overall, these data strongly suggest the action of balancing selection on these 5 IL genes reported in Tables II and III; it is worth mentioning that the MLHKA test failed to reject neutrality for *IL1F7* and *IL18RAP*, suggesting that relaxed constraint rather than balancing selection might be responsible for the observed high D_T, D*, and F* values. Still, we noticed that the 6 polymorphisms within *IL1F7* exons in YRI are all

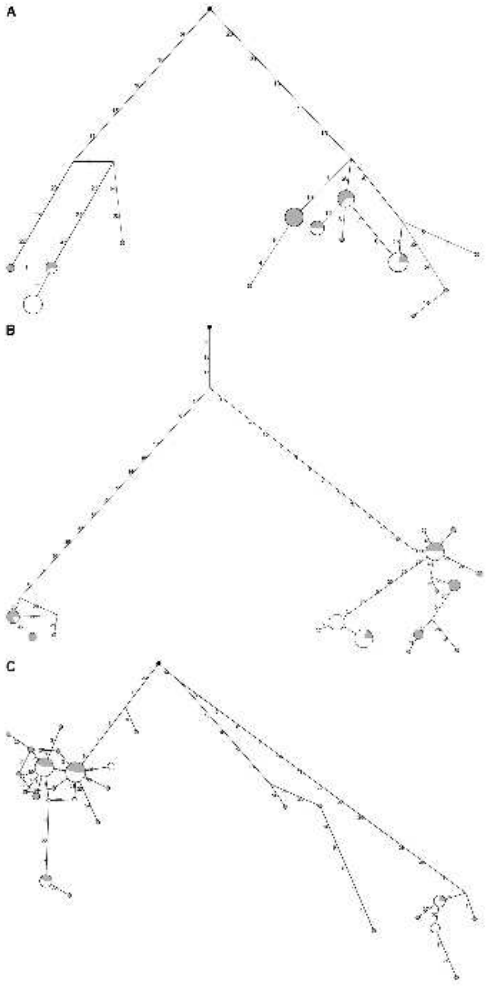


Figure 1. Haplotype genealogy for *IL1F5*, *IL1F7*, and *IL7R* gene regions. The analyzed regions correspond to the largest LD block for each gene (Fig. S2). Each node represents a different haplotype, with the size of the circle proportional to the haplotype frequency. Nucleotide differences between haplotypes are indicated on the branches of the network. Circles are color-coded according to population (gray, AA or YRI; white, EU). The chimpanzee sequence is also shown (black). Fig. S2 is available at http://www.jem.org/cgi/content/full/jem.20082779/DC1.

Published May 25, 2009
JEM

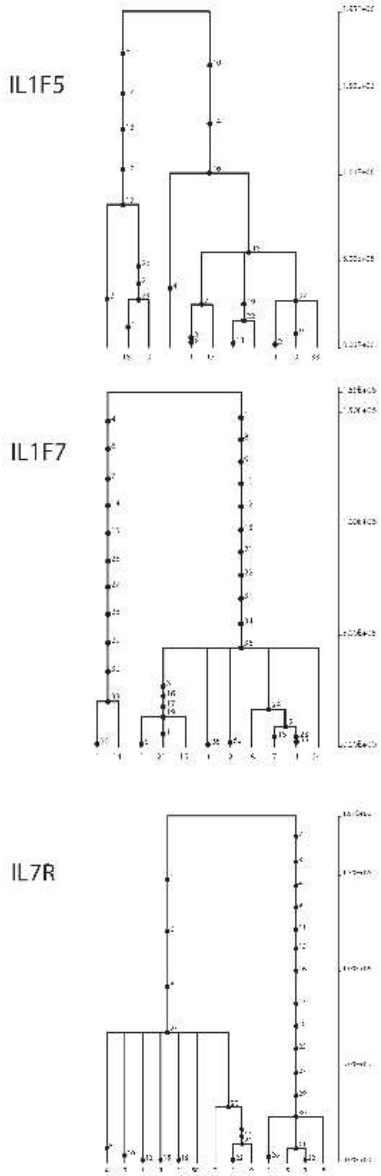


Figure 2. Estimated trees for *IL1F5*, *IL1F7*, and *IL7R* gene regions. The analyzed regions correspond to the largest LD block for each gene (Fig. S2). Mutations are represented as black dots and named for their physical position along the regions. The absolute frequency of each haplotype is also reported. Fig. S2 is available at http://www.jem.org/cgi/content/full/jem.20082779/DC1.

accounted for by nonsynonymous substitutions; application of the MK tests (35) using orangutan for divergence yielded a p-value of 0.058. Although not fully significant, this result indicates an unusual frequency of replacement polymorphic variants and, together with the deeper-than-average TMRCA, is in line with balancing selection rather than with functional relaxation.

With respect to *IL18RAP*, the extremely low F_{ST} value we observed (Table III) is not consistent with relaxed functional constraints, but instead supports the idea that a balanced polymorphism is maintained in AA and EU, suggesting response to a widespread selective pressure (i.e., not local adaptation). Finally, it should be noted that *IL1F5* and *IL1F10* are located nearby, raising the possibility that the signatures we observe at both gene regions are caused by hitchhiking with a single functional variant; yet, LD is low across the region because of the presence of a recombination hotspot within the *IL1F10* gene region. Moreover, although hitchhiking has the potential to affect large genomic regions, the signatures of balancing selection are predicted to extend over relatively short distances (36, 37). We therefore consider that the two genes might be independent selection targets.

Risk alleles for inflammatory bowel and celiac disease correlate with pathogen-richness

Previous analyses showed that a polymorphism (rs917997) located ~1,500 bp downstream of *IL18RAP* is significantly associated with both celiac disease (CeD) and inflammatory bowel disease (IBD; with the SNP also influencing the level of gene expression (38, 39). This variant was not included in our analysis of pathogen correlations because of its location outside the gene region (as defined in Materials and methods). Still, we observed that rs917997 is in strong LD (D' = 1 and 0.87 in EU and YRI, respectively) with rs2272128, which we found to correlate with macro- and micropathogen richness. We therefore checked the presence of rs917997 and verified that the risk allele for CeD and IBD also correlates significantly with pathogen richness (Table IV). Stimulated by this finding, we next verified whether the frequency distribution of other

Table III. MLHKA test and F_{ST} for five IL genes

Gene	P[a]	MLHKA		F_{ST} (rank[b])
		k[c]	P-value	
IL1F5	YRI	2.92	0.0043	0.13 (0.53)
	EU	2.85	0.0032	
IL1F7	YRI	1.82	0.29	0.28 (0.89)
	EU	1.91	0.29	
IL1F10	YRI	2.38	0.017	0.16 (0.62)
	EU	1.46	0.56	
IL18RAP	AA	1.18	0.22	0 (<0.025)
	EU	1.62	0.38	
IL7R	AA	2.53	0.0042	0.019 (0.084)
	EU	2.14	0.028	

[a]Population.
[b]Percentile rank relative to the distribution of F_{ST} values calculated for 238 NIEHS genes.
[c]Selection parameter.

Published May 25, 2009

susceptibility alleles in IL genes has been influenced by pathogen richness. To analyze only variants that have been identified in an unbiased manner (i.e., without a priori hypothesis on the genes involved), we searched among published genome-wide association studies for SNPs in IL genes that have been associated with any trait. For available variants in the CEPH-HGDP panel, we next calculated correlations with micro- and macropathogen richness. Results reported in Table IV indicate that six out of nine risk variants for CeD or IDB/Crohn's disease (CD) were associated with micro- and macropathogen richness; in particular, two of them located within *IL23R* are in tight LD, and thus do not represent independent observations. Noticeably, in all six cases, the risk allele correlates with pathogen richness and correlations with micropathogens were stronger compared with macropathogens (with the exception of rs11465804), a situation different from the general observation that macropathogens have represented a more powerful selective pressure on IL genes.

DISCUSSION

We analyzed the recent evolutionary history of IL genes in humans by integrating information on environmental variables with classical population genetics and association studies. Results herein suggest that microbes and parasitic worms played a relevant role as selective agents, but the pressure imposed by helminths on IL genes has been stronger than the one caused by viral and microbial agents. Helminths were present among our ancestors before the emergence of humans as a species (for review see reference [40]). These parasites evolve at lower rates than viruses and bacteria and, in contrast to most viral/microbial agents, are able to maintain themselves in small human communities (41). Notably, by establishing chronic infections, parasitic worms affect the susceptibility of their host to viruses, bacteria, and protozoa (for review see reference [2]). Therefore, helminths might have represented a stable threat to human populations and their distribution, which is not associated with sudden epidemics and, as in the case of micropathogens, might have left stronger genetic signatures.

A limitation of our study is that we implicitly assumed that the number of different pathogen species/genera per country has been maintained proportionally unchanged along human evolutionary history. Although clearly an oversimplification, this might reflect reality to some degree, given that climatic variables (e.g., precipitation rates and temperature) have a primary importance in driving the spatial distribution of human micro- and macropathogens (21). Therefore, while the fitness cost imposed by specific species/genera might have evolved rapidly, the relative number of pathogen species per country might have changed proportionally less.

Another possible caveat of our results concerns the definition of "pathogen," in that we included any organism that can cause a disease irrespective of its virulence or pathogenicity. The reason for this choice is that the fitness of a pathogen is a direct measure of the ability of such pathogen to replicate within a given environment. Fitness is dependent on both the features of the pathogen and of the host. The features of the

Table IV. Correlations with micro- and macropathogen richness for SNPs associated with different traits

SNP	Gene/location	Allele		Micropathogen		Macropathogen		Trait/disease
		Risk	Anc.[a]	τ[b]	P-value[c]	τ[b]	P-value[c]	
rs6897932	*IL7R*	C	C	0.22	n.s.[d]	0.21	n.s.[d]	Multiple sclerosis/Type 1 diabetes
rs917997	*IL18RAP* (downstream)	A	G	0.42	<0.001	0.35	0.008	CeD/IBD
rs10045431	*IL12B*	C	C	0.43	<0.001	0.34	0.015	CD
rs7517847	*IL23R*	C	A	0.23	n.s.[d]	0.26	n.s.[d]	IBD
rs11209026	*IL23R*	G	G	0.47	<0.001	0.44	0.001	IBD
rs11465804	*IL23R*	T	T	0.39	0.004	0.44	<0.001	CD
rs6822844	*IL2/IL21* (intergenic)	G	G	0.40	0.004	0.39	0.006	CeD
rs3024505	*IL10*	T	C	−0.28	n.s.[d]	−0.28	n.s.[d]	Ulcerative colitis
rs17810546	*IL12A*	G	A	0.03	n.s.[d]	−0.11	n.s.[d]	CeD
rs13015714	*IL18R1*	C	A	0.47	<0.001	0.39	0.002	CeD
rs2250417	*IL18*	A	A	0.23	n.s.[d]	0.26	n.s.[d]	Protein quantitative trait loci
rs7626795	*IL1RAP*	G	A	0.20	n.s.[d]	0.04	n.s.[d]	Lung cancer
rs4129267	*IL6R*	C	C	−0.17	n.s.[d]	−0.24	n.s.[d]	Protein quantitative trait loci/pulmonary function
rs6761276	*IL1F10*[e]	n.r.[f]	T	−0.31	0.030	−0.51	<0.001	Protein quantitative trait loci
rs12251307	*IL2RA*	T	T	−0.16	n.s.[d]	−0.13	n.s.[d]	Type 1 diabetes

[a]Ancestral state based on chimpanzee sequence.
[b]The correlation coefficient is calculated between pathogen richness and the risk allele frequency;.
[c]Bonferroni-corrected p-value (15 tests).
[d]Nonsignificant (P > 0.05).
[e]This SNP is located within *IL1F10*, but affects *IL1RN* protein levels.
[f]Not reported.

Published May 25, 2009
JEM

pathogen include its abilities to (a) replicate within the host, (b) select a proper cell target (tropism), (c) avoid the immune response, and (d) escape the obstacles posed by drugs. The features of the host include the immune response, the genetic background, and the availability of target cells. The reciprocal balance between these factors determines the virulence of a pathogen, a feature that, therefore, cannot be considered as a constant, but rather evolves dynamically over time.

Despite these limitations, we were able to identify several variants that are likely candidates for pathogen-driven selection, and the suitability of this approach is confirmed by the previous demonstration that variability at *HLA* genes correlates with a similar measure of pathogen richness (10). Our data point to the SNPs in Table I as good candidates for experimental analyses aimed at inferring their role in IL gene function, given that a signature of natural selection necessarily implies the presence of a functional variant (either the correlated variant itself or a linked one). Also, genes subjected to a selective pressure from infectious diseases should be regarded (42) as obvious candidates for genetic epidemiology studies (e.g., case-control studies).

It is worth noting that most members of the IL-1 signaling pathway were observed to correlate with pathogen richness. IL-1A and IL-1B are pleiotropic cytokines with central roles in immune and inflammatory responses (43) required for the development of Th2-mediated immunity and protection against chronic infection in mice (44). The observation that SNPs strongly correlated with micro- and macro-pathogen richness map to *IL1RAPL1* is more puzzling, as this gene is not known to be associated with infection and immune response, but is involved in brain development and function (45); nonetheless, this gene is expressed in tissues that are different from brain (46), suggesting that it plays other roles apart from neurodevelopment.

Strong correlations with macropathogen richness were obtained for *IL4*, *IL4R*, and *IL10*. These molecules are pivotal to the elicitation of Th2 response, which is the immune response central to helminth resistance (40, 47). The strongest correlation with macropathogens was observed for rs12044804 in *IL19* ($\tau = -0.62$). This gene is located within an IL cluster, which also comprises *IL10* and *IL20*, and is encompassed by a low-frequency CNV. The low levels of LD across the IL cluster (Fig. S2) suggest that *IL19*, *IL20*, and *IL10* SNPs are independently associated with pathogen richness. *IL19*, *IL20*, *IL22*, *IL24*, and *IL26* all belong to the IL-20 subfamily, are all produced by different leukocyte populations, and all bind to partially shared receptors that are mainly expressed by epithelial cells and known to promote keratinocyte growth and to induce skin inflammatory responses (48, 49). The correlation of SNPs in most of these genes with micropathogen and helminth richness suggests that modulation of skin immunological properties might represent an adaptive response to parasite species that infect humans through the skin.

SNPs in *IL2RB*, *IL15*, and *IL15RA* also correlated with macropathogen richness; these genes converge on the same pathway as IL-2RB and IL-15RA are part of a trimeric complex that binds IL-15 (50). IL-15 has a central role in intestinal inflammatory processes and in the pathogenesis of CD and CeD (51), possibly participating in the immune protection of gut tissues. An interesting observation is that IL-2RB, via IL-15 binding, regulates the intestinal epithelial barrier by transducing signals that result in tight junction formation (52). Variation of mucosal permeability after nematode infection is a host defense response and is important for efficient parasite expulsion (53); the correlation we observed between SNPs in these genes and macropathogen richness might indicate selection for improved intestinal clearance of nematodes. Again, mouse models will prove central in addressing the role of these loci in parasite expulsion; e.g., in the case of *IL4R*, which also correlates with helminth richness (Table I), treatment with antibodies or use of *il4r*-deficient mice prevented expulsion of *Heligmosomoides polygyrus*, *Trichuris muris*, and *Nippostrongylus brasiliensis* (54).

A major locus for susceptibility to *Schistosoma mansoni* infection (*SM1*) has previously been mapped to 5q31–q33 (55). This region covers *IL4*, *IL5*, *IL3*, *IL13*, *IL9*, and *IL12B*, as well as other candidate loci (*IRF1* and *CSF2*). Although our data do not allow inference on which gene is responsible for increased susceptibility to schistosomiasis, it is worth noting that four SNPs in *IL4* display very strong correlation with helminth richness. *IL4* might therefore be regarded as a candidate locus for susceptibility to *S. mansoni* infection, with dedicated association studies being required to verify this prediction. Conversely, the *SM2* region (6q22–q23), where a locus controlling hepatic fibrosis in *S. mansoni* infection has been mapped (56), harbors no IL genes.

Pathogen-driven variations in allele frequencies can occur under different selection scenarios, such as directional or balancing selection; the latter itself is the result of an initial stage of positive selection that favors the spread in a population of a new allele until selection opposes its fixation and a balanced situation is established. Common causes of balancing selection include heterozygote advantage, changing environmental conditions, and frequency-dependent selection, all of these possibly applying to host–pathogen interactions. Also, given the pleiotropic roles of many IL genes, selective pressures different from pathogen richness might affect the evolutionary fate of these loci. Classical population genetics analyses indicated that five IL genes are likely targets of balancing selection. *IL1F5*, *IL1F7*, and *IL1F10* are recently discovered IL-1 family members (57) that are located within the *IL1* gene cluster; based on comparative analyses, IL-1F5 and IL-1F10 are predicted to act as antagonists (58). *IL1F5* is a regulator of skin and brain inflammation (59, 60) and it is expressed in many different human tissues (57). Interestingly, an excess of heterozygotes was observed for *IL1F5*, suggesting that overdominance might underly the maintenance of a balanced polymorphism in the gene. Overdominance is rare in humans (36, 61) and is hypothesized to enhance immune response flexibility by modulating allele-specific gene expression in different cell types and in response to diverse stimuli/cytokines (62). Whether this is the case for *IL1F5* remains to be verified.

Published May 25, 2009

IL1F10 is a still relatively unknown protein mainly expressed in skin, proliferating B cells, and tonsils (63). One of the intermediate frequency SNPs in the gene is accounted for by a missense substitution that replaces an aspartic acid residue with an alanine (Asp51Ala); the presence of a negatively charged residue at this position is conserved among mammals (Fig. S4), possibly suggesting functional significance and awaiting experimental testing.

In analogy to *IL1F5*, the other IL1 family member we identified as rejecting neutral evolution, i.e., *IL1F7*, also acts as an antiinflammatory molecule (64).

IL1F5, *IL1F7*, and *IL1F10* are not known to be involved in human diseases; in contrast, *IL18RAP* and *IL7R* play a role in triggering immune responses. The IL-7/IL-7R ligand-receptor pair is central to the proliferation and survival of B and T leukocytes. We identified one SNP in the gene as highly correlated with macropathogen richness (Table I). This is not surprising given the role of Th2 responses in helminth infection and the involvement of IL-7R in the TSLP signaling pathway (65), which in turn regulates Th2-mediated inflammatory responses (66).

Similar to *IL7R*, *IL18RAP* plays a known role in human pathology. The gene encodes a component of the protein complex involved in transducing IL-18 signal, resulting in the activation of NF-κB (67). The IL-18 receptor complex is expressed in the intestine (38), and one SNP immediately downstream *IL18RAP* (rs917997) has been associated with both CeD and IBD (38, 39). We found the predisposing allele of rs917997 and a linked variant (rs2272128) to correlate with pathogen richness. The location of rs2272128 in the 5′ gene region and its strong correlation with pathogens might suggest that it (rather than rs917997) represents or is in close LD with the functional allele. Moreover, the correlation of a risk allele for autoimmune diseases with pathogen-richness suggests an interesting link between adaptation and disease. Indeed, we observed that five more risk alleles for either IBD or CeD significantly correlate with micro- and macropathogen richness. Albeit preliminary, these data suggest that infectious agents have shaped the genetic variation at IL loci involved in intestinal inflammatory processes and, as a consequence, the genetic predisposition to both CeD and CD/IBD.

A north–south gradient for IBD prevalence has been described in both the US and Europe (68). This observation, together with the increase of IBD prevalence in the last 40 yr (68) and the hypothesis that helminths elicit Th2-mediated responses, led to the proposal that lower exposure to parasitic worms in the setting of industrialized countries results in unbalanced immune response, and eventually predisposes to IBD (68, 69). The so-called hygiene hypothesis, which clearly implies evolutionary considerations concerning human–pathogen interactions, has been supported by recent studies in both humans and mice (40, 68–70). Data herein seem to indicate that a portion of CeD- and IBD-predisposing alleles have been selected by micropathogen richness, pointing to an adaptive role for these variants. Although not directly supportive of the "IBD hygiene hypothesis," these results indicate a higher disease predisposition in subjects carrying IL SNP variants that confer stronger protection against viruses/bacteria and therefore likely elicit more vigorous Th1 responses. Living conditions in industrialized countries have resulted in a reduction of both helminth and bacterial/viral infection. The effect of this environmental change on the homeostasis of immune responses might be difficult to reconcile with simple theories (71). In this complex scenario, we consider that evolutionary studies and population genetics approaches, such as the one proposed here, provide some insight into the genetic basis of predisposition to infectious and autoimmune diseases.

MATERIALS AND METHODS

Data retrieval and haplotype construction. Data concerning the HGDP-CEPH panel derive from a previous work (20). A SNP was ascribed to a specific gene if it was located within the transcribed region or no further than 500 bp upstream the transcription start site.

Genotype data for resequenced IL genes were retrieved from the SeattleSNPs (http://pga.mbt.washington.edu) and Innate Immunity PGA (http://innateimmunity.net/) web sites. A total of 68 genes were available for analysis. For each gene, genotypes deriving from 24 subjects of African ancestry and 23 of Caucasian ancestry were retrieved.

Genotype data for 238 resequenced human genes were derived from the NIEHS SNPs Program web site (http://egp.gs.washington.edu). We specifically selected genes that had been resequenced in populations of defined ethnicity (NIEHS panel 2).

Haplotypes were inferred using PHASE version 2.1 (72), a program for reconstructing haplotypes from unrelated genotype data through a Bayesian statistical method.

Recombination rates were derived from the University of California at Santa Cruz genome browser web site (http://genome.ucsc.edu). Information concerning CNVs was derived from the database of genomic variants (http://projects.tcag.ca/variation/).

Variants and risk alleles identified in genome-wide association studies were retrieved from the National Human Genome Research Institute web site (http://www.genome.gov/) updated on December 1, 2008.

Statistical analysis. D_T (23), Fu and Li's D* and F* (24) statistics, and diversity parameters θ_W (25) and π (26) were calculated using libsequence (73). Coalescent simulations were performed using the cosi package (27) and its best-fit parameters for YRI, AA, and EU populations with 10^4 iterations. cosi is a simulation package based on a population genetics model calibrated on empirical data; it therefore allows incorporation of demographic scenarios in simulations.

The F_{ST} statistic (74) estimates genetic differentiation among populations and was calculated as previously proposed (75).

The maximum likelihood ratio HKA test was performed using the ML-HKA software (29) as previously described (18). In brief, we used multilocus data of 16 selected genes and *Pan troglodytes* (NCBI panTro2) as an outgroup (except for *IL1F7*, where *Pongo pygmaeus abelii*, NCBI ponAbe2, was used as the outgroup). The 16 reference genes were randomly selected among NIEHS loci <20 kb that have been resequenced across panel 2; the only criterion was that no reference gene rejected the neutral model (i.e., that no gene yielded significant D_T). The reference loci used were as follows: *VNN3*, *PLA2G2D*, *MB*, *MAD2L2*, *HRAS*, *CYP17A1*, *ATOX1*, *BNIP3*, *CDC20*, *NGB*, *TUBA1*, *MT3*, *NUDT1*, *PRDX5*, *RETN*, and *JUND*.

LD analyses were performed using Haploview (76), and haplotype blocks were identified through an implemented method.

Median-joining networks to infer haplotype genealogy were constructed using NETWORK 4.5 (31). Estimate of the TMRCA was obtained using a phylogeny-based approach implemented in NETWORK using a mutation rate based on the number of fixed differences between human and chimpanzee or orangutan and assuming a separation time from humans of 6 MY and 13 MY ago, respectively. A second TMRCA estimate was derived from application of a maximum-likelihood coalescent method implemented

Published May 25, 2009

JEM

in GENETREE (32). Again, the mutation rate μ was obtained on the basis of the divergence between human and a primate, assuming a generation time of 25 yr. Using this μ and the maximum likelihood θ (θ_{ML}), we estimated the effective population size parameter (N_e). With these assumptions, the coalescence time, scaled in $2N_e$ units, was converted into years. For the coalescence process, 10^6 simulations were performed. All calculations were performed in the R environment (www.r-project.org).

Environmental variables. Pathogen absence/presence matrices for the 21 countries where HGDP-CEPH populations are located were derived from the Gideon database (http://www.gideononline.com) following previous methods (10, 18). Information in Gideon is updated weekly and derives from WHO reports, National Health Ministries, PubMed searches, and epidemiology meetings. The Gideon Epidemiology module follows the status of known infectious diseases globally, as well as in individual countries, with specific notes indicating the disease's history, incidence, and distribution per country. We manually curated pathogen absence/presence matrices by extracting information from single Gideon entries. These may refer to either species or genera (in case data are not available for different species of a same genus). Following previous suggestions (10, 18), we recorded only species/genera that are transmitted in the 21 countries, meaning that cases of transmission caused by tourism and immigration were not taken into account; also, species that have recently been eradicated as a result, for example, of vaccination campaigns, were recorded as present in the matrix. A total of 283 pathogen species were retrieved (Table S6). Other environmental variables such as average annual mean and maximum temperature, precipitation rate, and short-wave radiation flux were derived for the geographic coordinates corresponding to HGDP-CEPH populations from the NCEP/NCAR database (http://www.cdc.noaa.gov/PublicData/).

Online supplemental material. Table S1 shows correlations between the richness of viruses, bacteria, protozoa, and fungi. Table S2 is a list of IL genes analyzed in the study. Table S3 shows correlations with climatic variables. Table S4 provides diversity indexes and summary statistics for 238 human genes resequenced by the NIEHS program. Table S5 shows GENETREE estimates for *IL1F5*, *IL1F7*, and *IL7R*. Table S6 is a list of pathogen species/genera identified in at least one population. Fig. S1 shows the geographic location and pathogen richness estimates for the 52 HGDP-CEPH populations. Fig. S2 shows LD structure for IL gene clusters. Fig. S3 reports LD blocks for *IL1F5*, *IL1F7*, and *IL7R*. Fig. S4 shows multiple protein alignment for IL-1F10. Online supplemental material is available at http://www.jem.org/cgi/content/full/jem.20082779/DC1.

We are grateful to Dr. Roberto Giorda for helpful discussion about the manuscript. We also wish to thank Dr. Daniele Sampietro for technical assistance in retrieving data on climatic variables.

The authors have no conflicting financial interests.

Submitted: 10 December 2008
Accepted: 28 April 2009

REFERENCES

1. Kapp, C. 1999. WHO warns of microbial threat. *Lancet*. 353:2222.
2. Hotez, P.J., P.J. Brindley, J.M. Bethony, C.H. King, E.J. Pearce, and J. Jacobson. 2008. Helminth infections: the great neglected tropical diseases. *J. Clin. Invest.* 118:1311–1321.
3. Colley, D.G., P.T. LoVerde, and L. Savioli. 2001. Infectious disease. Medical helminthology in the 21st century. *Science*. 293:1437–1438.
4. Callender, J.E., S. Grantham-McGregor, S. Walker, and E.S. Cooper. 1992. Trichuris infection and mental development in children. *Lancet*. 339:181.
5. Nokes, C., S.M. Grantham-McGregor, A.W. Sawyer, E.S. Cooper, and D.A. Bundy. 1992. Parasitic helminth infection and cognitive function in school children. *Proc. Biol. Sci.* 247:77–81.
6. Strachan, D.P. 1989. Hay fever, hygiene, and household size. *BMJ*. 299:1259–1260.
7. Zaccone, P., O.T. Burton, and A. Cooke. 2008. Interplay of parasite-driven immune responses and autoimmunity. *Trends Parasitol.* 24:35–42.
8. Charlesworth, D. 2006. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet.* 2:e64.
9. Hurst, L.D. 2009. Fundamental concepts in genetics: genetics and the understanding of selection. *Nat. Rev. Genet.* 10:83–93.
10. Prugnolle, F., A. Manica, M. Charpentier, J.F. Guegan, V. Guernier, and F. Balloux. 2005. Pathogen-driven selection and worldwide HLA class I diversity. *Curr. Biol.* 15:1022–1027.
11. Quintana-Murci, L., A. Alcais, L. Abel, and J.L. Casanova. 2007. Immunology in natura: clinical, epidemiological and evolutionary genetics of infectious diseases. *Nat. Immunol.* 8:1165–1171.
12. Picard, C., J.L. Casanova, and L. Abel. 2006. Mendelian traits that confer predisposition or resistance to specific infections in humans. *Curr. Opin. Immunol.* 18:383–390.
13. Frodsham, A.J., and A.V. Hill. 2004. Genetics of infectious diseases. *Hum. Mol. Genet.* 13 Spec No 2:R187–R194.
14. Lettre, G., and J.D. Rioux. 2008. Autoimmune diseases: insights from genome-wide association studies. *Hum. Mol. Genet.* 17:R116–R121.
15. Wu, X., A. Di Rienzo, and C. Ober. 2001. A population genetics study of single nucleotide polymorphisms in the interleukin 4 receptor alpha (IL4RA) gene. *Genes Immun.* 2:128–134.
16. Sakagami, T., D.J. Witherspoon, T. Nakajima, N. Jinnai, S. Wooding, L.B. Jorde, T. Hasegawa, E. Suzuki, F. Gejyo, and I. Inoue. 2004. Local adaptation and population differentiation at the interleukin 13 and interleukin 4 loci. *Genes Immun.* 5:389–397.
17. Akey, J.M., M.A. Eberle, M.J. Rieder, C.S. Carlson, M.D. Shriver, D.A. Nickerson, and L. Kruglyak. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.* 2:e286.
18. Fumagalli, M., R. Cagliani, U. Pozzoli, S. Riva, G.P. Comi, G. Menozzi, N. Bresolin, and M. Sironi. 2009. Widespread balancing selection and pathogen-driven selection at blood group antigen genes. *Genome Res.* 19:199–212.
19. Pullan, R., and S. Brooker. 2008. The health impact of polyparasitism in humans: are we under-estimating the burden of parasitic diseases? *Parasitology*. 135:783–794.
20. Li, J.Z., D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza, and R.M. Myers. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*. 319:1100–1104.
21. Guernier, V., M.E. Hochberg, and J.F. Guegan. 2004. Ecology drives the worldwide distribution of human diseases. *PLoS Biol.* 2:e141.
22. Handley, L.J., A. Manica, J. Goudet, and F. Balloux. 2007. Going the distance: human population genetics in a clinal world. *Trends Genet.* 23:432–439.
23. Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. 123:585–595.
24. Fu, Y.X., and W.H. Li. 1993. Statistical tests of neutrality of mutations. *Genetics*. 133:693–709.
25. Watterson, G.A. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7:256–276.
26. Nei, M., and W.H. Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA*. 76:5269–5273.
27. Schaffner, S.F., C. Foo, S. Gabriel, D. Reich, M.J. Daly, and D. Altshuler. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 15:1576–1583.
28. Hudson, R.R., M. Kreitman, and M. Aguade. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics*. 116:153–159.
29. Wright, S.I., and B. Charlesworth. 2004. The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. *Genetics*. 168:1071–1076.
30. Takahata, N. 1990. A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. *Proc. Natl. Acad. Sci. USA*. 87:2419–2423.
31. Bandelt, H.J., P. Forster, and A. Rohl. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* 16:37–48.
32. Griffiths, R.C., and S. Tavare. 1995. Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Math. Biosci.* 127:77–98.

Published May 25, 2009

33. Tishkoff, S.A., and B.C. Verrelli. 2003. Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu. Rev. Genomics Hum. Genet.* 4:293–340.
34. Alonso, S., S. Lopez, N. Izagirre, and C. de la Rua. 2008. Overdominance in the human genome and olfactory receptor activity. *Mol. Biol. Evol.* 25:997–1001.
35. McDonald, J.H., and M. Kreitman. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature.* 351:652–654.
36. Bubb, K.L., D. Bovee, D. Buckley, E. Haugen, M. Kibukawa, M. Paddock, A. Palmieri, S. Subramanian, Y. Zhou, R. Kaul, et al. 2006. Scan of human genome reveals no new Loci under ancient balancing selection. *Genetics.* 173:2165–2177.
37. Wiuf, C., K. Zhao, H. Innan, and M. Nordborg. 2004. The probability and chromosomal extent of trans-specific polymorphism. *Genetics.* 168:2363–2372.
38. Hunt, K.A., A. Zhernakova, G. Turner, G.A. Heap, L. Franke, M. Bruinenberg, J. Romanos, L.C. Dinesen, A.W. Ryan, D. Panesar, et al. 2008. Newly identified genetic risk variants for celiac disease related to the immune response. *Nat. Genet.* 40:395–402.
39. Zhernakova, A., E.M. Festen, L. Franke, G. Trynka, C.C. van Diemen, A.J. Monsuur, M. Bevova, R.M. Nijmeijer, R. van 't Slot, R. Heijmans, et al. 2008. Genetic analysis of innate immunity in Crohn's disease and ulcerative colitis identifies two susceptibility loci harboring CARD9 and IL18RAP. *Am. J. Hum. Genet.* 82:1202–1210.
40. Dunne, D.W., and A. Cooke. 2005. A worm's eye view of the immune system: consequences for evolution of human autoimmune disease. *Nat. Rev. Immunol.* 5:420–426.
41. Dobson, A. 1992. People and disease. *In* The Cambridge Encyclopedia of Human Evolution. S. Jones, R. Martin, and D. Pilbeam, editors. Cambridge University Press, Cambridge. 411–420.
42. Burgner, D., S.E. Jamieson, and J.M. Blackwell. 2006. Genetic susceptibility to infectious diseases: big is beautiful, but will bigger be even better? *Lancet Infect. Dis.* 6:653–663.
43. Dinarello, C.A. 1994. The biological properties of interleukin-1. *Eur. Cytokine Netw.* 5:517–531.
44. Helmby, H., and R.K. Grencis. 2004. Interleukin 1 plays a major role in the development of Th2-mediated immunity. *Eur. J. Immunol.* 34:3674–3681.
45. Carrie, A., L. Jun, T. Bienvenu, M.C. Vinet, N. McDonell, P. Couvert, R. Zemni, A. Cardona, G. Van Buggenhout, S. Frints, et al. 1999. A new member of the IL-1 receptor family highly expressed in hippocampus and involved in X-linked mental retardation. *Nat. Genet.* 23:25–31.
46. Born, T.L., D.E. Smith, K.E. Garka, B.R. Renshaw, J.S. Bertles, and J.E. Sims. 2000. Identification and characterization of two members of a novel class of the interleukin-1 receptor (IL-1R) family. Delineation of a new class of IL-1R-related proteins based on signaling. *J. Biol. Chem.* 275:29946–29954.
47. Maizels, R.M., and M. Yazdanbakhsh. 2003. Immune regulation by helminth parasites: cellular and molecular mechanisms. *Nat. Rev. Immunol.* 3:733–744.
48. Parrish-Novak, J., W. Xu, T. Brender, L. Yao, C. Jones, J. West, C. Brandt, L. Jelinek, K. Madden, P.A. McKernan, et al. 2002. Interleukins 19, 20, and 24 signal through two distinct receptor complexes. Differences in receptor-ligand interactions mediate unique biological functions. *J. Biol. Chem.* 277:47517–47523.
49. Kotenko, S.V., L.S. Izotova, O.V. Mirochnitchenko, E. Esterova, H. Dickensheets, R.P. Donnelly, and S. Pestka. 2001. Identification of the functional interleukin-22 (IL-22) receptor complex: the IL-10R2 chain (IL-10Rbeta) is a common chain of both the IL-10 and IL-22 (IL-10-related T cell-derived inducible factor, IL-TIF) receptor complexes. *J. Biol. Chem.* 276:2725–2732.
50. Burton, J.D., R.N. Bamford, C. Peters, A.J. Grant, G. Kurys, C.K. Goldman, J. Brennan, E. Roessler, and T.A. Waldmann. 1994. A lymphokine, provisionally designated interleukin T and produced by a human adult T-cell leukemia line, stimulates T-cell proliferation and the induction of lymphokine-activated killer cells. *Proc. Natl. Acad. Sci. USA.* 91:4935–4939.
51. van Heel, D.A., L. Franke, K.A. Hunt, R. Gwilliam, A. Zhernakova, M. Inouye, M.C. Wapenaar, M.C. Barnardo, G. Bethel, G.K. Holmes, et al. 2007. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat. Genet.* 39:827–829.
52. Nishiyama, R., T. Sakaguchi, T. Kinugasa, X. Gu, R.P. MacDermott, D.K. Podolsky, and H.C. Reinecker. 2001. Interleukin-2 receptor beta subunit-dependent and -independent regulation of intestinal epithelial tight junctions. *J. Biol. Chem.* 276:35571–35580.
53. Murray, M., W.F. Jarrett, and F.W. Jennings. 1971. Mast cells and macromolecular leak in intestinal immunological reactions. The influence of sex of rats infected with Nippostrongylus brasiliensis. *Immunology.* 21:17–31.
54. Urban, J.F. Jr., N. Noben-Trauth, D.D. Donaldson, K.B. Madden, S.C. Morris, M. Collins, and F.D. Finkelman. 1998. IL-13, IL-4Ralpha, and Stat6 are required for the expulsion of the gastrointestinal nematode parasite *Nippostrongylus brasiliensis*. *Immunity.* 8:255–264.
55. Marquet, S., L. Abel, D. Hillaire, H. Dessein, J. Kalil, J. Feingold, J. Weissenbach, and A.J. Dessein. 1996. Genetic localization of a locus controlling the intensity of infection by *Schistosoma mansoni* on chromosome 5q31-q33. *Nat. Genet.* 14:181–184.
56. Dessein, A.J., D. Hillaire, N.E. Elwali, S. Marquet, Q. Mohamed-Ali, A. Mirghani, S. Henri, A.A. Abdelhameed, O.K. Saeed, M.M. Magzoub, and L. Abel. 1999. Severe hepatic fibrosis in Schistosoma mansoni infection is controlled by a major locus that is closely linked to the interferon-gamma receptor gene. *Am. J. Hum. Genet.* 65:709–721.
57. Smith, D.E., B.R. Renshaw, R.R. Ketchem, M. Kubin, K.E. Garka, and J.E. Sims. 2000. Four new members expand the interleukin-1 superfamily. *J. Biol. Chem.* 275:1169–1175.
58. Nicklin, M.J., J.L. Barton, M. Nguyen, M.G. FitzGerald, G.W. Duff, and K. Kornman. 2002. A sequence-based map of the nine genes of the human interleukin-1 cluster. *Genomics.* 79:718–725.
59. Blumberg, H., H. Dinh, E.S. Trueblood, J. Pretorius, D. Kugler, N. Weng, S.T. Kanaly, J.E. Towne, C.R. Willis, M.K. Kuechle, et al. 2007. Opposing activities of two novel members of the IL-1 ligand family regulate skin inflammation. *J. Exp. Med.* 204:2603–2614.
60. Costelloe, C., M. Watson, A. Murphy, K. McQuillan, C. Loscher, M.E. Armstrong, C. Garlanda, A. Mantovani, L.A. O'Neill, K.H. Mills, and M.A. Lynch. 2008. IL-1F5 mediates anti-inflammatory activity in the brain through induction of IL-4 following interaction with SIGIRR/TIR8. *J. Neurochem.* 105:1960–1969.
61. Asthana, S., S. Schmidt, and S. Sunyaev. 2005. A limited role for balancing selection. *Trends Genet.* 21:30–32.
62. Beaty, J.S., K.A. West, and G.T. Nepom. 1995. Functional effects of a natural polymorphism in the transcriptional regulatory sequence of HLA-DQB1. *Mol. Cell. Biol.* 15:4771–4782.
63. Lin, H., A.S. Ho, D. Haley-Vicente, J. Zhang, J. Bernal-Fussell, A.M. Pace, D. Hansen, K. Schweighofer, N.K. Mize, and J.E. Ford. 2001. Cloning and characterization of IL-1HY2, a novel interleukin-1 family member. *J. Biol. Chem.* 276:20597–20602.
64. Sharma, S., N. Kulk, M.F. Nold, R. Graf, S.H. Kim, D. Reinhardt, C.A. Dinarello, and P. Bufler. 2008. The IL-1 family member 7b translocates to the nucleus and down-regulates proinflammatory cytokines. *J. Immunol.* 180:5477–5482.
65. Al-Shami, A., R. Spolski, J. Kelly, T. Fry, P.L. Schwartzberg, A. Pandey, C.L. Mackall, and W.J. Leonard. 2004. A role for thymic stromal lymphopoietin in CD4(+) T cell development. *J. Exp. Med.* 200:159–168.
66. Huston, D.P., and Y.J. Liu. 2006. Thymic stromal lymphopoietin: a potential therapeutic target for allergy and asthma. *Curr. Allergy Asthma Rep.* 6:372–376.
67. Wu, C., P. Sakorafas, R. Miller, D. McCarthy, S. Scesney, R. Dixon, and T. Ghayur. 2003. IL-18 receptor beta-induced changes in the presentation of IL-18 binding sites affect ligand binding and signal transduction. *J. Immunol.* 170:5571–5577.
68. Elliott, D.E. Jr., J.F. Urban, C.K. Argo, and J.V. Weinstock. 2000. Does the failure to acquire helminthic parasites predispose to Crohn's disease? *FASEB J.* 14:1848–1855.
69. Oliva-Hemker, M., and C. Fiocchi. 2002. Etiopathogenesis of inflammatory bowel disease: the importance of the pediatric perspective. *Inflamm. Bowel Dis.* 8:112–128.
70. Weinstock, J.V., and D.E. Elliott. 2008. Helminths and the IBD hygiene hypothesis. *Inflamm. Bowel Dis.* 15:128–133.

Published May 25, 2009

JEM

71. Radford-Smith, G.L. 2005. Will worms really cure Crohn's disease? *Gut.* 54:6–8.
72. Stephens, M., and P. Scheet. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.* 76:449–462.
73. Thornton, K. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics.* 19:2325–2327.
74. Wright, S. 1950. Genetical structure of populations. *Nature.* 166:247–249.
75. Hudson, R.R., M. Slatkin, and W.P. Madison. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics.* 132:583–589.
76. Barrett, J.C., B. Fry, J. Maller, and M.J. Daly. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics.* 21:263–265.

2.3 The role of protozoa-driven selection in shaping human genetic variability

Update

Cell PRESS

Genome Analysis

The role of protozoa-driven selection in shaping human genetic variability

Uberto Pozzoli[1], Matteo Fumagalli[1,2], Rachele Cagliani[1], Giacomo P. Comi[3], Nereo Bresolin[1,3], Mario Clerici[4,5] and Manuela Sironi[1]

[1] Scientific Institute IRCCS E. Medea, Bioinformatic Laboratory, Via don L. Monza 20, 23842 Bosisio Parini (LC), Italy
[2] Bioengineering Department, Politecnico di Milano, P.zza L. da Vinci, 32, 20133 Milan, Italy
[3] Dino Ferrari Centre, Department of Neurological Sciences, University of Milan, IRCCS Ospedale Maggiore Policlinico, Mangiagalli and Regina Elena Foundation, Via F. Sforza 35, 20100 Milan, Italy
[4] Department of Biomedical Sciences and Technologies LITA Segrate, University of Milan, Via F.lli Cervi 93, 20090 Milan, Italy
[5] Don C. Gnocchi ONLUS Foundation IRCCS, Via Capecelatro 66, 20148 Milan, Italy

Protozoa exert a strong selective pressure in humans. The selection signatures left by these pathogens can be exploited to identify genetic modulators of infection susceptibility. We show that protozoa diversity in different geographic locations is a good measure of protozoa-driven selective pressure; protozoa diversity captured selection signatures at known malaria resistance loci and identified several selected single nucleotide polymorphisms in immune and hemolytic anemia genes. A genome-wide search enabled us to identify 5180 variants mapping to 1145 genes that are subjected to protozoa-driven selective pressure. We provide a genome-wide estimate of protozoa-driven selective pressure and identify candidate susceptibility genes for protozoa-borne diseases.

Protozoa as a selective force

In the late 1940s Haldane hypothesized that the prevalence of thalassemia in the Mediterranean region was the result of a selective pressure imposed by malaria [1]; we now know his proposition to be true. In addition to thalassemia, several other hemoglobinopathies and disorders of red blood cells are thought to be maintained as balanced polymorphisms because of their protective effects against *Plasmodium* [2]. Indeed, malaria is now considered to have exerted the strongest selective pressure in the recent history of humans [2], which is not surprising because 300–500 million people develop malaria and 1.5–3 million die from this disease each year [3]. Notably, other protozoan genera such as *Leishmania* and *Trypanosoma* are widespread in many geographic regions and the prevalence of protozoan infection has likely been high throughout human history [2]. Despite the relevance of protozoan-borne diseases, few susceptibility loci have been identified. A possible approach for the identification of gene variants that modulate susceptibility to protozoan infection is to exploit the selection signatures left by these pathogens on human genes [2]. In general, pathogen-driven selection is a situation whereby infectious agents shape the genetic variability at a locus. This occurs because one or more alleles influence the susceptibility to be infected or the severity of the resulting disease.

Corresponding author: Sironi, M. (manuela.sironi@bp.lnf.it).

Protozoa diversity is a reliable estimator of protozoa-driven selective pressure

Owing to the strong selective pressure imposed by protozoa, polymorphisms that protect against these agents are expected to be at high frequencies in heavily affected populations [2]. As an example, HbS, the allele responsible for sickle hemoglobin, is maintained by balancing selection at a frequency of about 10% in regions where *Plasmodium* is endemic because heterozygotes have a greatly reduced risk of severe malaria [2]. Therefore, one possible way to identify susceptibility alleles for protozoa-borne diseases is to search for correlations between genetic variability and an estimate of the selective pressure exerted by the infectious agents in different human populations. To do this, we analyzed genotype data and measured protozoan-driven selective pressure. We exploited the availability of 660,832 single nucleotide polymorphisms (SNPs) genotyped in 52 human populations distributed worldwide (from the HGDP–CEPH panel; Table S1 in the supplementary material online) [4].

It was previously shown that pathogen diversity (i.e. number of different pathogen species) in a given geographic location is a good estimator of the pathogen-driven pressure imposed on populations living in that area [5–7]. We therefore attempted to determine whether a variable based on the estimated protozoa diversity in different geographic locations could also capture selection signatures at genes known to affect resistance to malaria (number of genes = 31, Table S2 in the supplementary material online).

Data on the prevalence of malaria were derived from two different sources: the World Health Organization (WHO, http://www.who.int) and Gideon (http://www.gideononline.com) databases, which provide non-overlapping information. The data for protozoa diversity estimate were retrieved from Gideon (see the supplementary material online); all parameters were calculated for the 21 countries where 52 HGDP–CEPH populations are located (Table S1 in the supplementary material online).

We found these two prevalence estimates to be significantly correlated (Kendall's $\tau = 0.42$, $P = 7.1 \times 10^{-5}$), and

Table 1. Correlation analysis of genes involved in resistance to malaria, hemolytic anemia and immune response

Estimate	Gene list	Number of genes[a]	Number of associated genes[b]	*P* value[c]	Contributing genes[d]
Malaria prevalence (WHO)	Malaria	31	0	n.d	–
Malaria prevalence (Gideon)	Malaria	31	0	n.d	–
Protozoa diversity	Malaria	31	11	n.d.	*PKLR, CR1, IL1B, GYPC, IL4, IL12B, CD36, ABO, IFNG, SLC4A1, ICAM1*
Protozoa diversity	Malaria plus HA	60	23	0.016	Malaria contributing genes plus *ENO1, EPB41, CFH, CD46, ADD2, GCLC, BPGM, ANK1, HK1, CD59, GSS, SPTB*
Protozoa diversity	ImmPort	2287	300	1×10^{-4}	*CD163, IFNG, IL7, STAT4, LY9, SERPINA3, FAS, CD163L1, TYK2, JAK1* and see Table S4 in the supplementary material online

[a]Number of genes with at least one SNP genotyped in the HGDP-CEPH panel.
[b]Number of genes with at least one SNP showing a significant correlation with either malaria prevalence or protozoa diversity.
[c]Empirical *P* value calculated by performing 10,000 re-samplings of randomly chosen genes (for details, see the supplementary material online). n.d., not determined.
[d]Genes showing at least one SNP significantly correlated with protozoa diversity: *CR1*, complement component (3b/4b) receptor 1; *IL1B*, interleukin 1, β; *CD36*, CD36 molecule; *ABO*, ABO blood group; *ENO1*, enolase 1; *EPB41*, erythrocyte membrane protein band 4.1; CFH, complement factor H; *CD46*, CD 46 molecule; *ADD2*, adducin 2; *GCLC*, glutamate-cysteine ligase, catalytic subunit; *BPGM*, 2,3-bisphosphoglycerate mutase; *ANK1*, ankirin 1; *HK1*, hexokinase 1; *CD59*, CD59 molecule; *GSS*, glutathione synthetase; *SPTB*, spectrin β, erythrocytic; IL7, interleukin 7; *STAT4*, signal transducer and activator of transcription 4; *LY9*, lymphocyte antigen 9; *SERPINA3*, serpin peptidase inhibitor, clade A; *FAS*, Fas (TNF receptor superfamily, member 6); *CD163L1*, CD163 molecule-like 1; TYK2, tyrosine kinase 2; *JAK1*, Janus kinase 1.

both correlated with protozoa diversity (Kendall's $\tau = 0.39$, $P = 4.2 \times 10^{-4}$ and $\tau = 0.37$, $P = 6.9 \times 10^{-4}$ for WHO and Gideon prevalence, respectively). We calculated Kendall's τ rank correlation between allele frequencies of SNPs in malaria resistance genes and either prevalence or protozoa diversity estimates. Also, given the demographic history of human populations and their non-independence, we wished to correct for isolation by distance [8,9]. Therefore, we used partial Mantel tests to account for the association between genetic and geographic distances between populations and then test for the effect of a third variable (protozoa diversity or malaria prevalence) above and beyond isolation by distance (see Methods). Specifically, we calculated partial Mantel coefficients (r_M) for the three estimates and for all SNPs in the data set. A SNP was defined as being associated with a given estimate if it displayed a significant Kendall correlation (*P* value after Bonferroni correction <0.01) and r_M higher than the 95th percentile in the distribution of all SNPs.

Eleven malaria resistance genes carried SNPs significantly associated with protozoa diversity, whereas no variant was identified using the prevalence of malaria obtained from Gideon or the WHO as a correlate (Table 1; Table S3 in the supplementary material online); these data suggest that protozoa diversity represents a better estimate of malaria-imposed selective pressure than either estimate of malaria prevalence. Notably, in addition to SNPs in genes known to be involved in immune response (e.g. interleukin 4, interleukin 12B, interleukin 1B and interferon gamma), polymorphisms in genes associated with erythrocyte homeostasis such as pyruvate kinase (*PKLR*), glycophorin C (*GYPC*) and erythrocyte membrane protein band 3 (*SLC4A1*) significantly correlated with protozoa diversity (Table 1). Given the pathogenesis of malaria infection, these results suggest that *Plasmodium* rather than other protozoa species acted as the selective agent for these variants.

Two observations might explain why protozoa diversity is a better measure of malaria-driven selective pressure than prevalence. First, prevalence estimates might be more heavily affected by poor accuracy and report biases than protozoa diversity. Second, despite our effort to account for historical notes, present-day malaria prevalence is likely to represent a weak proxy for long-term infection intensity. Conversely, it has been shown that protozoa diversity is strongly influenced by climatic variables [10], which in turn might be considered as relatively stable over time; therefore, protozoa diversity might reflect historical pressures better than the prevalence of specific infections.

Mutations in *PKLR*, *SLC4A1* and other genes result in increased resistance to malaria, but are also responsible for hemolytic anemia (HA), a condition characterized by abnormal lysis of red blood cells due to membrane or metabolic defects. We therefore explored whether variants in genes that cause HA (number of genes with at least one genotyped SNP in the HGDP-CEPH panel = 29, Table S2 in the supplementary material online) carried SNPs significantly associated with protozoa diversity. As shown in Table 1, 12 HA genes showed SNPs significantly correlated with protozoa diversity (see also Table S3 in the supplementary material online), suggesting that these loci have been shaped by the selective pressure imposed by *Plasmodium*, and therefore represent likely susceptibility/resistance genes. Our data suggest the existence of non-null alleles in HA genes that might confer a modest (compared to total deficiency) but significant protection against malaria with little fitness reduction in populations.

A simple expectation of this prediction is that HA and malaria resistance genes more often display SNPs significantly correlated with protozoa diversity compared to other human genes. Indeed this was the case. Thus, by performing 10,000 random re-samplings of 60 genes (the number of malaria plus HA genes) covered by at least one SNP in the HGDP-CEPH panel (see the supplementary material online) we verified that the probability of obtaining 23 genes with at least one significantly associated SNP amounts to 0.016 (Table 1). Notably, non-significant empirical *P* values were obtained (all *P* values >0.05), when this same analysis was performed using the diversity

of helminths, viruses and bacteria, suggesting that protozoa diversity is a specific measure for the selective pressure imposed by these organisms.

Infection with *Plasmodium*, as well as with other protozoa, elicits a strong immune response in the host; therefore, we expected SNPs in genes involved in the immune response to be associated with protozoa diversity more frequently than we observed for randomly sampled loci. We verified this prediction by analyzing the ImmPort list (The Immunology Database and Analysis Portal, http://www.immport.org) which contains 2287 genes involved in immune response and covered in the HGDP-CEPH panel. Among these genes, 300 contained at least one SNP significantly associated with protozoa diversity (Table 1; and see Table S4 in the supplementary material online), corresponding to an empirical probability of 0.0001 and confirming our prediction.

Genome-wide search for variants subjected to protozoa-driven selective pressure

These findings indicate that protozoa diversity is a reliable estimator of the selective pressure imposed by protozoa and warrant its use for a genome-wide search of significantly associated SNPs. We therefore calculated Kendall's

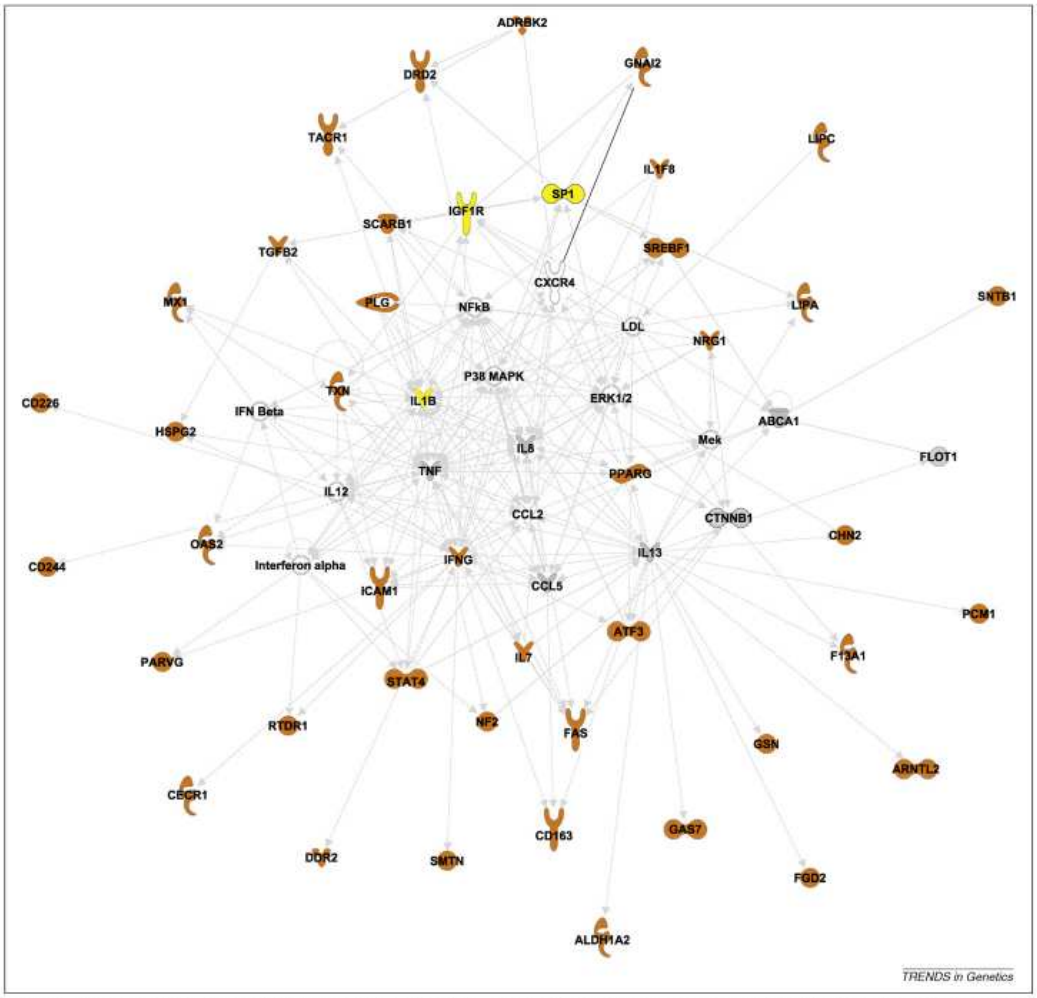


Figure 1. Network analysis of genes associated with protozoa diversity. Networks were constructed through unsupervised Ingenuity Pathway Analysis (IPA) using all genes with at least one SNP significantly associated with protozoa diversity as the input. Genes are represented as nodes, and edges indicate known interactions between proteins (unbroken lines depict direct interaction, and broken lines depict indirect interaction). Genes are color-coded as follows: orange, genes with at least one SNP significantly associated with protozoa diversity; yellow, genes with at least one SNP that did not withstand genome-wide Bonferroni correction but displayed an η_M rank higher than the 95th percentile and a *P* value lower than 10^{-6} (these genes were not included in the input IPA list used to generate networks); gray, genes with at least one SNP typed in the HGDP-CEPH panel that showed no association with protozoa diversity; white, genes with no SNP in the panel.

rank correlations between allele frequency and protozoa diversity for all SNPs ($n = 660,832$) typed in the HGDP-CEPH panel. We next searched for instances that withstood Bonferroni correction (with $\alpha = 0.05$) and displayed an r_M percentile rank higher than the 95th. A total of 5180 SNPs mapping to 1145 distinct genes satisfied both requirements (see Table S5 and Figure S1 in the supplementary material online). These SNPs are expected to be either causal or in linkage disequilibrium with the causal variant.

We wished to verify whether climatic variables could be responsible for the correlations detected between these SNPs and protozoa diversity. Hence, for all countries where HGDP-CEPH populations are located we obtained the following parameters: annual temperature (minimum, maximum and mean), short-wave radiation flux, annual precipitation rate (minimum, maximum and mean), annual relative humidity (minimum, maximum and mean) and latitude. No SNP that we identified was correlated significantly with these variables after Bonferroni correction (Table S6 in the supplementary material online) with the only exception being rs296721 (intergenic), which was correlated with both protozoa diversity and maximum precipitation rate.

Analysis of the 5180 SNPs identified indicated that they displayed, on average, a greater population genetic differentiation (F_{ST}) compared to variants with similar minor allele frequency (MAF), with the exception of variants with very low MAF (Figure S2 in the supplementary material online). These variants were located within gene regions more frequently than expected (χ^2 test, $P = 5.09 \times 10^{-5}$). This finding was verified using a MAF-matched control SNP set for comparison (χ^2 test, $P = 1.17 \times 10^{-5}$); these data are consistent with the suggestion that selection preferentially targets genic over non-genic regions [11,12].

Cellular pathways involved in response to protozoa-borne infections

Next we explored the functional relationship between genes associated with protozoa diversity. Unsupervised Ingenuity Pathway Analysis retrieved two high-scoring networks ($P < 10^{-10}$) and four networks with lower scores ($P < 10^{-8}$) (see Figure S3 in the supplementary material online). When networks 1 and 2 were merged, the resulting network (Figure 1) was organized around three major hubs: interleukin 13 (*IL13*), tumor necrosis factor (*TNF*) and interferon gamma (*IFNG*). The two latter molecules are central mediators of immunity and pathogenesis for malaria [2], whereas IL-13 is a key factor in determining susceptibility to *Leishmania major* infection in mice [13]. Among the genes in the merged network, the strongest associations with protozoa diversity were obtained for the CD163 molecule (*CD163*, $\tau = 0.72$) and the intercellular adhesion molecule 1 (*ICAM1*, $\tau = 0.68$). *CD163* encodes a scavenger receptor that mediates the internalization of both free hemoglobin and hemoglobin–haptoglobin complexes [14] and might therefore exert an important protective role in malaria by preventing intravascular hemolysis-associated tissue damage. ICAM1 mediates the adhesion of erythrocytes infected with *Plasmodium falciparum* to the endothelium, thus playing a major role in the pathogenesis of severe malaria [2].

Among the genes more strongly associated with protozoa diversity, we identified E2F transcription factor 1 (*E2F1*), which encodes a transcriptional activator recruited by NF-kB upon activation of toll-like receptor 4 (TLR4) [15], a known receptor for the glycosylphosphatidylinositol anchors of *P. falciparum* and other protozoa [16]. Although no SNP within the *TLR4* gene region is correlated with protozoa diversity, we noticed a cluster of 13 associated variants (Table S5 in the supplementary material online) scattered across a 373 kb region downstream from the transcription end site of the gene (Figure S4 in the supplementary material online). This region is therefore a good candidate to harbor regulatory variants for *TLR4* expression. Another interesting association involves *G6PC3* (glucose-6-phosphatase, catalytic subunit 3), a pivotal gene for neutrophil survival and function [17,18]. Notably, people of African ancestry tend to display lower neutrophil counts than other ethnic groups due to a high prevalence of Duffy null alleles, in turn resulting from *Plasmodium* counter-selecting the expression of this antigen [19]. Given the relevance of neutrophils for protection against malaria [20], it is possible that variants in *G6PC3* have been selected to compensate for the Duffy defect.

Finally, four genes (zinc finger homeobox 3, *ZFHX3;* CUB and Sushi multiple domains 1, *CSMD1;* calcium/calmodulin-dependent protein kinase II delta, *CAMK2D*; and diacylglycerol kinase β, *DGKB*) with at least one SNP significantly associated with protozoa diversity (Table S5 in the supplementary material online) were involved in the susceptibility to Kawasaki disease [21]. In line with these findings, *CD40LG* (CD40 ligand) was shown to be a susceptibility gene for both severe malaria [2] and Kawasaki disease [22], possibly underlying the central role that inflammation and vascular endothelia integrity plays in both conditions.

Concluding remarks

The concept whereby protozoa (and *Plasmodium* in particular) have exerted a strong selective pressure on the human genome is based mainly on the observation that erythrocyte defects conferring protection from malaria are highly prevalent in human populations [2]. Our results provide the first genome-wide estimate of protozoa-driven selective pressure, confirm that protozoa played a selective pressure of utmost importance in shaping the human genome, and indicate that previously known genetic variants represent only the tip of the iceberg [2]. In this respect, it should be noted that the approach we adopted in this work is intended to identify genes/variants that have adaptively evolved in response to protozoa (i.e. under balancing or positive selection) but is not expected to retrieve loci that have been evolutionarily constrained by protozoa-driven selective pressure. Also, the power of our method depends largely on the relative strength of selection versus gene flow/drift and is likely to be low in cases of convergent evolution (i.e. when parallel mutations with similar effects arise or are selected in different populations).

It was suggested recently that a small portion of the genetic susceptibility to malaria could be attributed to known hemoglobin gene defects [23]. Resistance to this

disease is also under a complex, multigenic control with single loci playing a small protective role [23]. The same probably applies to many other infectious diseases, suggesting that classic genome-wide association studies (GWAS) might overlook many susceptibility loci, unless extremely large cohorts are recruited. Conversely, if the selective pressure is sufficiently ancient, even a small fitness advantage can leave a signature on the allele frequency spectrum allowing inference of its role in modulating infection susceptibility or disease progression. In addition to providing insight into the evolutionary history of our species, approaches like that used here might complement and integrate GWA studies in identifying the genetic basis of resistance/susceptibility to disease.

Acknowledgments

We thank Dr Daniele Sampietro for technical assistance in retrieving data on climatic variables. M.S. is a member of the Doctorate School in Molecular Medicine, University of Milan. M.C. is supported by grants from Istituto Superiore di Sanita Programma Nazionale di Ricerca sull'AIDS, the EMPRO and AVIP EC WP6 Projects, the nGIN EC WP7 Project, the Japan Health Science Foundation, 2008 Ricerca Finalizzata (Italian Ministry of Health), 2008 Ricerca Corrente (Italian Ministry of Health), Progetto FIRB RETI: Rete Italiana Chimica Farmaceutica CHEM-PROFARMA-NET [RBPR05NWWC], and Fondazione CARIPLO.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2009.12.010.

References

1 Dronamraju, K. (ed.) (1990) *Selected Genetic Papers of J.B.S. Haldane*, Garland Publishing
2 Kwiatkowski, D.P. (2005) How malaria has affected the human genome and what human genetics can teach us about malaria. *Am. J. Hum. Genet.* 77, 171–192
3 Snow, R.W. *et al.* (2005) The global distribution of clinical episodes of *Plasmodium falciparum* malaria. *Nature* 434, 214–217
4 Li, J.Z. *et al.* (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319, 1100–1104
5 Prugnolle, F. *et al.* (2005) Pathogen-driven selection and worldwide HLA class I diversity. *Curr. Biol.* 15, 1022–1027
6 Fumagalli, M. *et al.* (2009) Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions. *J. Exp. Med.* 206, 1395–1408
7 Fumagalli, M. *et al.* (2009) Widespread balancing selection and pathogen-driven selection at blood group antigen genes. *Genome Res.* 19, 199–212
8 Handley, L.J. *et al.* (2007) Going the distance: human population genetics in a clinal world. *Trends Genet.* 23, 432–439
9 Prugnolle, F. *et al.* (2005) Geography predicts neutral genetic diversity of human populations. *Curr. Biol.* 15, R159–R160
10 Guernier, V. *et al.* (2004) Ecology drives the worldwide distribution of human diseases. *PLoS Biol.* 2, e141
11 Coop, G. *et al.* (2009) The role of geography in human adaptation. *PLoS Genet.* 5, e1000500
12 Barreiro, L.B. *et al.* (2008) Natural selection has driven population differentiation in modern humans. *Nat. Genet.* 40, 340–345
13 Matthews, D.J. *et al.* (2000) IL-13 is a susceptibility factor for Leishmania major infection. *J. Immunol.* 164, 1458–1462
14 Nielsen, M.J. and Moestrup, S.K. (2009) Receptor targeting of hemoglobin mediated by the haptoglobins: roles beyond heme scavenging. *Blood* 114, 764–771
15 Lim, C.A. *et al.* (2007) Genome-wide mapping of RELA(p65) binding identifies E2F1 as a transcriptional activator recruited by NF-kappaB upon TLR4 activation. *Mol. Cell* 27, 622–635
16 Gazzinelli, R.T. and Denkers, E.Y. (2006) Protozoan encounters with Toll-like receptor signalling pathways: implications for host parasitism. *Nat. Rev. Immunol.* 6, 895–906
17 Cheung, Y.Y. *et al.* (2007) Impaired neutrophil activity and increased susceptibility to bacterial infection in mice lacking glucose-6-phosphatase-beta. *J. Clin. Invest.* 117, 784–793
18 Boztug, K. *et al.* (2009) A syndrome with congenital neutropenia and mutations in G6PC3. *N. Engl. J. Med.* 360, 32–43
19 Reich, D. *et al.* (2009) Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. *PLoS Genet.* 5, e1000360
20 Pierrot, C. *et al.* (2007) Contribution of T cells and neutrophils in protection of young susceptible rats from fatal experimental malaria. *J. Immunol.* 178, 1713–1722
21 Burgner, D. *et al.* (2009) A genome-wide association study identifies novel and functionally related susceptibility Loci for Kawasaki disease. *PLoS Genet.* 5, e1000319
22 Onouchi, Y. *et al.* (2004) CD40 ligand gene and Kawasaki disease. *Eur. J. Hum. Genet.* 12, 1062–1068
23 Mackinnon, M.J. *et al.* (2005) Heritability of malaria in Africa. *PLoS Med.* 2, e340

0168-9525/$ – see front matter
doi:10.1016/j.tig.2009.12.010 Available online 25 January 2010

Letters

An ontology to clarify homology-related concepts

Julien Roux[1,2] and Marc Robinson-Rechavi[1,2]

[1] Université de Lausanne, Département d'Ecologie et d'Evolution, Quartier Sorge, 1015 Lausanne, Switzerland
[2] Swiss Institute of Bioinformatics, Lausanne, Switzerland

Although homology is a fundamental concept in biology and is one of the shared channels of communication universal to all biology [1], it is difficult to find a consensus definition [2]. Indeed, the interpretations of homology have changed as biology has progressed. New terms, such as paramorphism [3], have been introduced into the literature with mixed success. In addition, different research fields operate with different definitions of homology, for example the mechanistic usage of evo-devo [4] is not strictly historical and would not be acceptable in cladistics. This makes a global understanding of homology complex, whereas the integration of evolutionary concepts into bioinformatics and genomics is increasingly important. We propose an ontology organizing homology and related concepts and hope this solution will also facilitate the integration and sharing of knowledge among the community.

Corresponding author: Robinson-Rechavi, M. (marc.robinson-rechavi@unil.ch).

2.4 Genome-wide identification of susceptibility alleles for viral infections through a population genetics approach

OPEN ACCESS Freely available online PLoS GENETICS

Genome-Wide Identification of Susceptibility Alleles for Viral Infections through a Population Genetics Approach

Matteo Fumagalli[1,2], Uberto Pozzoli[1], Rachele Cagliani[1], Giacomo P. Comi[3], Nereo Bresolin[1,3], Mario Clerici[4,5☯], Manuela Sironi[1*☯]

1 Scientific Institute IRCCS E. Medea, Bioinformatic Lab, Bosisio Parini (LC), Italy, **2** Bioengineering Department, Politecnico di Milano, Milan, Italy, **3** Dino Ferrari Centre, Department of Neurological Sciences, University of Milan, IRCCS Ospedale Maggiore Policlinico, Mangiagalli and Regina Elena Foundation, Milan, Italy, **4** Department of Biomedical sciences and Technologies LITA Segrate, University of Milan, Milan, Italy, **5** Don C. Gnocchi ONLUS Foundation IRCCS, Milan, Italy

Abstract

Viruses have exerted a constant and potent selective pressure on human genes throughout evolution. We utilized the marks left by selection on allele frequency to identify viral infection-associated allelic variants. Virus diversity (the number of different viruses in a geographic region) was used to measure virus-driven selective pressure. Results showed an excess of variants correlated with virus diversity in genes involved in immune response and in the biosynthesis of glycan structures functioning as viral receptors; a significantly higher than expected number of variants was also seen in genes encoding proteins that directly interact with viral components. Genome-wide analyses identified 441 variants significantly associated with virus-diversity; these are more frequently located within gene regions than expected, and they map to 139 human genes. Analysis of functional relationships among genes subjected to virus-driven selective pressure identified a complex network enriched in viral products-interacting proteins. The novel approach to the study of infectious disease epidemiology presented herein may represent an alternative to classic genome-wide association studies and provides a large set of candidate susceptibility variants for viral infections.

Citation: Fumagalli M, Pozzoli U, Cagliani R, Comi GP, Bresolin N, et al. (2010) Genome-Wide Identification of Susceptibility Alleles for Viral Infections through a Population Genetics Approach. PLoS Genet 6(2): e1000849. doi:10.1371/journal.pgen.1000849

Editor: Harmit S. Malik, Fred Hutchinson Cancer Research Center, United States of America

Received August 24, 2009; **Accepted** January 18, 2010; **Published** February 19, 2010

Funding: MC is supported by grants from Istituto Superiore di Sanita "Programma Nazionale di Ricerca sull" AIDS, the EMPRO and AVIP EC WP6 Projects, the nGIN EC WP7 Project, the Japan Health Science Foundation, 2008 Ricerca Finalizzata [Italian Ministry of Health], 2008 Ricerca Corrente [Italian Ministry of Health], Progetto FIRB RETI: Rete Italiana Chimica Farmaceutica CHEM-PROFARMA-NET [RBPR05NWWC], and Fondazione CARIPLO. 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 contributed equally to this work.

Introduction

Infectious diseases represent one of the major threats to human populations, are still the first cause of death in developing countries [1], and are therefore a powerful selective force. In particular, viruses have affected humans before they emerged as a species, as testified by the fact that roughly 8% of the human genome is represented by recognizable endogenous retroviruses [2] which represent the fossil remnants of past infections. Also, viruses have probably acted as a formidable challenge to our immune system due to their fast evolutionary rates [3]. Indeed, higher eukaryotes have evolved mechanisms to sense and oppose viral infections; the recent identification of the antiviral activity of particular proteins such as APOBEC, tetherin, and TRIM5 has shed light on some of these mechanisms. Genes involved in anti-viral response have therefore been presumably subjected to an enormous, continuous selective pressure.

Despite the relevance of viral infection for human health, only few genome-wide association studies (GWAS) have been performed in the attempt to identify variants associated with increased susceptibility to infection or faster disease progression [4–5]. These studies have shown the presence of a small number of variants, mostly located in the HLA region. This possibly reflects the low power of GWAS to identify variants with a small effect. An alternative approach to discover variants that modulate susceptibility to viral infection is based on the identification of SNPs subjected to virus-driven selective pressure. Indeed, even a small fitness advantage can, on an evolutionary timescale, leave a signature on the allele frequency spectrum and allow identification of candidate polymorphisms. To this aim we exploited the availability of more than 660,000 SNPs genotyped in 52 human populations distributed world-wide (HGDP-CEPH panel) [6] and of epidemiological data stored in the Gideon database.

Results

Virus diversity is a reliable estimator of virus-driven selective pressure

Previous studies [7–9] have suggested that the number of the different pathogen species transmitted in a given geographic location is a good estimate of pathogen-driven selection for populations living in that area. Indeed, pathogen diversity is largely dependent on climatic factors [10] and might more closely reflect historical pressures than other estimates such as the prevalence of specific infections. We therefore reasoned that virus diversity can be used as a measure of the selective pressure exerted by virus-borne diseases on human populations and, as a consequence, that SNPs showing an unusually strong correlation

Author Summary

Viruses have represented a constant threat to human communities throughout their history, therefore, human genes involved in anti-viral response can be thought of as targets of virus-driven selective pressure. Here we utilized the marks left by selection to identify viral infection-associated allelic variants. We analyzed more than 660,000 single nucleotide polymorphisms (SNPs) genotyped in 52 human populations, and we used virus diversity (the number of different viruses in a geographic region) to measure virus-driven selective pressure. Results showed that genes involved in immune response and in the biosynthesis of glycan structures functioning as viral receptors display more variants associated with virus diversity than expected by chance. The same holds true for genes encoding proteins that directly interact with viral components. Genome-wide analysis identified 441 variants, mapping to 139 human genes, significantly associated with virus-diversity. We analyzed the functional relationships among genes subjected to virus-driven selective pressure and identified a complex interaction network enriched in viral products-interacting proteins. Therefore, we describe a novel approach for the identification of gene variants that may be involved in the susceptibility to viral infections.

with virus diversity can be considered genetic modulators of infection susceptibility or progression. To explore this possibility we used a large set of SNPs that have been genotyped in the HGDP-CEPH panel, a collection of DNAs from almost 950 individuals sampled throughout the world (Table 1). Virus diversity estimates were derived from the Global Infectious Disease and Epidemiology Network database: for each country where HGDP-CEPH populations are located we counted the number of different virus species (or genera/family as described in materials and methods) that are naturally transmitted (Table 1).

One simple prediction of our hypothesis whereby virus diversity is a reliable estimator of virus-driven selective pressure is that genes known to be involved in immune response are enriched in SNPs significantly associated with virus richness. In order to verify whether this is the case we analysed the InnateDB gene list which contains 2,915 genes involved in immune response and showing the presence of at least one SNP in the HGDP-CEPH panel. Correlations with virus richness were calculated using Kendall's partial rank correlation; since allele frequency spectra in human populations are known to be affected by demographic factors in addition to selective forces [11–12], each SNP was 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 analysed. A SNP was considered to be significantly associated with virus diversity if it displayed a significant correlation (after Bonferroni correction with $\alpha = 0.01$) and a rank higher than 0.99. As shown in Table 2, 104 SNPs in InnateDB genes showed a significant association with virus diversity. All SNPs in InnateDB genes that correlated with virus diversity are listed in Table S1. By performing 10,000 re-samplings of 2,915 randomly selected human genes (see materials and methods for details) we verified that the empirical probability of obtaining 104 significantly associated SNPs amounts to 0.010, indicating that genes in the InnateDB list display more virus-associated SNPs than expected.

It is worth mentioning that amongst these genes, *UNG* (MIM 191525), encoding uracil DNA glycosylase, functions downstream of *APOBEC3G* (MIM 607113) to mediate the degradation of nascent HIV-1 DNA [13]. *SERPING1* (MIM 606860), a regulator of the complement cascade, is also involved in HIV-1 infection (MIM 609423) as its expression is dysregulated in immature dendritic cells by Tat [14]; moreover, the protein product of *SERPING1* is cleaved by HCV and HIV-1 proteases [15–16].

Genes involved in the biosynthesis of glycan structures have also been considered as possible modulators of infection susceptibility. Indeed, since Haldane's prediction in 1949 [17] that antigens constituted of protein-carbohydrates molecules modulate the resistance/susceptibility to pathogen infection, protein glycolsylation has been shown to play a pivotal role in viral recognition of host targets [18], as well as in antigen uptake and processing and in immune modulation [19–20]. We therefore computed a list of genes involved in glycan biosynthesis from KEGG pathways and Gene Ontology annotations. Again these genes displayed significantly more virus-associated SNPs than expected if randomness alone were responsible (empirical $p = 0.0138$) (Table 2 and Table S2). Several virus-associated SNPs were located in genes coding for sialyltransferases (*ST6GAL1* (MIM 109675), *ST3GAL3* (MIM 606494), *ST6GALNAC3* (MIM 610133), *ST8SIA1* (MIM 601123), *ST3GAL1* (MIM 607187) and *ST8SIA6* (MIM 610139)). Notably, sialic acids represent the most prevalent terminal monosaccharides on the surface of human cells and determine the host range of different viruses including influenza A [21–22], polyomaviruses (i.e JCV and BKV in humans) [23], and rotaviruses (the leading cause of childhood diarrhea) [24].

Sialyltransferases also play central roles in B and T cell communication and function. In particular, the generation of influenza-specific humoral responses is impaired in mice lacking *ST6GAL1* [25], while *ST3GAL1* regulates apoptosis of CD8+ T cells [20]. Interestingly, *ST8SIA6* is expressed in NK cells, possibly playing a role in the regulation of Siglec-7 lectin inhibitory function in these cells [26]. Four other genes (*XYLT1* (MIM 608124), *HS3ST3A1* (MIM 604057), *UST* (MIM 610752) and *CHST3* (MIM 609963)) carrying SNPs associated with virus diversity are involved in the biosynthesis of either heparan sulphate or chondroitin sulphate. The former is an ubiquitously expressed glycosaminoglycan serving as the cell entry route for herpesviruses [27], HTLV-1 [28] and papillomaviruses [29]. Chondroitin sulphate is similarly expressed on a wide array of cell types and functions as an auxiliary receptor for binding of herpes simplex virus [30] as well as a facilitator of HIV-1 entry into brain microvascular endothelial cells [31]. Finally, we identified *LARGE* (MIM 603590) among the genes subjected to virus-driven selective pressure (Table 2). Recent studies have demonstrated that the post-translational modification of α-dystroglycan by LARGE is critical for the binding of arenaviruses of different phylogenetic origin including Lassa fever virus and lymphocytic-choriomeningitis virus [32–33]. Therefore our data support the previously proposed hypothesis whereby viruses represent the selective pressure underlying the strong signal of positive selection at the *LARGE* locus [34].

Since genes involved in immune response and in the biosynthesis of glycan structures are likely to be subjected to selective pressures exerted by pathogens other than viruses, we verified whether a set of genes directly involved in interaction with viral proteins also displays more SNPs significantly correlated with virus diversity. To this aim we retrieved a list of 1,916 genes known to interact with at least one viral product and displaying at least one genotyped SNP in the HGDP-CEPH panel (see materials and methods). In order to perform a non-redundant analysis, genes included in the InnateDB list and involved in glycan biosynthesis were removed; the remaining 987 genes displayed 80 SNPs correlated with virus diversity, corresponding to an empirical

Table 1. Populations in the HGDP-CEPH panel and virus diversity estimates.

Population	Country	Sampled individuals	Virus diversity
Bantu North East	Kenya	11	49
Bantu South East	South Africa	8	46
Biaka Pygmies	Central African Republic	23	54
Mandenka	Senegal	22	51
Mbuti Pygmies	Democratic Republic of Congo	13	50
San	Namibia	5	42
Yoruba	Nigeria	21	54
Colombians	Colombia	7	49
Karitiana	Brazil	14	55
Maya	Mexico	21	49
Pima	Mexico	14	49
Surui	Brazil	8	55
Balochi	Pakistan	24	45
Brahui	Pakistan	25	45
Burusho	Pakistan	25	45
Hazara	Pakistan	22	45
Kalash	Pakistan	23	45
Makrani	Pakistan	25	45
Pathan	Pakistan	23	45
Sindhi	Pakistan	24	45
Uygur	China	10	47
Cambodians	Cambodia	10	42
Dai	China	10	47
Daur	China	9	47
Han	China	44	47
Hezhen	China	9	47
Japanese	Japan	29	41
Lahu	China	8	47
Miaozu	China	10	47
Mongola	China	10	47
Naxi	China	8	47
Oroqen	China	9	47
She	China	10	47
Tu	China	10	47
Tujia	China	10	47
Xibo	China	9	47
Yakut	Russia	25	48
Yizu	China	10	47
Adygei	Russia	17	48
French	France	28	42
French Basque	France	24	42
North Italian	Italy	13	43
Orcadian	Orkney Islands (Scotland)	15	39
Russian	Russia	25	48
Sardinian	Italy	28	43
Tuscan	Italy	8	43
Bedouin	Israel	46	41
Druze	Israel	42	41
Mozabite	Algeria	29	39
Palestinian	Israel	46	41
NAN Melanesian	Papua New Guinea	11	45
Papuan	Papua New Guinea	17	45

doi:10.1371/journal.pgen.1000849.t001

p value of 0.017 (Table 2 and Table S3). Notably, when this same analysis was performed using the diversity of pathogens other than viruses (bacteria, protozoa and helminths), no significant excess of correlated SNPs was found (all empirical p values>0.05).

Genome-wide identification of variants subjected to virus-driven selective pressure

Given these results, we wished to identify SNPs significantly associated with virus richness on a genome-wide base. We therefore calculated Kendall's rank correlations between allele frequency and virus diversity for all the SNPs (n = 660,832) typed in the HGDP-CEPH panel. We next searched for instances which withstood Bonferroni correction (with $\alpha = 0.05$) and displayed a τ percentile rank higher than the 99th among MAF-matched SNPs. A total of 441 SNPs mapping to 139 distinct genes satisfied both requirements. Table 3 shows the 30 top SNPs (or SNP clusters) located within genic regions and associated with virus diversity, while the full list of SNPs subjected to virus-driven selective pressure is available on Table S4. It is worth noting that the SNP dataset we used contains less than 200 variants mapping to *HLA* genes (both class I and II), therefore covering a minor fraction of genetic variability at these loci; as a consequence *HLA* genes cannot be expected to be identified as targets of virus-driven selective pressure using the approach we describe herein.

We next verified whether the correlations detected between the SNPs we identified and virus diversity could be secondary to climatic variables. Hence, for all countries where HGDP-CEPH populations are located we obtained (see materials and methods) the following parameters: average annual minimum and maximum temperature, and short wave (UV) radiation flux. Results showed that none of the SNPs associated with virus diversity significantly correlated with any of these variables (Table S5).

Previous works have reported an enrichment of selection signatures within or in close proximity to human genes [12,35]. In line with these data we verified that virus-associated SNPs are more frequently located within gene regions compared to a control set of MAF-matched variants (χ^2 test, $p = 0.026$).

Functional characterization of genes subjected to virus-driven selective pressure

We investigated the role and functional relationship among genes subjected to virus-driven selective pressure using the Ingenuity Pathway Analysis (IPA, Ingenuity Systems) and the PANTHER classification system [36–37]. Unsupervised IPA analysis retrieved two networks with significant scores ($p = 10^{-17}$ and $p = 10^{-12}$) which were merged into a single interaction network (Figure 1). The network contains 23 genes showing a significant correlation with virus diversity and, among these, 10 encode proteins interacting with viral products (Figure 1). Based on the number of observed human-virus interactions, this finding

Table 2. Enrichment of SNPs significantly associated with virus diversity in different gene lists.

Gene list	Genes	SNPs	Corr. SNPs[a]	*p* value[b]	Contributing genes[c]
InnateDB	2915	59783	104	0.0105	*TNFRSF1B, HSPG2, KIAA0319L, PSMB2, NEGR1, CHIA, ARHGEF11, FCRLA, DDR2, HMCN1, IL19, LAMB3, TGFB2PRKCE, CLEC4F, POLR1A, LRP1B, LRP2, HDAC4, CNTN4, CLDN18, LPP, MAEA, C1QTNF7, PPP3CA, DCHS2, SEMA5A, PDZD2, SQSTM1, GMDS, GPLD1, CCND3, LAMA4, MMD2, CNTNAP2, TNFRSF10C, FREM1, COL5A1, NELL1, SERPING1, CTNND1, FCHSD2, CCND2, SCNN1A, ST8SIA1, PPFIBP1, PKP2, LIN7A, UNG, GALNTL1, BDKRB2, AQP9, IL16, CDH13, CBFA2T3, CDH15, SLFN5, DCC, FXYD5, CLDN14, DSCAM, ADARB1, TOM1, PARVG, CLDN2*
Glycan biosynthesis	200	5343	50	0.0138	*ST3GAL3, ST6GALNAC3, GALNT14, GALNT13, ST6GALT, GALNT10, UST, WBSCR17, GALNTL5, ST3GAL1, UGCG, GALNTL4, B4GALNT3, ST8SIA1, GALNT6, GALNTL1, XYLT1, CHST6, HS3ST3A1, FUT6, LARGE, ST8SIA6, CHSY3, MGAT5B, TUSC3*
Host-virus interaction	1916	14746	80	0.0172	*ENO1, CAPZB, SFRS4, SFPQ, PDE4B, MSH2, PCAF, TMEM110, GTF2E1, ADCY5, PLS1, NUP43, AKAP12, RPA3, PDE1C, ABP1, MTDH, EIF3S3, SNTB1, PCSK5, GSN, VAV2, POLR3A, PDE2A, CENTD2, RPS3, GRIN2B, PTPRO, PDE3A, ITPR2, NR4A1, POMP, RFC3, PCCA, SIPA1L1, SPTBN5, PLA2G4F, CAPN3, GTF3C1, KARS, NF1, MGAT5B, GAA, IL4I1, VPS16, PTPRA, PLCB4, SREBF2*

[a]Number of SNPs showing significant correlation with virus diversity.
[b]The empirical *p* value was calculated as described in the text and in Materials and Methods.
[c]Genes showing at least one SNP significantly correlated with virus diversity.
doi:10.1371/journal.pgen.1000849.t002

is unlikely to occur by chance (χ^2 test, $p=0.0013$) as 2.88 human-virus interactions would be expected for 23 genes. Analysis of the whole network indicated that a 31 of 66 genes encode proteins interacting with viral products (Figure 1): again this number is higher than expected (expected interactions $=8.27$; χ^2 test, $p=2.8\times10^{-10}$). Thus, the interaction network we have identified is enriched in genes subjected to virus-driven selective pressure and in genes coding for proteins interacting with viral products. It is worth mentioning that, in agreement with previous findings [38], many viral-interacting proteins represent hubs in the network. Conversely, most of the genes we found to be subjected to virus-driven selective pressure, irrespective of their ability to interact with viral proteins, tend to display very low connectivity (low-degree nodes). This observation might be consistent with previous indications [39–41] that in eukaryotes hub genes are more selectively constrained compared to low-degree nodes, these latter being more likely to evolve in response to environmental pressures.

In addition to proteins directly interacting with viral products, several network genes showing correlation with virus diversity might play central roles during viral infection. *DNMT1* (MIM 126375) and *MGMT* (MIM 156569) are involved in DNA methylation and repair, respectively, two processes that are often dysregulated during viral infection. In particular, altered expression of *DNMT1* is induced by diverse viruses including HIV-1 [42], EBV [43], BKV and adenovirsuses [44]; also, *DNMT1* plays a pivotal role in the expansion of effector CD8+ T cell following viral infection [45]. A relevant role in HIV-1 infection is also played by *HSPG2* (MIM 142461), the gene coding for perlecan, a cell surface heparan sulfate proteoglycan which mediates the internalization of Tat protein [46].

We next investigated the over-representation of PANTHER classification categories among genes subjected to virus-driven selective pressure. Table 4 shows the significantly over-represented PANTHER molecular functions and biological processes with the contributing genes. In line with the results we reported above, genes involved in immune response, as well as genes coding for proteins involved in cell adhesion and extracellular matrix components, resulted to be over-represented; these latter genes might mediate viral-cellular interaction and facilitate viral entry.

Discussion

The identification of non-neutrally evolving loci with a role in immunity can be regarded as a strategy complementary to classic clinical and epidemiological studies in providing insight into the mechanisms of host defense [47]. Here we propose that susceptibility genes for viral infections can be identified by searching for SNPs that display a strong correlation with the diversity of virus species/genera transmitted in different geographic areas. Similar approaches have previously been applied to study the adaptation to climate for genes involved in metabolism and sodium handling [48–50]. These analyses, including the one we describe herein, rely on similar assumptions and imply some caveats. First, we implicitly considered virus diversity, as we measure it nowadays, a good proxy for long-term selective pressure. This clearly represents an oversimplification, as new viral pathogens have recently emerged and the virulence of different viral species or genera might have changed over time. Still, previous studies have indicated that the geographic distribution of virus diversity is strongly influenced by climatic variables such as temperature and precipitation rates [10], suggesting that, despite significant changes in prevalence and virulence, virus diversity might have remained relatively constant across different geographic areas, possibly representing the best possible estimate of long-standing pressure. In line with these considerations, we calculated virus diversity as the number of all viral species (or genera/families) that can cause a disease in humans, irrespective of virulence or pathogenicity (Table S6).

The second issue relevant to the data we present herein is that environmental variables tend to co-vary across geographic regions: the distribution of different pathogens (e.g. parasitic worms and viruses/bacteria/protozoa) is correlated across HGDP-CEPH populations [9] and, as reported above, virus diversity is influenced by climatic factors. Therefore, our genome-wide search was preceded by analyses aimed at verifying whether virus diversity is a reliable and specific estimator of virus-driven selective pressure. In particular, we verified that genes involved in immune response and in the biosynthesis of glycans display significantly more variants associated with virus diversity than randomly selected human genes; this finding supports the idea that pathogens rather than climate or demography has driven the genetic variability at these

Table 3. Top 30 SNPs (or SNP clusters) correlated with virus diversity.

SNP	Gene symbol	Description	Annotation*	τ
rs10511316	*CCDC80*	coiled-coil domain containing 80	intron	0.627
rs1135029; rs189332; rs11235559	*PDE2A*	phosphodiesterase 2A, cGMP-stimulated	A867A; intron; intron	0.615
rs1011051; rs2278295	*MYO5C*	myosin VC	intron; intron	0.609
rs993715; rs2189883	*CNTNAP2*	contactin associated protein-like 2	intron; intron	0.609
rs11581	*KIAA1529*	-	Q1642Q	0.607
rs3785415	*CDH15*	cadherin 15, type 1, M-cadherin	intron	0.603
rs17256082	*SCRN3*	secernin 3	intron	0.600
rs4852988	*ANXA4*	annexin A4	intron	0.597
rs4575989; rs4629443	*C1QTNF7*	C1q and tumor necrosis factor related protein 7	intron; intron	0.597
rs7637370	*CLDN18*	claudin 18	intron	0.596
rs519332	*EYA4*	eyes absent 4 homolog	intron	0.596
rs2188172; rs11760238	*LHFPL3*	lipoma HMGIC fusion partner-like 3	intron; intron	0.595
rs1650893	*LOC51149*	-	Q42R	0.594
rs1322633	*RNF217*	ring finger protein 217	intron	0.593
rs7927476	*NELL1*	NEL-like 1	intron	0.593
rs2615666	*TMEM132B*	transmembrane protein 132B	intron	0.593
rs13020779	*DIS3L2*	DIS3 mitotic control homolog (S. cerevisiae)-like 2	intron	0.589
rs1719596	*LEPREL1*	leprecan-like 1	intron	0.589
rs1065154	*SQSTM1*	sequestosome 1	3' UTR	0.589
rs12145973	*IL19*	Interleukin 19	intron	0.589
rs1890139	*PCCA*	propionyl Coenzyme A carboxylase, alpha polypeptide	intron	0.588
rs6505045	*ANKFN1*	ankyrin-repeat and fibronectin type III domain containing 1	intron	0.587
rs4953260	*PRKCE*	protein kinase C, epsilon	intron	0.587
rs4077341	*TNFRSF10C*	tumor necrosis factor receptor superfamily, member 10c, decoy without an intracellular domain	intron	0.587
rs2793434	*GPLD1*	glycosylphosphatidylinositol specific phospholipase D1	intron	0.587
rs6599300	*MAEA*	macrophage erythroblast attacher	intron	0.584
rs13340461	*CCND3*	cyclin D3	intron	0.584
rs11784487	*ANK1*	Ankyrin 1	intron	0.584
rs10849446	*SCNN1A*	sodium channel, nonvoltage-gated 1 alpha	intron	0.583
rs12186418	*PDZD2*	PDZ domain containing 2	intron	0.583

*For nonsynonymous substitutions the aminoacid change is reported.
SNPs are ranked according to τ values. For multiple correlating SNPs in the same gene, the correlation coefficient is only shown for the strongest SNP.
doi:10.1371/journal.pgen.1000849.t003

loci. Notably, we also analysed genes that encode proteins interacting with viral components: since loci involved in immune response and in glycan biosynthesis were removed from this list, the remaining genes are expected to be specific targets of viral-driven selective pressure; consistently, we verified that a significant excess of SNPs correlating with virus diversity map to these loci. Conversely, a SNP excess was not noticed when the diversity of other human pathogens was used for the analysis, suggesting that, despite the correlation among different pathogen species across geographic locations [9], the selective pressure imposed by viruses can be distinguished from that exerted by other organisms.

As a further control for the possible confounding effects of other environmental factors, we verified that the variants we identified at the genome-wide level do not correlate with climate (temperature) and UV radiation. This analysis was motivated by the known association of virus diversity and biodiversity in general, with temperature [10,51] and by the fact that both climate and UV exposure have long been considered among the strongest selective pressures in humans [52]. Since none of the SNPs we identified correlated with either short wave radiation flux or temperature, we consider that their geographic distribution is likely to have been shaped by virus-driven selective pressure. In this respect it is worth mentioning that UV irradiation has been shown to be immunosuppressive in mice (reviewed in [53–54]), but the effect of sun exposure on immune functions in humans is still poorly understood. Yet, herpes viruses (both simplex and zoster) and some papillomavirus types have been shown to be reactivated by UV exposure, suggesting that the link between short wave radiation flux and virus-driven selective pressure might be more complex than simply predicted on the basis of geographic variation.

Our genome wide search for genes subjected to virus-driven selection allowed the identification of a gene interaction network that is enriched in both genes associated with virus diversity and in genes encoding proteins that interact with viral products. Many of the genes included in the identified network are of great interest as they are known to be involved in the activation of mechanisms that have direct or indirect protective effects against viruses. Thus,

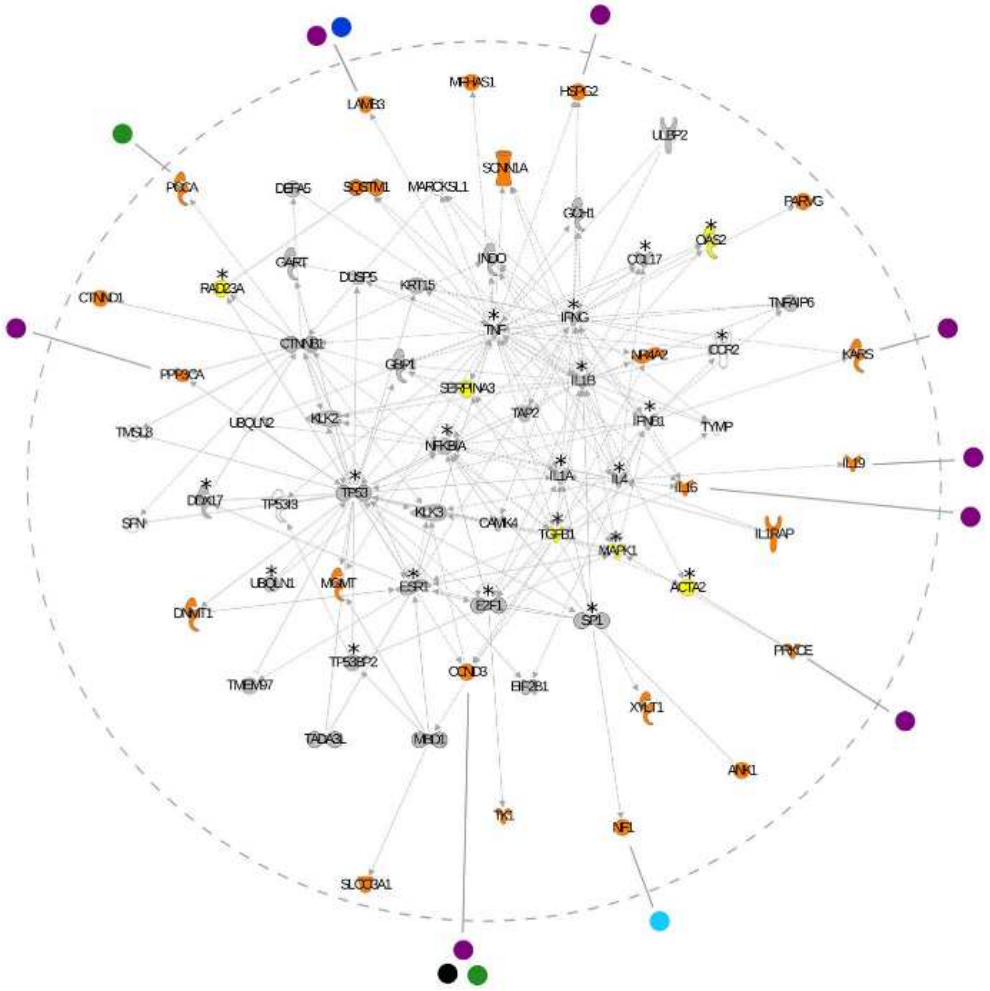


Figure 1. Network analysis of genes associated with virus diversity. Interactions between human proteins are delimited by the hatched grey circle. Genes are represented as nodes; edges indicate known interactions (sold lines depicts direct and hatched lines depict indirect interaction). Human genes are colour-coded as follows: orange, genes with at least one SNP significantly associated with virus diversity; yellow, genes with at least one SNP that did not withstand genome-wide Bonferroni correction but displayed a rank higher than the 99th and a p value lower than 10^{-5} (these genes were not included in the input IPA list used to generate networks); grey, genes covered by at least one SNP in the HGDP-CEPH panel; white, genes with no SNPs in the panel. Virus-host interactions are shown for genes subjected to virus-driven selection only; genes interacting with viral products that display no SNP significantly associated with virus diversity are denoted with an asterisk. Viral products are reported outside the hatched circle and colour coded as follows: purple, HIV-1; green, Human herpesvirus; blue, Human rotavirus G3; cyan, Human adenovirus 2; black, Human T-lymphotropic virus 1.
doi:10.1371/journal.pgen.1000849.g001

beside the well known activities of *IL1A* (MIM 147760) and *B* (MIM 147720), *IL4* (MIM 147780), *TGFB1* (MIM 190180), *IL16* (MIM 603035), *IFNG* (MIM 147570) and *TNF* (MIM 191160), *OAS2* (MIM 603350) encodes a protein that activates latent RNases, resulting in the degradation of viral RNA and in the inhibition of viral replication [55]. CCL17 (MIM 601520) induces T lymphocytes chemotaxis, thus potentiating the immune responses, and PPP3CA (MIM 114105), also known as calcineurin, activates NFATc [56], a key factor in the up-regulation of IL2 (MIM 147680) [57], the main cytokine responsible for T lymphocytes growth and differentiation. Finally, *ULBP2* (MIM 605698) encodes an MHC1-related protein that binds to NKG2D

Table 4. Significantly over-represented PANTHER categories.

PANTHER category	PANTHER description	Number of genes[a]	p value[b]
Biological process	Signal transduction	61	1.74×10^{-9}
	Cell adhesion-mediated signalling	16	9.10×10^{-6}
	Cell adhesion	16	7.98×10^{-4}
	Cell communication	24	2.79×10^{-3}
	Neuronal activities	13	1.43×10^{-2}
	Carbohydrate metabolism	13	2.05×10^{-2}
	Extracellular matrix protein-mediated signalling	5	2.47×10^{-2}
	Immunity and defense	21	3.56×10^{-2}
Molecular function	Receptor	30	4.27×10^{-5}
	Other receptor	11	3.19×10^{-4}
	Extracellular matrix linker protein	4	5.27×10^{-3}
	Extracellular matrix	10	2.29×10^{-2}

[a]Number of genes that correlate with virus diversity in each PANTHER category.
[b]*p* values are Bonferroni corrected.
doi:10.1371/journal.pgen.1000849.t004

(MIM 602893) [58], an activating receptor expressed on CD8 T cells as well as on NK cells, NKT cells and γδ T cells. In the light of the viral pathogenesis of a growing number of neoplasia, it is very interesting that other members of the network play a well described role in the inhibition of tumoral growth. In particular, *E2F1* (MIM 189971) is known to have a pivotal role in the control of cell cycle and in the activation of tumour suppressor proteins and, together with TP53I3, TADA3L, and TP53BP2 mediates p53-dependent and independent apoptosis [59–60]. *CCND3* (MIM 123834) is involved in cell cycle progression through the G2 phase, whereas *RAD23A* (MIM 600061) up-regulates the nucleotide excision activity of 3-methyladenine-DNA glycosylase [61], therefore playing a role in DNA damage recognition in base excision repair. Finally, *NR4A2* (MIM 601828) encodes a nuclear orphan receptor expressed in T cells and involved in apoptosis [62]. *NR4A2* is also known to play a central role in eliciting the production of inflammatory cytokines in multiple sclerosis (MS (MIM 126200)) [63]. Notably, variants in *PPP3CA* (Figure 1) have recently been reported to correlate with MS severity as well [64]. We therefore investigated whether other genes carrying SNPs which correlate with virus diversity have been identified in GWAS for MS susceptibility or severity. Three additional genes, *JMJD2C* (MIM 605469), *C20orf133* (also known as *MACROD2*, (MIM 611567)) and *CSMD1* (MIM 608397) have been associated with MS [64] and display SNPs significantly correlated with virus diversity (Table S1). While the function of *C20orf133* is unknown, *JMJD2C* encodes a histone demethylase expressed at very high levels in B cells and cytotoxic lymphocytes (see materials and methods), a pattern consistent with its being subjected to virus-driven selective pressure. Finally, *CSMD1*, in analogy to the aforementioned *SERPING1*, acts as a regulator of the complement system [65]; notably, complement activation plays a central role in both response to viruses and inflammatory reactions, particularly in the central nervous system [66].

Analysis of the 30 stronger associations (Table 3) indicated that several genes are part of the network described above or have been involved in immune response (see InnateDB gene list, Table 2). Conversely, others encode relatively unknown products (e.g. *KIAA1529* (MIM 611258), *LHFPL3* (MIM 609719), *LOC51149*, *RNF217*, *TMEM132B*, *LEPREL1* (MIM 610341), *ANKEN1*, *MYO5C* (MIM 610022), *ANXA4* (MIM 106491) and *SCRN3*). Among these genes, *MYO5C*, *ANXA4* and *SCRN3* are involved in membrane trafficking events along exocytotic and endocytotic pathways, suggesting that they might play a role in either viral cell entry [67] or lytic granule exocytosis; this might be the case for *ANXA4* which is expressed at high levels in NK cells (see materials and methods). Most interestingly, *ETA4* (MIM 603550) (Table 3) has recently been described as a phosphatase involved in triggering innate immune responses against viruses [68]. Finally, both *PDE2A* (MIM 602658) and *SCNN1A* (MIM 600228) might play a role in maintaining lung epithelial barrier homoeostasis during viral infection. Indeed, both genes can be induced by TNF-alpha in lung epithelial cells [69–70] and can influence lung fluid reabsorption and, therefore, edema formation. In line with these observations, expression of the amiloride-sensitive epithelial Na+ channel (*SCNN1A* codes for the α subunit) is affected by infection with influenza virus, severe acute respiratory syndrome coronavirus and respiratory syncitial virus.

In humans, resistance to infectious diseases is thought to be under complex, multigenic control with single loci playing a small protective role [47]. This concept also holds for viral infection as demonstrated by the role of genetic variants in modulating the susceptibility to HIV infection or disease progression (reviewed in [71]). Classic GWAS offer a powerful resource to identify susceptibility loci for infectious diseases; yet GWAS typically have limited power to detect variants with a low frequency or a small effect. Indeed, recent GWAS for SNPs determining the host control of HIV-1 [4–5] failed to identify most known loci with a role in AIDS progression. The alternative approach we have proposed here is based on the identification of variants subjected to virus-driven selective pressure. Similarly to the GWAS results mentioned above we did not identify well known antiviral-response genes. Still, we noticed that variants in *TRIM5* (MIM 608487) (rs2291845, $\tau = 0.44$, $p = 1.86\times10^{-5}$, rank = 0.97) and *IFIH1* (MIM 606951) (also known as *MDA5*, rs10439256, $\tau = 0.51$, $p = 5.4\times10^{-7}$, rank = 0.99) showed significant associations with virus-diversity, although they did not withstood genome-wide analysis. Also, it is worth mentioning that variants with a well established role in resistance to viral infections may be neutrally evolving; this is the case for the Δ32 allele of *CCR5* (MIM 601373) for example, which confers protection against HIV-1 infection and possibly against other pathogens, but displays no selection

signature [72]. This is possibly due to how long and how strong the selective pressure has been exerted. Conversely, variants subjected to selective pressure must have (or have had along human history) some selective advantage, indicating that the SNPs we have identified can be regarded as candidate modulators of infection susceptibility or disease progression.

Materials and Methods

Environmental variables

Virus absence/presence matrices for the 21 countries where HGDP-CEPH populations are located were derived from the Global Infectious Disease and Epidemiology Network database (Gideon, http://www.gideononline.com), a global infectious disease knowledge tool. Information in Gideon is weekly updated and derives from World Health Organization reports, National Health Ministries, PubMed searches and epidemiology meetings. The Gideon Epidemiology module follows the status of known infectious diseases globally, as well as in individual countries, with specific notes indicating the disease's history, incidence and distribution per country. We manually curated virus absence/presence matrices by extracting information from single Gideon entries. These may refer to either species, genera or families (in case data are not available for different species of a same genus/family). Following previous suggestions [7–9], we recorded only viruses that are transmitted in the 21 countries, meaning that cases of transmission due to tourism and immigration were not taken into account; also, species that have recently been eradicated as a result, for example, of vaccination campaigns, were recorded as present in the matrix. A total of 81 virus species/genera/families were retrieved (Table S6). The same approach was applied to calculate the diversity of other pathogens, namely bacteria, protozoa and helminths [9]. The annual minimum and maximum temperature were retrieved from the NCEP/NCAR database (http://www.ngdc.noaa.gov/ecosys/cdroms/ged_iia/datasets/a04/, Legates and Willmott Average, re-gridded dataset) using the geographic coordinates reported by HGDP-CEPH website for each population (http://www.cephb.fr/en/hgdp/table.php). Similarly, net short wave radiation flux data were obtained from NCEP/NCAR (http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.surfaceflux.html, Reanalysis 1: Surface Flux); these data were read using Grid Analysis and Display System (GrADS, http://www.iges.org/grads/). Daily values for four years (1948–1951) were averaged to obtain an annual mean.

Since virus diversity, due to data organization in Gideon, can only be calculated per country (rather than per population), the same procedure was applied to climatic variables. Therefore the values of annual temperature and radiation flux were averaged for populations located in the same country. This assures that a similar number of ties is maintained in all correlation analyses.

Data retrieval and statistical analysis

Data concerning the HGDP-CEPH panel derive from a previous work [6]. Atypical or duplicated samples and pairs of close relatives were removed [73].

A SNP was ascribed to a specific gene if it was located within the transcribed region or no farther than 500 bp upstream the transcription start site. MAF for any single SNP was calculated as the average over all populations. The list of immune response genes was derived from the InnateDB website (http://www.innatedb.com/) and it contains a non-redundant list of 5,070 immune genes derived from ImmPort, IRIS, Septic Shock Group, MAPK/NFKB Network and Immunome Database; it only includes genes derived from curated immune gene lists.

Genes involved in glycan biosynthesis were obtained by merging genes from two KEGG pathways ("Glycan structures - biosynthesis 1" and "Glycan structures - biosynthesis 2"). Additional genes were identified by searching Gene Ontology categories for genes that act as glycosyltransferases (GO:0016757) and are located in either the Golgi or the endoplasmic reticulum (GO:0005783, GO:0005793 and GO:0005794). The list of human genes coding for proteins interacting with viral products was derived from three sources: a previously published study [38], the VirHostNet website [74] (http://pbildb1.univ-lyon1.fr/virhostnet/) and the HIV-1 Human Protein Interaction Database [75] (http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/).

Expression data were obtained from SymAtlas (http://symatlas.gnf.org/). The location of genomic elements that are highly conserved among vertebrates was derived from UCSC annotation tables (http://genome.ucsc.edu/; "PhastCons Conserved Elements, 44-way Vertebrate Multiz Alignment" track).

All correlations were calculated by Kendall's rank correlation coefficient (τ), a non-parametric statistic used to measure the degree of correspondence between two rankings. The reason for using this test is that even in the presence of ties, the sampling distribution of τ satisfactorily converges to a normal distribution for values of n larger than 10 [76].

In order to estimate the probability of obtaining n SNPs located within m genes and significantly associated with virus diversity, we applied a re-sampling approach: samples of m genes were randomly extracted from a list of all genes covered by at least one SNP in the HGDP-CEPH panel (number of genes = 15,280) and for each sample the number of SNPs significantly associated with virus diversity was counted. The empirical probability of obtaining n SNPs was then calculated from the distribution of counts deriving from 10,000 random samples. A SNP was ascribed to a gene if it was located within the transcribed region or in the 500 upstream nucleotides.

Analysis of PANTHER over-represented functional categories and pathways was performed using the "Compare Classifications of Lists" tool available at the PANTHER classification system website [77] (http://www.pantherdb.org/). Briefly, gene lists are compared to the reference list using the binomial test for each molecular function, biological process, or pathway term in PANTHER.

All calculation were performed in the R environment [78] (http://www.r-project.org/).

Network construction

Biological network analysis was performed with Ingenuity Pathways Analysis (IPA) software using an unsupervised analysis (www.ingenuity.com). IPA builds networks by querying the Ingenuity Pathways Knowledge Base for interactions between the identified genes and all other gene objects stored in the knowledge base; it then generates networks with a maximum network size of 35 genes/proteins. We used all genes showing at least one significantly associated SNP as the input set; in this case a SNP was ascribed to a gene if it was located within the transcribed region or in the 25 kb upstream. All network edges are supported by at least one published reference or from canonical information stored in the Ingenuity Pathways Knowledge Base. To determine the probability of the analysed genes to be found together in a network from Ingenuity Pathways Knowledge Base due to random chance alone, IPA applies a Fisher's exact test. The network score represents the -log (p value).

Supporting Information

Table S1 SNPs in InnateDB genes that significantly correlate with virus diversity.

Found at: doi:10.1371/journal.pgen.1000849.s001 (0.05 MB PDF)

Table S2 SNPs in genes involved in glycan biosynthesis that significantly correlate with virus diversity.
Found at: doi:10.1371/journal.pgen.1000849.s002 (0.03 MB DOC)

Table S3 SNPs in genes coding for proteins interacting with viral products that significantly correlate with virus diversity.
Found at: doi:10.1371/journal.pgen.1000849.s003 (0.05 MB PDF)

Table S4 SNPs significantly associated with virus diversity. The table reports all SNPs that withstood Bonferroni correction at the genome-wide level (with $\alpha = 0.05$) and displayed a Tau percentile rank higher than the 99th among MAF-matched SNPs, as described in the main text and in material and methods. SNPs are ranked according to the value of Tau. If the SNP is located within a genic region (or in the 500 upstream nucleotides) the gene symbol is reported. Also, the gene closest to the SNP and its distance (in bp) are indicated. The aminoacid substitution is reported for nonsynonymous variants; SNPs annotated as "phastCons element" are located within non-coding genomic regions that display high sequence conservation among mammals (as described in the text).
Found at: doi:10.1371/journal.pgen.1000849.s004 (0.21 MB DOC)

Table S5 Correlations between SNPs associated with virus diversity and other climatic variables. The table shows correlation coefficients between each SNP associated with virus diversity and the following climatic variables: average annual maximum temperature (Tmax), average annual minimum temperature (Tmin), short wave radiation flux (Irradiation SW). After Bonferroni correction all p values were >0.05.
Found at: doi:10.1371/journal.pgen.1000849.s005 (0.42 MB DOC)

Table S6 List of viruses identified in at least one country (n = 81)
Found at: doi:10.1371/journal.pgen.1000849.s006 (0.01 MB DOC)

Acknowledgments

We wish to thank Dr. Daniele Sampietro for technical assistance in retrieving data on climatic variables. MS is a member of the Doctorate School in Molecular Medicine, University of Milan.

Author Contributions

Conceived and designed the experiments: MC MS. Performed the experiments: MF UP RC. Analyzed the data: MF UP GPC NB MS. Wrote the paper: MC MS.

References

1. Morens DM, Folkers GK, Fauci AS (2004) The challenge of emerging and re-emerging infectious diseases. Nature 430(6996): 242–249.
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409(6822): 860–921.
3. Beutler B, Eidenschenk C, Crozat K, Imler JL, Takeuchi O, et al. (2007) Genetic analysis of resistance to viral infection. Nat Rev Immunol 7(10): 753–766.
4. Limou S, Le Clerc S, Coulonges C, Carpentier W, Dina C, et al. (2009) Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS genomewide association study 02). J Infect Dis 199(3): 419–426.
5. Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, et al. (2007) A whole-genome association study of major determinants for host control of HIV-1. Science 317(5840): 944–947.
6. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319(5866): 1100–1104.
7. Prugnolle F, Manica A, Charpentier M, Guegan JF, Guernier V, et al. (2005) Pathogen-driven selection and worldwide HLA class I diversity. Curr Biol 15(11): 1022–1027.
8. 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.
9. Fumagalli M, Pozzoli U, Cagliani R, Comi GP, Riva S, et al. (2009) Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions. J Exp Med 206(6): 1395–1408.
10. Guernier V, Hochberg ME, Guegan JF (2004) Ecology drives the worldwide distribution of human diseases. PLoS Biol 2: e141. doi:10.1371/journal.pbio.0020141.
11. Handley LJ, Manica A, Goudet J, Balloux F (2007) Going the distance: Human population genetics in a clinal world. Trends Genet 23(9): 432–439.
12. Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, et al. (2009) The role of geography in human adaptation. PLoS Genet 5: e1000500. doi:10.1371/journal.pgen.1000500.
13. Yang B, Chen K, Zhang C, Huang S, Zhang H (2007) Virion-associated uracil DNA glycosylase-2 and apurinic/apyrimidinic endonuclease are involved in the degradation of APOBEC3G-edited nascent HIV-1 DNA. J Biol Chem 282(16): 11667–11675.
14. Izmailova E, Bertley FM, Huang Q, Makori N, Miller CJ, et al. (2003) HIV-1 tat reprograms immature dendritic cells to express chemoattractants for activated T cells and macrophages. Nat Med 9(2): 191–197.
15. Gerencer M, Turek V (2004) Identification of HIV-1 protease cleavage site in human C1-inhibitor. Virus Res 105(1): 97–100.
16. Drouet C, Bouillet L, Csopaki F, Colomb MG (1999) Hepatitis C virus NS3 serine protease interacts with the serpin C1 inhibitor. FEBS Lett 458(3): 415–418.
17. Dronamraju K, ed (1990) Selected genetic papers of J.B.S. haldane. New York/London: Garland Publishing.
18. Imberty A, Varrot A (2008) Microbial recognition of human cell surface glycoconjugates. Curr Opin Struct Biol 18(5): 567–576.
19. Erbacher A, Gieseke F, Handgretinger R, Muller I (2009) Dendritic cells: Functional aspects of glycosylation and lectins. Hum Immunol 70(5): 308–312.
20. Van Dyken SJ, Green RS, Marth JD (2007) Structural and mechanistic features of protein O glycosylation linked to CD8+ T-cell apoptosis. Mol Cell Biol 27(3): 1096–1111.
21. Srinivasan A, Viswanathan K, Raman R, Chandrasekaran A, Raguram S, et al. (2008) Quantitative biochemical rationale for differences in transmissibility of 1918 pandemic influenza A viruses. Proc Natl Acad Sci U S A 105(8): 2800–2805.
22. Chandrasekaran A, Srinivasan A, Raman R, Viswanathan K, Raguram S, et al. (2008) Glycan topology determines human adaptation of avian H5N1 virus hemagglutinin. Nat Biotechnol 26(1): 107–113.
23. Neu U, Stehle T, Atwood WJ (2009) The polyomaviridae: Contributions of virus structure to our understanding of virus receptors and infectious entry. Virology 384(2): 389–399.
24. Isa P, Arias CF, Lopez S (2006) Role of sialic acids in rotavirus infection. Glycoconj J 23(1-2): 27–37.
25. Zeng J, Joo HM, Rajini B, Wrammert JP, Sangster MY, et al. (2009) The generation of influenza-specific humoral responses is impaired in ST6Gal I-deficient mice. J Immunol 182(8): 4721–4727.
26. Avril T, North SJ, Haslam SM, Willison HJ, Crocker PR (2006) Probing the cis interactions of the inhibitory receptor siglec-7 with alpha2,8-disialylated ligands on natural killer cells and other leukocytes using glycan-specific antibodies and by analysis of alpha2,8-sialyltransferase gene expression. J Leukoc Biol 80(4): 787–796.
27. Shukla D, Spear PG (2001) Herpesviruses and heparan sulfate: An intimate relationship in aid of viral entry. J Clin Invest 108(4): 503–510.
28. Lambert S, Bouttier M, Vassy R, Seigneuret M, Petrow-Sadowski C, et al. (2009) HTLV-1 uses HSPG and neuropilin-1 for entry by molecular mimicry of VEGF165. Blood 113(21): 5176–5185.
29. Johnson KM, Kines RC, Roberts JN, Lowy DR, Schiller JT, et al. (2009) Role of heparan sulfate in attachment to and infection of the murine female genital tract by human papillomavirus. J Virol 83(5): 2067–2074.
30. Mardberg K, Trybala E, Tufaro F, Bergstrom T (2002) Herpes simplex virus type 1 glycoprotein C is necessary for efficient infection of chondroitin sulfate-expressing gro2C cells. J Gen Virol 83(Pt 2): 291–300.
31. Argyris EG, Acheampong E, Nunnari G, Mukhtar M, Williams KJ, et al. (2003) Human immunodeficiency virus type 1 enters primary human brain microvascular endothelial cells by a mechanism involving cell surface proteoglycans independent of lipid rafts. J Virol 77(22): 12140–12151.
32. Rojek JM, Spiropoulou CF, Campbell KP, Kunz S (2007) Old world and clade C new world arenaviruses mimic the molecular mechanism of receptor recognition used by alpha-dystroglycan's host-derived ligands. J Virol 81(11): 5685–5695.
33. Kunz S, Rojek JM, Kanagawa M, Spiropoulou CF, Barresi R, et al. (2005) Posttranslational modification of alpha-dystroglycan, the cellular receptor for arenaviruses, by the glycosyltransferase LARGE is critical for virus binding. J Virol 79(22): 14282–14296.

34. 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.
35. 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.
36. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, et al. (2003) PANTHER: A library of protein families and subfamilies indexed by function. Genome Res 13(9): 2129–2141.
37. Thomas PD, Kejariwal A, Guo N, Mi H, Campbell MJ, et al. (2006) Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. Nucleic Acids Res 34(Web Server issue): W645–50.
38. Dyer MD, Murali TM, Sobral BW (2008) The landscape of human proteins interacting with viruses and other pathogens. PLoS Pathog 4: e32. doi:10.1371/journal.ppat.0040032.
39. Albert R (2005) Scale-free networks in cell biology. J Cell Sci 118(Pt 21): 4947–4957.
40. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein interaction network. Science 296(5568): 750–752.
41. Pagel M, Meade A, Scott D (2007) Assembly rules for protein networks derived from phylogenetic-statistical analysis of whole genomes. BMC Evol Biol 7 Suppl 1: S16.
42. Youngblood B, Reich NO (2008) The early expressed HIV-1 genes regulate DNMT1 expression. Epigenetics 3(3): 149–156.
43. Hino R, Uozaki H, Murakami N, Ushiku T, Shinozaki A, et al. (2009) Activation of DNA methyltransferase 1 by EBV latent membrane protein 2A leads to promoter hypermethylation of PTEN gene in gastric carcinoma. Cancer Res 69(7): 2766–2774.
44. McCabe MT, Low JA, Imperiale MJ, Day ML (2006) Human polyomavirus BKV transcriptionally activates DNA methyltransferase 1 through the pRb/E2F pathway. Oncogene 25(19): 2727–2735.
45. Chappell C, Beard C, Altman J, Jaenisch R, Jacob J (2006) DNA methylation by DNA methyltransferase 1 is critical for effector CD8 T cell expansion. J Immunol 176(8): 4562–4572.
46. Argyris EG, Kulkosky J, Meyer ME, Xu Y, Mukhtar M, et al. (2004) The perlecan heparan sulfate proteoglycan mediates cellular uptake of HIV-1 tat through a pathway responsible for biological activity. Virology 330(2): 481–486.
47. Quintana-Murci L, Alcais A, Abel L, Casanova JL (2007) Immunology in natura: Clinical, epidemiological and evolutionary genetics of infectious diseases. Nat Immunol 8(11): 1165–1171.
48. Hancock AM, Witonsky DB, Gordon AS, Eshel G, Pritchard JK, et al. (2008) Adaptations to climate in candidate genes for common metabolic disorders. PLoS Genet 4: e32. doi:10.1371/journal.pgen.0040032.
49. Thompson EE, Kuttab-Boulos H, Witonsky D, Yang L, Roe BA, et al. (2004) CYP3A variation and the evolution of salt-sensitivity variants. Am J Hum Genet 75(6): 1059–1069.
50. Young JH, Chang YP, Kim JD, Chretien JP, Klag MJ, et al. (2005) Differential susceptibility to hypertension is due to selection during the out-of-africa expansion. PLoS Genet 1: e82. doi:10.1371/journal.pgen.0010082.
51. Allen AP, Brown JH, Gillooly JF (2002) Global biodiversity, biochemical kinetics, and the energetic-equivalence rule. Science 297(5586): 1545–1548.
52. Novembre J, Di Rienzo A (2009) Spatial patterns of variation due to natural selection in humans. Nat Rev Genet 10(11): 745–755.
53. Sleijfers A, Garssen J, Van Loveren H (2002) Ultraviolet radiation, resistance to infectious diseases, and vaccination responses. Methods 28(1): 111–121.
54. Norval M (2006) The effect of ultraviolet radiation on human viral infections. Photochem Photobiol 82(6): 1495–1504.
55. Justesen J, Hartmann R, Kjeldgaard NO (2000) Gene structure and function of the 2'-5'-oligoadenylate synthetase family. Cell Mol Life Sci 57(11): 1593–1612.
56. Crabtree GR, Olson EN (2002) NFAT signaling: Choreographing the social lives of cells. Cell 109 Suppl: S67–79.
57. Shaw JP, Utz PJ, Durand DB, Toole JJ, Emmel EA, et al. (1988) Identification of a putative regulator of early T cell activation genes. Science 241(4862): 202–205.
58. Sutherland CL, Chalupny NJ, Schooley K, VandenBos T, Kubin M, et al. (2002) UL16-binding proteins, novel MHC class I-related proteins, bind to NKG2D and activate multiple signaling pathways in primary NK cells. J Immunol 168(2): 671–679.
59. Sherr CJ (1998) Tumor surveillance via the ARF-p53 pathway. Genes Dev 12(19): 2984–2991.
60. Irwin M, Marin MC, Phillips AC, Seelan RS, Smith DI, et al. (2000) Role for the p53 homologue p73 in E2F-1-induced apoptosis. Nature 407(6804): 645–648.
61. Miao F, Bouziane M, Dammann R, Masutani C, Hanaoka F, et al. (2000) 3-methyladenine-DNA glycosylase (MPG protein) interacts with human RAD23 proteins. J Biol Chem 275(37): 28433–28438.
62. Cheng LE, Chan FK, Cado D, Winoto A (1997) Functional redundancy of the Nur77 and Nor-1 orphan steroid receptors in T-cell apoptosis. EMBO J 16(8): 1865–1875.
63. Doi Y, Oki S, Ozawa T, Hohjoh H, Miyake S, et al. (2008) Orphan nuclear receptor NR4A2 expressed in T cells from multiple sclerosis mediates production of inflammatory cytokines. Proc Natl Acad Sci U S A 105(24): 8381–8386.
64. Baranzini SE, Wang J, Gibson RA, Galwey N, Naegelin Y, et al. (2009) Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. Hum Mol Genet 18(4): 767–778.
65. Kraus DM, Elliott GS, Chute H, Horan T, Pfenninger KH, et al. (2006) CSMD1 is a novel multiple domain complement-regulatory protein highly expressed in the central nervous system and epithelial tissues. J Immunol 176(7): 4419–4430.
66. Speth C, Dierich MP, Gasque P (2002) Neuroinvasion by pathogens: A key role of the complement system. Mol Immunol 38(9): 669–679.
67. Mercer J, Helenius A (2009) Virus entry by macropinocytosis. Nat Cell Biol 11(5): 510–520.
68. Okabe Y, Sano T, Nagata S (2009) Regulation of the innate immune response by threonine-phosphatase of eyes absent. Nature.
69. Dagenais A, Frechette R, Clermont ME, Masse C, Prive A, et al. (2006) Dexamethasone inhibits the action of TNF on ENaC expression and activity. Am J Physiol Lung Cell Mol Physiol 291(6): L1220–31.
70. Seybold J, Thomas D, Witzenrath M, Boral S, Hocke AC, et al. (2005) Tumor necrosis factor-alpha-dependent expression of phosphodiesterase 2: Role in endothelial hyperpermeability. Blood 105(9): 3569–3576.
71. Piacentini L, Biasin M, Fenizia C, Clerici M (2009) Genetic correlates of protection against HIV infection: The ally within. J Intern Med 265(1): 110–124.
72. Sabeti PC, Walsh E, Schaffner SF, Varilly P, Fry B, et al. (2005) The case for selection at CCR5-Delta32. PLoS Biol 3: e378. doi:10.1371/journal.pbio.0030378.
73. 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(Pt 6): 841–847.
74. Navratil V, de Chassey B, Meyniel L, Delmotte S, Gautier C, et al. (2009) VirHostNet: A knowledge base for the management and the analysis of proteome-wide virus-host interaction networks. Nucleic Acids Res 37(Database issue): D661–8.
75. Fu W, Sanders-Beer BE, Katz KS, Maglott DR, Pruitt KD, et al. (2009) Human immunodeficiency virus type 1, human protein interaction database at NCBI. Nucleic Acids Res 37(Database issue): D417–22.
76. Salkind NJ, ed (2007) Encyclopedia of measurement and statistics. Thousand Oaks, CA: Sage Publications.
77. Cho RJ, Campbell MJ (2000) Transcription, genomes, function. Trends Genet 16(9): 409–415.
78. R Development Core Team (2008) R: A language and environment for statistical computing. Vienna, Austria.

2.5 Population genetics of *IFIH1*: ancient population structure, local selection and implications for susceptibility to type 1 diabetes

MOLECULAR BIOLOGY AND EVOLUTION

MBE Advance Access published June 10, 2010

Population genetics of *IFIH1*: ancient population structure, local selection and implications for susceptibility to type 1 diabetes

Type: research article

Matteo Fumagalli[1,2], Rachele Cagliani[1], Stefania Riva[1], Uberto Pozzoli[1], Mara Biasin[3], Luca Piacentini[3], Giacomo P. Comi[4], Nereo Bresolin[1,4], Mario Clerici[5,6], Manuela Sironi[1*]

[1]Scientific Institute IRCCS E. Medea, Via don L. Monza 20, 23842 Bosisio Parini (LC), Italy.

[2]Bioengineering Department, Politecnico di Milano, P.zza L. da Vinci, 32, 20133 Milan, Italy.

[3]Chair of Immunology, DISP LITA Vialba, University of Milano, Milano, Italy

[4]Dino Ferrari Centre, Department of Neurological Sciences, University of Milan, IRCCS Ospedale Maggiore Policlinico, Mangiagalli and Regina Elena Foundation, Via F. Sforza 35, 20100 Milan, Italy.

[5]Chair of Immunology, Department of Biomedical Sciences and Technologies LITA Segrate, University of Milano, Milano, Italy

[6]Fondazione Don C. Gnocchi, IRCCS, Milano, Italy.

* Address for correspondence:

Manuela Sironi

Scientific Institute IRCCS E. Medea, Bioinformatic Lab, Via don L. Monza 20, 23842 Bosisio Parini (LC), Italy.

Tel: +39031877915

Fax: +39031877499

E-mail: manuela.sironi@BP.LNF.it.

Abstract

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

Introduction

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

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

Recent studies have shown that polymorphisms in *IFIH1* also play a relevant role in the pathogenesis of autoimmune diseases. Both common and rare polymorphic variants in the gene have been reproducibly associated with the susceptibility to type 1 diabetes (T1D) (Jermendy et al. 2010; Liu et al. 2009; Nejentsev et al. 2009; Shigemoto et al. 2009; Smyth et al. 2006). In particular, the derived T allele of rs1990760 (exon 15, Ala946Thr) is common in Europeans and correlates with an increased risk to develop the disease (Jermendy et al. 2010; Liu et al. 2009; Nejentsev et al. 2009; Smyth et al. 2006), while the association of this same variant with susceptibility to rheumatoid arthritis, autoimmune thyroid disease and multiple sclerosis is more

controversial (Couturier et al. 2009; Enevold et al. 2009; Marinou et al. 2007; Martinez et al. 2008a; Martinez et al. 2008b; Penna-Martinez et al. 2009; Sutherland et al. 2007). In the case of T1D, rare variants that decrease or disable *IFIH1* expression have a protective role (Nejentsev et al. 2009), while higher gene expression is observed in peripheral blood mononuclear cells of individuals carrying the common susceptible genotype (Liu et al. 2009).

On the one hand, these data suggest that increased efficiency of *IFIH1* transcription or protein function may be associated with the development of autoimmunity. On the other hand, given the central role of *IFIH1* in antiviral response, it is conceivable that sustained activity of this helicase might confer strong protection against infections and therefore be favoured by natural selection. Recent studies addressing the relationship between adaptation and immune diseases in humans (Barreiro and Quintana-Murci 2010; Fumagalli et al. 2009b) suggested that a portion of risk alleles has been selected because they could provide protection against infectious diseases.

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

Materials and Methods

DNA samples and sequencing/genotyping

Human genomic DNA for Europeans (CEU), Yoruba (YRI), Asians (AS) and South Americans (SAM) was obtained from the Coriell Institute for Medical Research. The analyzed region was PCR amplified in overlapping fragments and directly sequenced; primer sequences are available upon request. 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), inspected manually by two distinct operators and singletons were re-amplified and resequenced.

Data retrieval and haplotype construction

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

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

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

Data concerning the HGDP-CEPH panel derive from a previous work (Li et al. 2008). Atypical or duplicated samples and pairs of close relatives were removed (Rosenberg 2006). Following

previous indications (Fumagalli et al. 2009a; Fumagalli et al. 2009b), Bantu individuals (South Africa) were considered as one population.

Annotation regarding conserved sequences among different species was derived from UCSC Genome Browser (http://genome.ucsc.edu/, phastConsElements28wayPlacMammal table).

We calculated CNS densities (proportion of conserved bases per intron length) for a set of 20.978 non alternative introns. Density distributions (percentiles) were calculated independently for 10 intron-length classes (breaks: 1,128,294,549,886,1300,1880,2700,4080,8020,53600). CNS densities for *IFIH1* introns 3 and 4 have been compared to the corresponding class distribution based on their length and correspond to the 97th and 99th percentiles, respectively.

Statistical analysis

A detailed description of all tests we applied and their meaning is available as supplementary material (Table S2).

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

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

Composite-likelihood-ratio test and coalescent simulations under a selective sweep regime were performed using *clsw* and ss*w* programs kindly provided by Yuseob Kim.

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

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

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

The maximum-likelihood-ratio HKA test was performed using the MLHKA software (Wright and Charlesworth 2004), as previously proposed (Fumagalli et al. 2009a). Briefly, 16 reference loci were randomly selected among NIEHS loci shorter than 20 kb that have been resequenced in the 3 populations; the only criterion was that Tajima's D did not suggest the action of natural selection (i.e. Tajima's D is higher than the 5th and lower than the 95th percentiles in the distribution of NIEHS genes). The reference set was accounted for by the following genes: *VNN3* (MIM 606592), *PLA2G2D* (MIM 605630), *MB* (MIM 160000), *MAD2L2* (MIM 604094), *HRAS* (MIM 190020), *CYP17A1* (MIM 609300), *ATOX1* (MIM 602270), *BNIP3* (MIM 603293), *CDC20* (MIM 603618), *NGB* (MIM 605304), *TUBA1* (MIM 191110), *MT3* (MIM 139255), *NUDT1* (MIM 600312), *PRDX5* (MIM 606583), *RETN* (MIM 605565) and *JUND* (MIM 165162).

The *MWUhigh* test was performed as previously described (Andres et al. 2009; Nielsen et al. 2009). Significance was assessed by performing 10000 coalescent simulations conditioned on the number of segregating sites and incorporating demographic scenarios (Schaffner et al. 2005).

Median-joining networks to infer haplotype genealogy was constructed using NETWORK 4.5 (Bandelt, Forster, Rohl 1999). Estimate of the time to the most common ancestor (TMRCA) was obtained using a phylogeny based approach implemented in NETWORK based on the average

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

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

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

The maximum-likelihood-ratio HKA test was performed using the MLHKA software (Wright and Charlesworth 2004), as previously proposed (Fumagalli et al. 2009a). Briefly, 16 reference loci were randomly selected among NIEHS loci shorter than 20 kb that have been resequenced in the 3 populations; the only criterion was that Tajima's D did not suggest the action of natural selection (i.e. Tajima's D is higher than the 5th and lower than the 95th percentiles in the distribution of NIEHS genes). The reference set was accounted for by the following genes: *VNN3* (MIM 606592), *PLA2G2D* (MIM 605630), *MB* (MIM 160000), *MAD2L2* (MIM 604094), *HRAS* (MIM 190020), *CYP17A1* (MIM 609300), *ATOX1* (MIM 602270), *BNIP3* (MIM 603293), *CDC20* (MIM 603618), *NGB* (MIM 605304), *TUBA1* (MIM 191110), *MT3* (MIM 139255), *NUDT1* (MIM 600312), *PRDX5* (MIM 606583), *RETN* (MIM 605565) and *JUND* (MIM 165162).

The *MWUhigh* test was performed as previously described (Andres et al. 2009; Nielsen et al. 2009). Significance was assessed by performing 10000 coalescent simulations conditioned on the number of segregating sites and incorporating demographic scenarios (Schaffner et al. 2005).

Median-joining networks to infer haplotype genealogy was constructed using NETWORK 4.5 (Bandelt, Forster, Rohl 1999). Estimate of the time to the most common ancestor (TMRCA) was obtained using a phylogeny based approach implemented in NETWORK based on the average

distance from the root (Morral et al. 1994; Saillard et al. 2000) and using a mutation rate based on the number of fixed differences between human and chimpanzee and assuming a separation time from humans of 6 MY ago (Glazko and Nei 2003). A second TMRCA estimate derived from application of a maximum-likelihood coalescent method implemented in GENETREE (Griffiths and Tavare 1994; Griffiths and Tavare 1995). Again, the mutation rate μ was obtained on the basis of the divergence between human and chimpanzee and under the assumption of a generation time of 25 years. Using this μ and θ maximum likelihood (θ_{ML}), we estimated the effective population size parameter (Ne). With these assumptions, the coalescence time, scaled in 2Ne units, was converted into years. For the coalescence process, 10^6 simulations were performed.

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

Results

Nucleotide diversity and haplotype structure

We had previously identified a SNP in *IFIH1* that strongly correlates with virus diversity in human populations, suggesting that this variant or a linked one has been subjected to virus-driven selective pressure (Fumagalli et al. 2010). The SNP (rs10439256) is located within intron 4 and falls within a region that is highly conserved among mammals. In general, the region surrounding exon 4 harbours several conserved non-coding sequences (CNS) (Fig. 1); a comparison with the intronic density of CNSs in human genes indicated that both intron 3 and 4 display an exceptionally high number of conserved sequences (see methods), suggesting the presence of gene regulatory elements. As shown in figure 2, the allele frequency distribution for rs10439256 displays high continental differentiation: the ancestral T allele is fixed or almost fixed in Europe, the Middle East and Asia, while the C derived allele shows intermediate or high frequency in Africa and Central/South America, respectively. In line with this observation, analysis of population genetic

differentiation (F_{ST}) across 52 human populations indicated that rs10439256 displays the highest F_{ST} among SNPs genotyped in the HGDP-CEPH panel and located within *IFIH1* (Fig. 1). In particular, a striking difference in allele frequency is observed between populations living in East Asia and those in Central/South America (Fig. 2), suggesting local adaptation. In order to explore this possibility we calculated the distribution of F_{ST} values for all SNPs in the HGDP-CEPH panel (more than 660,000 variants, (Li et al. 2008)) between the Yakut (located in Siberia and considered as the closest ancestors of modern Americans, (Li et al. 2008)) and the Maya. rs10439256 displayed an F_{ST} of 0.32, which corresponds to a percentile rank of 0.957 in the distribution of HGDP-CEPH SNPs, suggesting the action of local selective pressures on *IFIH1* in the Americas.

To gain further insight into the evolutionary history of the gene, we resequenced a ~10 kb region encompassing rs10439256 and covering several CNSs (Fig. 1) in 3 HapMap populations, namely Yoruba (YRI), Europeans (CEU) and East Asians (AS), as well as in South Americans (SAM). The number of SNPs identified in each population is reported in table 1, together with θ_W and π, two nucleotide diversity measures ((Nei and Li 1979; Watterson 1975)). As an empirical comparison, θ_W and π were also calculated for 5 kb windows deriving from 238 genes resequenced by the NIEHS program in YRI, CEU and AS (no extensive resequencing data are available for SAM) (see table S3 for a comparison with 10 kb windows). As shown in table 1, the CEU sample displayed a significant reduction of nucleotide diversity in this gene region.

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

In order to analyse the haplotype structure of the resequenced *IFIH1* gene region, we constructed a median-joining network (Fig. 3). A genealogy with two major, deeply separated clades (A and B) was evident. In line with the F_{ST} results, haplotype frequency is extremely diverse across

populations with all CEU chromosomes clustering in clade B together with the majority of Asian haplotypes. Conversely, most SAM chromosomes belong to clade A, and account for population-specific haplogroups.

In order to estimate the TMRCA (Time to the Most Recent Common Ancestor) of the *IFIH1* haplotype genealogy, we applied a phylogeny-based method (Bandelt, Forster, Rohl 1999). Using a mutation rate based on 63 fixed differences between chimpanzees and humans and a separation time of 6 million years (MY) (Glazko and Nei 2003), we estimated a TMRCA of 3.01 MY (SD: 0.531 MY). In order to obtain a more robust estimate and given the relatively low recombination rate in the region, we calculated a second TMRCA using GENETREE, which is based on a maximum-likelihood coalescent analysis (Griffiths and Tavare 1995). The method assumes an infinite-site model without recombination: 8 segregating sites and 1 haplotype had to be removed as they violated these assumptions. The resulting gene tree, rooted using the chimpanzee sequence, is partitioned into two deep branches (Fig. 4). A maximum-likelihood estimate of θ (θ_{ML}) of 9.4 was obtained, resulting in an estimated effective population size (N_e) of 17904. Using this method, the TMRCA of the *IFIH1* haplotype lineages amounted to 2.31 MY (SD: 0.537 MY). These TMRCA estimates are deeper than those obtained for the great majority of neutrally evolving autosomal loci (Garrigan and Hammer 2006; Tishkoff and Verrelli 2003).

Neutrality tests

We calculated Tajima's D (D_T) (Tajima 1989), Fu and Li's F* and D* (Fu and Li 1993), as well as Fay and Wu' s *H* (Fay and Wu 2000) for the resequenced *IFIH1* region and evaluated whether these statistics significantly deviate from expectations under neutrality using both coalescent simulations and the empirical distribution of 5 kb reference windows (see table S3 for a comparison with 10 kb windows). For coalescent simulations, we applied models that incorporate demographic scenarios (see methods). Results are summarized in table 2 and indicate that D_T is significantly high in YRI

but no other test rejects neutrality in this population. Conversely, low values of D_T, D* and F* were obtained for CEU with borderline significant *p* values. As for AS and SAM, D* was unusually high in both populations and *H* was significantly negative; this latter result indicates that AS and SAM display an excess of high frequency derived alleles.

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

Selection pattern in human populations and possible selection targets

As determined above, *IFIH1* haplotypes display an unusually deep coalescent time. There are at least two possible explanations for this finding: long-standing balancing selection and ancient population structure in the African continent. The two processes originate different gene tree topologies in that balancing selection elongates the whole neutral genealogy, while population structure results in a longer proportion of genealogical time occupied by single lineages (i.e. longer than expected basal branches) (Takahata 1990; Wall 2000). Therefore, admixture of structured populations is expected to result in a specific pattern of linkage disequilibrium (LD). Plagnol and Wall (Plagnol and Wall 2006) recently proposed a new measure of LD, denoted S*, that is specifically devised to test for an excess of congruent or almost congruent mutations. Calculation of S* for the YRI sample resulted in a value of 69,014; coalescent simulations with different demographic models yielded the following *p* values: 0.038 (Schaffner et al. 2005), 0.034 (Voight et al. 2005) and 0.017 (Gutenkunst et al. 2009) (see methods). These data are therefore consistent with

the idea that the two major clades of the *IFIH1* genealogy result from ancient population admixture in Africa. Yet, it is worth mentioning that a more complex situation of long-standing balancing selection with episodic selective sweeps or with multiple positively selected alleles (see below) might also result in unusual LD patterns.

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

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

population bottlenecks). For CEU, AS and SAM, the GOF *p* values suggest that rejection of neutrality by the CLR test is more likely due to a selective sweep than to demographic history (Tab. 3). These data therefore support the idea that different *IFIH1* haplotypes have increased in frequency in CEU/AS and SAM as a result of directional selection.

It has recently been shown that biased gene conversion (BGC) affects neutral substitution patterns (reviewed in (Duret and Galtier 2009)); the effect of BGC is particularly strong in sub-telomeric regions and in regions with high male-specific recombination rates (Dreszer et al. 2007; Duret and Arndt 2008; Webster et al. 2005). *IFIH1* is not sub-telomeric and male-specific recombination rate along the gene amounts to 0.4 cM/Mb, a value substantially lower than the genome average for autosomes (0.98 cM/Mb) (Kong et al. 2002). In the region we analysed 17 polymorphisms (those located on the branch leading to clade B haplotypes) display a high frequency of the derived allele (frequency > 0.6 averaged over all populations); of these only 7 are A/T -> G/C mutations, one is a T -> A and the remaining are G/C -> T/A substitutions. Overall, these data suggest that BGC does not play a major role in shaping nucleotide variability at *IFIH1*.

The rise in frequency of a selected allele may generate an extended haplotype which results from the long-range association with nearby polymorphisms - depending on the timing and strength of the selective event. To verify whether this is the case for *IFIH1*, we derived empirical *p* values for two haplotype-based tests, namely iHS (Voight et al. 2006) and XP-EHH (Sabeti et al. 2007) from the HGDP Selection Browser (http://hgdp.uchicago.edu/cgi-bin/gbrowse/HGDP/) (Pickrell et al. 2009). While no exceptional value was observed for CEU and AS, the two tests were significant for HGDP-CEPH populations living in America. Specifically, the iHS test was significant over a large genomic portion encompassing *IFIH1* and several flanking genes while the XP-EHH test showed significant *p* values over a narrower region encompassing *IFIH1* and partially extending into the *FAP* gene, with a peak being observed within *IFIH1* (Fig. 5).

As mentioned above, two nonsynonymous variants in *IFIH1*, Arg843His (rs3747517) and

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

Discussion

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

MWUhigh test calculated on the folded spectrum yielded a significant result for YRI (p= 0.021; see methods) but not for the other populations (not shown), suggesting a role for balancing selection in maintaining the two major haplotype clades. Therefore, while unusual LD patterns (as measured by S*) are not expected in a situation of balancing selection, an excess of intermediate frequency alleles (as obtained by the *MWUhigh* test) is not consistent with a scenario of ancient population structure. Nonetheless, complex evolutionary scenarios following admixture or balancing selection may affect both allele frequency and LD patterns. Also, these two possibilities are not necessarily mutually exclusive as a haplotype introduced by admixture may have higher chances to be detected in modern populations when subjected to a selective event (that opposes its chances of being lost by drift) (Hawks et al. 2008).

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

Asian and American populations, strongly suggesting that local selective pressures resulted in the spread of different alleles. The significant XP-EHH test is in line with this observation and suggests a relatively recent selective sweep event in SAM.

IFIH1 plays a central role in antiviral response and rs10439256 was previously indicated as being targeted by virus-driven selective pressure (Fumagalli et al. 2010). In the populations we analysed this variant is in strong linkage disequilibrium with the His460Arg SNP (Table S4), which affects a highly conserved residue located in the helicase domain. Notably, the helicase domain is directly involved in viral RNA binding (Takeuchi and Akira 2008), suggesting that variants in this region alter the specificity of IFIH1 against one or more viral species. Therefore, it is tempting to speculate that the derived 460His allele has been the target of positive selection in European and Asian populations, possibly because it confers increased resistance to one or more viruses in these geographic areas. This would also imply that the Arg460His polymorphism represents a susceptibility/protective variant against diverse viral infections. With respect to South American populations, we noticed that two variants (rs12474958 and rs12478730, positions 31 and 32 in figure 3) define all SAM chromosomes in clade A and occur within a CNS. Whether these variants affect *IFIH1* regulation remains to be evaluated, as well as the possibility that they represent the selection target in this population. In this respect, it is worth mentioning that selective sweeps typically affect large genomic regions due to genetic hitch-hiking; thus, inference on the real selection target is difficult unless functional information is available.

The structure and distribution of *IFIH1* haplotypes described herein may harbor consequences for association studies, especially when populations of non-European descent are being analysed. The derived allele of rs1990760 was shown to predispose to T1D in Caucasians (Jermendy et al. 2010; Liu et al. 2009; Nejentsev et al. 2009; Smyth et al. 2006). As shown in figure 3, all CEU chromosomes belong to clade B, and both the ancestral and risk alleles of rs1990760 occur on the same haplotype background in this population. This is not the case for YRI and AS given that

Asian and American populations, strongly suggesting that local selective pressures resulted in the spread of different alleles. The significant XP-EHH test is in line with this observation and suggests a relatively recent selective sweep event in SAM.

IFIH1 plays a central role in antiviral response and rs10439256 was previously indicated as being targeted by virus-driven selective pressure (Fumagalli et al. 2010). In the populations we analysed this variant is in strong linkage disequilibrium with the His460Arg SNP (Table S4), which affects a highly conserved residue located in the helicase domain. Notably, the helicase domain is directly involved in viral RNA binding (Takeuchi and Akira 2008), suggesting that variants in this region alter the specificity of IFIH1 against one or more viral species. Therefore, it is tempting to speculate that the derived 460His allele has been the target of positive selection in European and Asian populations, possibly because it confers increased resistance to one or more viruses in these geographic areas. This would also imply that the Arg460His polymorphism represents a susceptibility/protective variant against diverse viral infections. With respect to South American populations, we noticed that two variants (rs12474958 and rs12478730, positions 31 and 32 in figure 3) define all SAM chromosomes in clade A and occur within a CNS. Whether these variants affect *IFIH1* regulation remains to be evaluated, as well as the possibility that they represent the selection target in this population. In this respect, it is worth mentioning that selective sweeps typically affect large genomic regions due to genetic hitch-hiking; thus, inference on the real selection target is difficult unless functional information is available.

The structure and distribution of *IFIH1* haplotypes described herein may harbor consequences for association studies, especially when populations of non-European descent are being analysed. The derived allele of rs1990760 was shown to predispose to T1D in Caucasians (Jermendy et al. 2010; Liu et al. 2009; Nejentsev et al. 2009; Smyth et al. 2006). As shown in figure 3, all CEU chromosomes belong to clade B, and both the ancestral and risk alleles of rs1990760 occur on the same haplotype background in this population. This is not the case for YRI and AS given that

chromosomes harboring the ancestral allele are represented both in clade A and B. If, as expected, haplotypes in the two clades are functionally different, association studies in non-European populations using rs1990760 might be hindered by haplotype heterogeneity for the protective allele. Given that T1D is characterized by a juvenile onset and is a potentially lethal disease, it is unclear why alleles that predispose to this condition have not been eliminated by natural selection.

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

Supplementary material is available online

Acknowledgements

MC is supported by grants from Istituto Superiore di Sanita' "Programma Nazionale di Ricerca sull' AIDS", the EMPRO and AVIP EC WP6 Projects, the nGIN EC WP7 Project, the Japan Health Science Foundation, 2008 Ricerca Finalizzata [Italian Ministry of Health], 2008 Ricerca Corrente [Italian Ministry of Health], Progetto FIRB RETI: Rete Italiana Chimica Farmaceutica CHEM-PROFARMA-NET [RBPR05NWWC], and Fondazione CARIPLO.
MS is a member of the Doctorate School in Molecular Medicine, University of Milan.

Literature cited

Andrejeva J, Childs KS, Young DF, Carlos TS, Stock N, Goodbourn S, Randall RE. 2004. 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. U. S. A. 101:17264-17269.

Andres AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, Boyko AR, Gutenkunst RN, White TJ, Green ED, Bustamante CD et al. 2009. Targets of balancing selection in the human genome. Mol. Biol. Evol. 26:2755-2764.

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

Barreiro LB, Quintana-Murci L. 2010. From evolutionary genetics to human immunology: How selection shapes host defence genes. Nat. Rev. Genet. 11:17-30.

Barreiro LB, Patin E, Neyrolles O, Cann HM, Gicquel B, Quintana-Murci L. 2005. The heritage of pathogen pressures and ancient demography in the human innate-immunity CD209/CD209L

region. Am. J. Hum. Genet. 77:869-886.

Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet. 2:e64.

Couturier N, Gourraud PA, Cournu-Rebeix I, Gout C, Bucciarelli F, Edan G, Babron MC, Clerget-Darpoux F, Clanet M, Fontaine B et al. 2009. IFIH1-GCA-KCNH7 locus is not associated with genetic susceptibility to multiple sclerosis in french patients. Eur. J. Hum. Genet. 17:844-847.

Cox MP, Mendez FL, Karafet TM, Pilkington MM, Kingan SB, Destro-Bisol G, Strassmann BI, Hammer MF. 2008. Testing for archaic hominin admixture on the X chromosome: Model likelihoods for the modern human RRM2P4 region from summaries of genealogical topology under the structured coalescent. Genetics 178:427-437.

Dreszer TR, Wall GD, Haussler D, Pollard KS. 2007. Biased clustered substitutions in the human genome: The footprints of male-driven biased gene conversion. Genome Res. 17:1420-1430.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu. Rev. Genomics Hum. Genet. 10:285-311.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. 4:e1000071.

Enevold C, Oturai AB, Sorensen PS, Ryder LP, Koch-Henriksen N, Bendtzen K. 2009. Multiple sclerosis and polymorphisms of innate pattern recognition receptors TLR1-10, NOD1-2, DDX58, and IFIH1. J. Neuroimmunol. 212:125-131.

Fay JC, Wu CI. 2000. Hitchhiking under positive darwinian selection. Genetics 155:1405-1413.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. Genetics 133:693-709.

Fumagalli M, Pozzoli U, Cagliani R, Comi GP, Bresolin N, Clerici M, Sironi M. 2010. Genome-wide identification of susceptibility alleles for viral infections through a population genetics approach. PLoS Genet. 6:e1000849.

Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi GP, Menozzi G, Bresolin N, Sironi M. 2009a. Widespread balancing selection and pathogen-driven selection at blood group antigen genes. Genome Res. 19:199-212.

Fumagalli M, Pozzoli U, Cagliani R, Comi GP, Riva S, Clerici M, Bresolin N, Sironi M. 2009b. Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions. J. Exp. Med. 206:1395-1408.

Garrigan D, Hammer MF. 2006. Reconstructing human origins in the genomic era. Nat. Rev. Genet. 7:669-680.

Garrigan D, Mobasher Z, Kingan SB, Wilder JA, Hammer MF. 2005. Deep haplotype divergence and long-range linkage disequilibrium at xp21.1 provide evidence that humans

descend from a structured ancestral population. Genetics 170:1849-1856.

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 :184-186.

Griffiths RC, Tavare S. 1995. Unrooted genealogical tree probabilities in the infinitely-many-sites model. Math. Biosci. 127:77-98.

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.

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.

Hawks J, Cochran G, Harpending HC, Lahn BT. 2008. A genetic legacy from archaic homo. Trends Genet. 24:19-23.

Jensen JD, Kim Y, DuMont VB, Aquadro CF, Bustamante CD. 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170:1401-1410.

Jermendy A, Szatmari I, Laine AP, Lukacs K, Horvath KH, Korner A, Madacsy L, Veijola R,

Simell O, Knip M et al. 2010. The interferon-induced helicase IFIH1 Ala946Thr polymorphism is associated with type 1 diabetes in both the high-incidence finnish and the medium-incidence hungarian populations. Diabetologia 53:98-102.

Kato H, Takeuchi O, Sato S, Yoneyama M, Yamamoto M, Matsui K, Uematsu S, Jung A, Kawai T, Ishii KJ et al. 2006. Differential roles of MDA5 and RIG-I helicases in the recognition of RNA viruses. Nature 441:101-105.

Kim HL, Satta Y. 2008. Population genetic analysis of the N-acylsphingosine amidohydrolase gene associated with mental activity in humans. Genetics 178:1505-1515.

Kim Y, Stephan W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160:765-777.

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

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G et al. 2002. A high-resolution recombination map of the human genome. Nat. Genet. 31:241-247.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100-1104.

Liu S, Wang H, Jin Y, Podolsky R, Reddy MV, Pedersen J, Bode B, Reed J, Steed D, Anderson S et al. 2009. IFIH1 polymorphisms are significantly associated with type 1 diabetes and IFIH1 gene expression in peripheral blood mononuclear cells. Hum. Mol. Genet. 18:358-365.

Marinou I, Montgomery DS, Dickson MC, Binks MH, Moore DJ, Bax DE, Wilson AG. 2007. The interferon induced with helicase domain 1 A946T polymorphism is not associated with rheumatoid arthritis. Arthritis Res. Ther. 9:R40.

Martinez A, Varade J, Lamas JR, Fernandez-Arquero M, Jover JA, de la Concha EG, Fernandez-Gutierrez B, Urcelay E. 2008a. Association of the IFIH1-GCA-KCNH7 chromosomal region with rheumatoid arthritis. Ann. Rheum. Dis. 67:137-138.

Martinez A, Santiago JL, Cenit MC, de Las Heras V, de la Calle H, Fernandez-Arquero M, Arroyo R, de la Concha EG, Urcelay E. 2008b. IFIH1-GCA-KCNH7 locus: Influence on multiple sclerosis risk. Eur. J. Hum. Genet. 16:861-864.

Meylan E, Tschopp J, Karin M. 2006. Intracellular pattern recognition receptors in the host response. Nature 442:39-44.

Morral N, Bertranpetit J, Estivill X, Nunes V, Casals T, Gimenez J, Reis A, Varon-Mateeva R, Macek M,Jr, Kalaydjieva L. 1994. The origin of the major cystic fibrosis mutation (delta F508) in european populations. Nat. Genet. 7:169-175.

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.

Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. 2009. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324:387-389.

Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andres AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Boyko A et al. 2009. Darwinian and demographic forces affecting human protein coding genes. Genome Res. 19:838-849.

Penna-Martinez M, Ramos-Lopez E, Robbers I, Kahles H, Hahner S, Willenberg H, Reisch N, Seidl C, Segni M, Badenhoop K. 2009. The rs1990760 polymorphism within the IFIH1 locus is not associated with graves' disease, hashimoto's thyroiditis and addison's disease. BMC Med. Genet. 10:126.

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. Genome Res. 19:826-837.

Plagnol V, Wall JD. 2006. Possible ancestral structure in human populations. PLoS Genet. 2:e105.

Ray N, Wegmann D, Fagundes NJ, Wang S, Ruiz-Linares A, Excoffier L. 2010. A statistical evaluation of models for the initial settlement of the american continent emphasizes the importance of gene flow with asia. Mol. Biol. Evol. 27:337-345.

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.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R et al. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449:913-918.

Saillard J, Forster P, Lynnerup N, Bandelt HJ, Norby S. 2000. mtDNA variation among greenland eskimos: The edge of the beringian expansion. Am. J. Hum. Genet. 67:718-726.

Satta Y, Takahata N. 2004. The distribution of the ancestral haplotype in finite stepping-stone models with population expansion. Mol. Ecol. 13:877-886.

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.

Shigemoto T, Kageyama M, Hirai R, Zheng J, Yoneyama M, Fujita T. 2009. Identification of loss of function mutations in human genes encoding RIG-I and MDA5: Implications for resistance to type I diabetes. J. Biol. Chem. 284:13348-13354.

Sironi M, Clerici M. 2010. The hygiene hypothesis: An evolutionary perspective. Microbes Infect., in press.

Smyth DJ, Cooper JD, Bailey R, Field S, Burren O, Smink LJ, Guja C, Ionescu-Tirgoviste C, Widmer B, Dunger DB et al. 2006. A genome-wide association study of nonsynonymous SNPs

identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. Nat. Genet. 38:617-619.

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.

Sutherland A, Davies J, Owen CJ, Vaikkakara S, Walker C, Cheetham TD, James RA, Perros P, Donaldson PT, Cordell HJ et al. 2007. Genomic polymorphism at the interferon-induced helicase (IFIH1) locus contributes to graves' disease susceptibility. J. Clin. Endocrinol. Metab. 92:3338-3341.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595.

Takahata N. 1990. A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. Proc. Natl. Acad. Sci. U. S. A. 87:2419-2423.

Takeuchi O, Akira S. 2008. MDA5/RIG-I and virus recognition. Curr. Opin. Immunol. 20:17-22.

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.

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

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. 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.

Wall JD. 2000. Detecting ancient admixture in humans using sequence polymorphism data. Genetics 154:1271-1279.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7:256-276.

Webster MT, Smith NG, Hultin-Rosenberg L, Arndt PF, Ellegren H. 2005. Male-driven biased gene conversion governs the evolution of base composition in human alu repeats. Mol. Biol. Evol. 22:1468-1474.

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.

Figure legends

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

Figure 2. Worldwide allele frequency distribution for rs10439256. Each pie represents one HGDP-CEPH population. The ancestral T allele is shown in blue, the derived C allele in yellow. Yakut and Maya are circled in red.

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

Figure 4. Estimated haplotype tree for the *IFIH1* gene region we resequenced. Mutations are represented as black dots and named for their physical position along the regions. The absolute frequency of each haplotype is also reported. Note that mutation numbering does not correspond to that reported in figure 3.

Figure 5. Long-range haplotype analysis for populations living in the Americas.

Sliding window analysis of iHS (-log10 value, blue) and XP-EHH (-log10 value, red) for a 500 kb region centered around *IFIH1*. The location of known genes in the region is shown, as well as the *IFIH1* region we analyzed (gray shading).

Tables

Table 1. Nucleotide diversity and F_{ST} for the *IFIH1* region we analyzed

Pop.	N[a]	S[b]	θ_W[c] (rank[d])	π[c] (rank[d])	F_{ST} (rank[d])
YRI	40	49	10.62 (0.69)	13.86 (0.93)	CEU: 0.41 (0.97); AS: 0.20 (0.72); SAM: 0.52
CEU	40	6	1.30 (<0.001)	0.74 (0.005)	AS: 0.16 (0.78); SAM: 0.88
AS	40	31	6.72 (0.68)	6.93 (0.69)	SAM: 0.72
SAM	32	31	5.34	7.09	

[a] Sample size (chromosomes)
[b] Number of segregating sites
[c] Nucleotide diversity measures (x 10^{-4})
[d] percentile rank relative to the distribution of 5kb windows from NIEHS genes

Table 2. Summary statistics and MLHKA test for the *IFIH1* gene region.

Pop.	Tajima's D			Fu & Li's D*			Fu & Li's F*			Fay and Wu's H		MLHKA	
	value	rank[a]	*p*[b]	value	rank[a]	*p*[b]	value	rank[a]	*p*[b]	value	*p*[b]	k[d]	p
YRI	1.16	0.975	0.014	-0.73	0.350	0.413	-0.081	0.621	0.248	0.73	0.284	3	0.006
CEU	-1.14	0.145	0.104	-2.10	0.067	0.057	-2.11	0.062	0.055	0.23	0.392	0.40	0.046
AS	0.11	0.548	0.498	1.32	0.962	0.029	1.07	0.849	0.127	-12.45	0.024	2.6	0.013
SAM	-0.91	n.a.[c]	0.148	1.53	n.a.[c]	0.010	0.86	n.a.[c]	0.187	-22.13	0.003	n.a.[c]	n.a.[c]

[a] percentile rank relative to the distribution of 5kb windows from NIEHS genes;
[b] *p* value obtained by coalescent simulations using demographic models.
[c] not available
[d] selection parameter (k>1 indicates an excess of polymorphism compared to divergence; k< 1 indicates the opposite situation)

Table 3. CLR test results

Pop.	Test 1[a]		Test 2[b]	
	LR[c]	GOF[d]	LR[c]	GOF[d]
CEU	6.95 (0.004)	10.19 (0.25)	7.57 (0.002)	12.83 (0.24)
AS	6.01 (0.007)	15.61 (0.81)	9.25 (0.007)	14.83 (0.77)
SAM	9.90 (0.006)	9.18 (0.87)	6.17 (0.011)	11.21 (0.84)

[a] Distinguishing ancestral/derived allele.
[b] Not distinguishing ancestral/derived allele.
[c] Likelihood ratio with *p* values in parenthesis.
[d] Goodness-of-fit test with *p* values in parenthesis.

Figures

Figure 1.

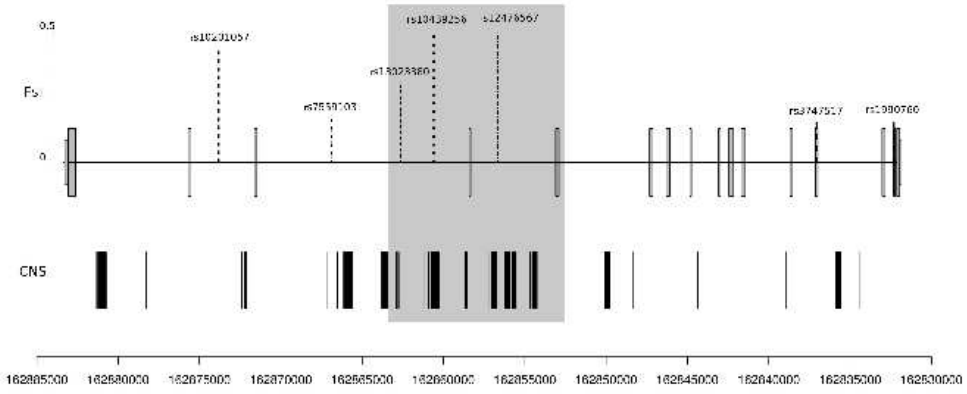


Figure 2.

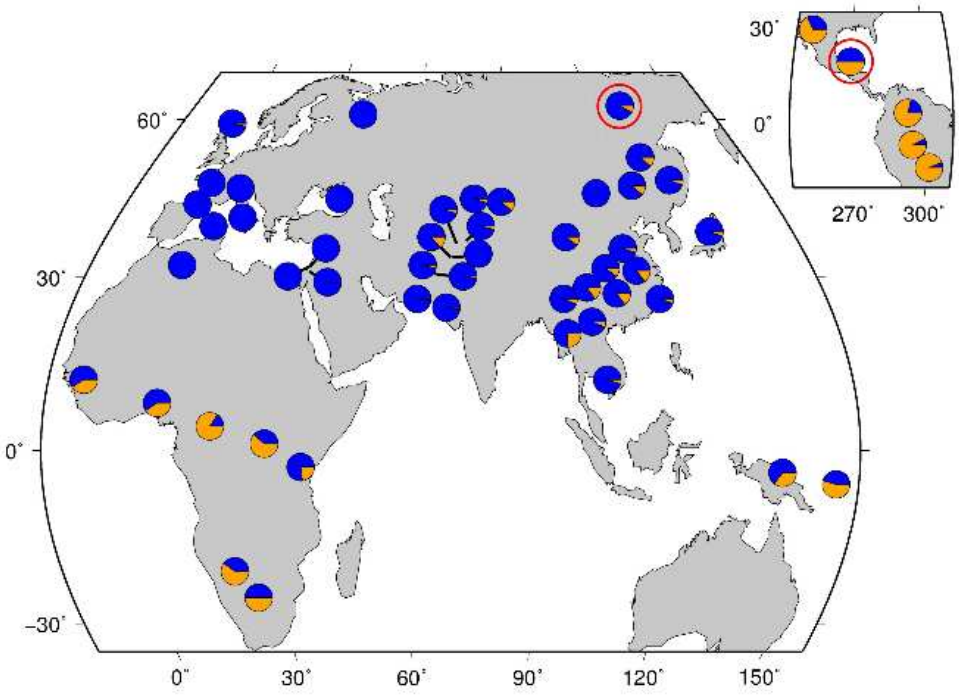


Figure 3.

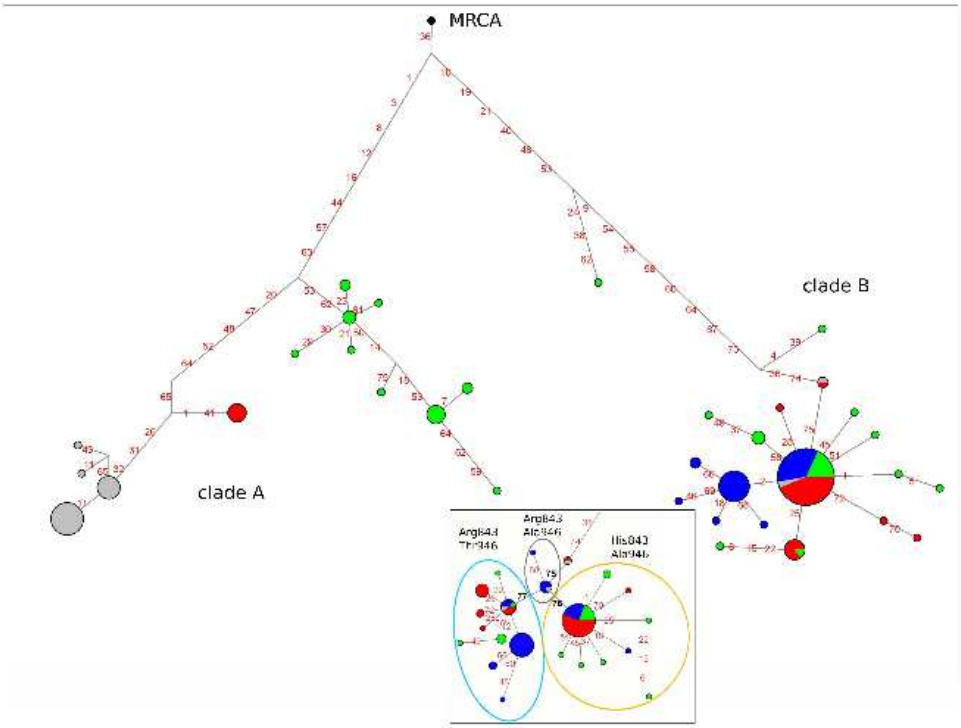


Figure 4.

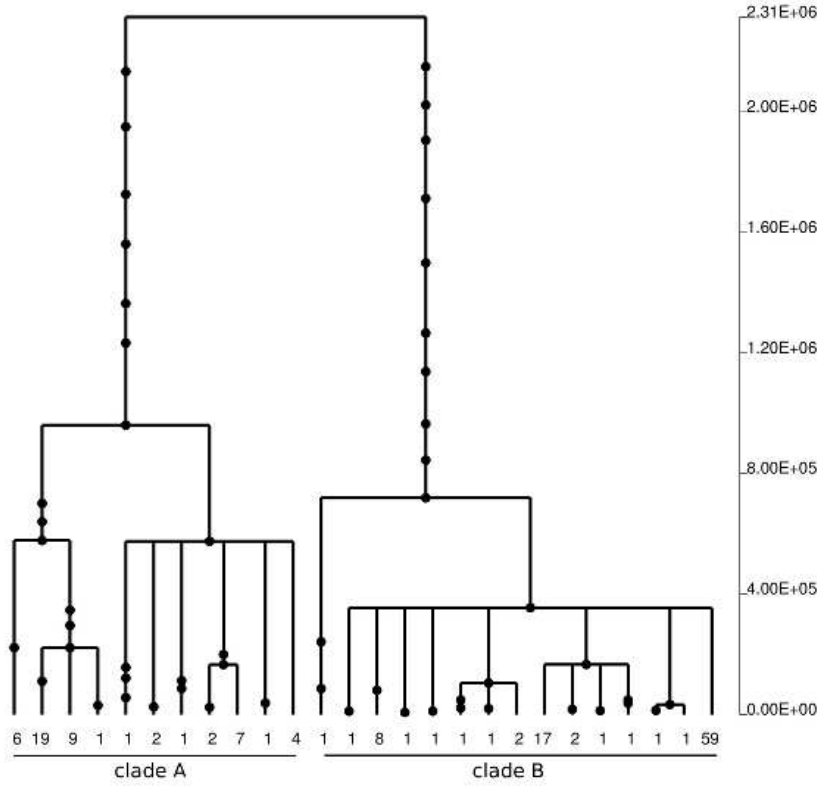
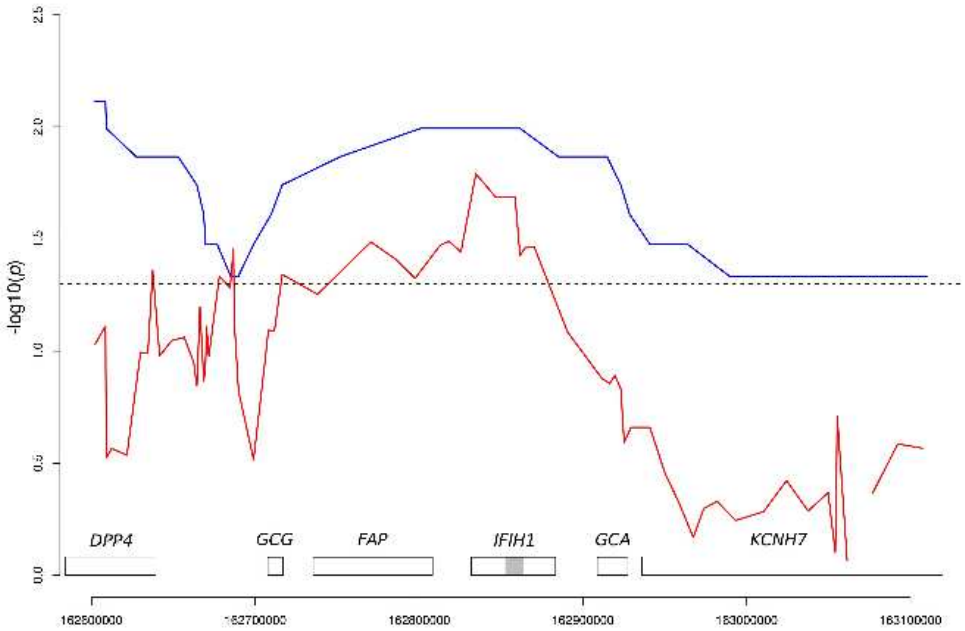


Figure 5.



37

2.6 The landscape of human genes involved in the immune response to parasitic worms

Fumagalli et al. BMC Evolutionary Biology 2010, 10:264
http://www.biomedcentral.com/1471-2148/10/264

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

The landscape of human genes involved in the immune response to parasitic worms

Matteo Fumagalli[1,2†], Uberto Pozzoli[1†], Rachele Cagliani[1], Giacomo P Comi[3], Nereo Bresolin[1,3], Mario Clerici[4,5], Manuela Sironi[1*]

Abstract

Background: More than 2 billion individuals worldwide suffer from helminth infections. The highest parasite burdens occur in children and helminth infection during pregnancy is a risk factor for preterm delivery and reduced birth weight. Therefore, helminth infections can be regarded as a strong selective pressure.

Results: Here we propose that candidate susceptibility genes for parasitic worm infections can be identified by searching for SNPs that display a strong correlation with the diversity of helminth species/genera transmitted in different geographic areas. By a genome-wide search we identified 3478 variants that correlate with helminth diversity. These SNPs map to 810 distinct human genes including loci involved in regulatory T cell function and in macrophage activation, as well as leukocyte integrins and co-inhibitory molecules. Analysis of functional relationships among these genes identified complex interaction networks centred around Th2 cytokines. Finally, several genes carrying candidate targets for helminth-driven selective pressure also harbour susceptibility alleles for asthma/allergy or are involved in airway hyper-responsiveness, therefore expanding the known parallelism between these conditions and parasitic infections.

Conclusions: Our data provide a landscape of human genes that modulate susceptibility to helminths and indicate parasitic worms as one of the major selective forces in humans.

Background

Helminth infections are estimated to infect about 2 billion individuals worldwide (reviewed in [1]). Although rarely fatal, these parasites cause high rates of morbidity by establishing chronic infections. In particular, the highest parasite burdens are observed in pre-school and school-aged children, often resulting in anemia, undernourishment and growth stunting (reviewed in [1]). During pregnancy helminth infection is a risk factor for preterm delivery, reduced birth weight and maternal mortality (reviewed in [1]). Moreover, by chronically infecting their host, parasitic worms increase the susceptibility to other pathogens such as viruses, bacteria and protozoa [1]. Previous works have indicated that the intensity of helminth infection is a heritable trait, although measures of heritability vary among studies and depend on the parasite beanalyzed [2].

These observations suggest that helminth infections have represented a very strong selective pressure for humans, a selective pressure that is very likely to also be remarkably constant over time. Indeed, most vertebrates have been hosting a wide range of parasitic worms for million years and humans have had their share before emerging as a species (reviewed in [3]). We have previously addressed the role of helminths as selective agents in human evolution analyzing a large set of human genes encoding interleukins and their receptors; we demonstrated that the pressure imposed by parasitic worms on these genes has been stronger than the one due to viral and microbial agents [4]. The reasons for this observation likely lie in the long-term relationship between humans and helminths, in the relatively slow evolutionary rates of these parasites and in their geographic distribution being considerably stable. Here we aimed at exploiting the selection signatures left by these pathogens on human genes to identify, at the genome-

* Correspondence: manuela.sironi@bp.lnf.it
† Contributed equally
[1]Scientific Institute IRCCS E. Medea, Bioinformatic Lab, Via don L. Monza 20, 23842 Bosisio Parini (LC), Italy
Full list of author information is available at the end of the article

wide level, candidate genes and variants that may have been subjected to helminth-driven selective pressure.

Results

Helminth diversity and prevalence correlate across geographic locations

In order to search for candidate variants subjected to helminth-driven selection, an estimate of the selective pressure exerted by these pathogens needs to be defined. It has been previously suggested [4-6] that the selective pressure exerted by infectious diseases in different geographic areas can be estimated by counting the number of different pathogen species/genera that are transmitted in these regions. In particular, in the case of helminthiases, we consider parasite diversity to be a better estimate of helminth-driven selective pressure than prevalence for different reasons. First of all, comprehensive data on prevalence are impossible to retrieve for many parasite species/genera or countries and even when available, prevalence data may vary considerably within the same country depending on the surveyed regions and the population surveyed (e.g. city dwellers rather than farmers/bushmen/nomads, children rather than adults). Moreover, the prevalence of specific helminthiases might have changed greatly over recent years as a result of eradication campaigns, and historical prevalence data are rarely available. Also, it should be considered that prevalence data are difficult to combine since in endemic regions polyparasitism is common [1]; indeed, in these regions subjects infected with multiple helminths tend to harbor the most intense infections, possibly due to an additive and/or multiplicative impact on nutrition and organ pathology (reviewed in [7]).

These same observations also apply to other possible measures such as parasite burden or infection pathogenicity, which are very difficult to quantify and that may have varied considerably along human evolutionary history. Conversely, the diversity of helminth species transmitted in different geographic locations has been shown to depend upon climatic variables [8] which, in turn, may be considered as relatively stable over time, suggesting that diversity may better describe long-term evolutionary pressures. Thus, we calculated helminth diversity from the Global Infectious Disease and Epidemiology Network database (Gideon); as described in the method section, all entries for single helminthiases were manually inspected and all species/genera transmitted in a given location were counted as present irrespective of their prevalence. The number of different helminth species/genera per country is reported in Additional file 1 (Table S1).

We next wished to verify whether the prevalence of the most common helminth infections (as reported in [1]) correlates with helminth diversity across the 52 populations genotyped in the HGDP-CEPH panel (Additional file 1, Table S1). Again, we retrieved prevalence and diversity data from Gideon by manually inspecting single and retrieving all prevalence surveys. The prevalence of single helminthiases per country was obtained by averaging all surveys; data obtained on HIV-seropositive individuals were not included in the analyses because of the known modulation of HIV on the susceptibility on helminth-based infections [1,9,10]. Finally, countries with no survey data were not included in the analysis for the specific helminth. Parasites were then grouped into five major classes (following [1]). As shown in table 1, the prevalence of all parasite groups correlated with helminth diversity.

SNPs associated with helminth diversity are over-represented in immune response genes

Previous analysis have indicated that several genes coding for interleukins and interleukin receptors are subjected to helminth-driven selective pressure [4]. Interleukins are central mediators of immunity and inflammation and, in general, we might expect genes with a role in immune response to be preferential targets of helminth-driven selective pressure. Specifically, we expect variants within these genes to be more frequently associated with helminth diversity than observed for randomly sampled loci. We verified this prediction by analyzing the ImmPort list which contains 2,287 genes involved in immune response and covered in the HGDP-CEPH panel, this latter containing data for more than 660,000 single nucleotide polymorphisms genotyped in almost 950 individuals sampled throughout the world (Additional file 1, Table S1) [11].

We calculated Kendall's τ rank correlation coefficient between allele frequencies of SNPs in ImmPort genes and helminth diversity. A SNP was defined as being significantly associated with helminth-diversity if it displayed a significant correlation (p value after Bonferroni correction < 0.01; uncorrected p value < 1.94×10^{-7}) and a τ value higher than the 95th percentile in the distribution of correlation coefficients calculated over all SNPs having minor allele frequency (MAF) similar (in the 1% range) to that of the SNP being analyzed. This latter requirement stems from the need to account for the influence of non-selective events, as a few HGDP-CEPH population result from recent or ancient admixture [11] and population history, migratory events and genetic drift have affected human genetic variability [12,13]. Among 2,287 ImmPort genes, 246 contained at least one SNP significantly associated with helminth diversity (Additional file 2, Table S2). The likelihood to obtain an equal or higher number of genes carrying significantly associated SNPs was assessed by a re-sampling approach. Specifically,

Table 1 Correlation between the prevalence of all parasite groups and helminth diversity

Parasite group	Kendall's τ	p value	Parasite species
Soil-transmitted nematodes	0.381	0.00037	*Ascaris lumbricoides, Trichuris trichiura, Necator americanus, Ancylostoma duodenale*
Filarial nematodes	0.387	0.00053	*Wuchereria bancrofti, Brugia malayi, Onchocerca volvulus, Loa loa*[a]
Schistosomes	0.482	2.8×10^{-5}	*Schistosoma mansoni, Schistosoma haematobium, Schistosoma intercalatum, Schistosoma japonicum, Schistosoma mekongi.*
Food-borne trematodes	0.617	2.6×10^{-7}	*Clonarchis sinensis, Opisthorchis viverrini, Paragonimus africanus, Paragonimus compactus, Paragonimus ecuadoriensis, Paragonimus hueltungensis, Paragonimus heterotremus, Paragonimus kellicotti, Paragonimus mexicanus, Paragonimus miyazakii, Paragonimus szechuanensis, Paragonimus tuanshanensis, Paragonimus uterobilateralis, Paragonimus westermani, Fasciolopsis buski, Fascicola (hepatica or gigantica)*
Taenia[b]	0.462	0.00633	*Taenia solium*

a *Dracunculus medinenis* was not included because prevalence data were only available in few countries;
b Prevalence data were available for 27 populations only.

we divided all genes with at least one SNP typed in the HGDP-CEPH panel in 24 intervals based on the number of typed SNPs; 10,000 re-samplings were then performed by selecting for each ImmPort gene, a randomly selected gene with a similar number of SNPs (i.e. a gene located in the same interval) (see methods). The number of SNPs in ImmPort genes did not differ significantly from the average number in the re-samplings (p = 0.22) and the empirical probability of obtaining 246 genes with at least one significant SNP resulted equal to 0.041, indicating that immune response genes more frequently display variants correlating with helminth diversity compared to randomly chosen loci.

When the same analysis was performed using the prevalence of soil-transmitted nematodes, filarial nematodes, schistosomes and food-borne trematodes no significant enrichment for ImmPort genes was observed (empirical p values = 0.83, 0.96, 0.45, and 0.39, respectively).

In agreement with previous suggestions [4], these data indicate that helminth diversity may be a reliable estimator helminth-driven selective pressure.

Genome-wide search for variants subjected to helminth-driven selective pressure

Given these results, we wished to identify SNPs significantly associated with helminth diversity on a genome-wide base. We therefore calculated Kendall's rank correlations between allele frequency and helminth diversity for all SNPs (n = 660,832) typed in the HGDP-CEPH panel. We next searched for instances which withstood Bonferroni correction with $\alpha = 0.05$ (i.e. uncorrected p value $< 7.6 \times 10^{-8}$) and displayed a τ percentile rank higher than the 95th among MAF-matched SNPs. A total of 3,478 SNPs mapping to 810 distinct genes satisfied both requirements (Additional file 3, Table S3). We next verified whether climatic variables could be responsible for the correlations detected between these SNPs and helminth diversity. Hence, for all countries where HGDP-CEPH populations are located we obtained the following parameters: average annual minimum and maximum temperature, short wave radiation flux and precipitation rate (annual maximum and mean). None of these SNPs withstood Bonferroni corrections in these analyses.

Previous works have reported an enrichment of selection signatures within or in close proximity to human genes [12,14-17]. In line with these data we verified that helminth-associated SNPs are more frequently located within gene regions compared to a control set of MAF-matched variants (χ^2 test, $p = 0.014$).

A full list of the 3,478 SNPs that showed a significant correlation with helminth diversity is available as Additional file 3 (Table S3). Table 2 shows the 20 strongest correlations together with a short comment on the possible role of candidate genes in immune response or helminth resistance.

Among the 810 genes subjected to helminth-driven selective pressure, we identified ectodysplasin A (*EDA*) and its receptor (*EDAR*). *EDAR* has been subjected to strong positive selection in populations of Asian descent [17] and is responsible for hair thickness in these populations [18]. One coding variant (rs3827760, 370Val/Ala) in *EDAR* is thought to be the selection target but it has not been genotyped in the HGDP-CEPH panel. We therefore wished to verify whether the variants we found to correlate with helminth diversity are located on the same haplotype as the selected 370Ala allele. We used the Sweep software [19] to analyze the haplotype structure in the genomic region encompassing *EDAR*; we selected a core containing the 370Ala/Val variant and used HapMap data from Asian individuals (Chinese and Japanese). The results indicated that most chromosomes carrying the putatively selected 370Ala variant also harbor four alleles we found to be significantly correlated with helminth diversity (Figure 1).

Table 2 Top 20 SNP (or SNP clusters) associated with helminth diversity

SNP	Candidate gene	Distance (bp)[a]	Annotation	τ	Description	Reference
rs6989916	*CSMD1*	14833	intergenic	0.702	CSMD1 acts as a regulator of the complement system	[77]
rs11614925; rs2082529	*NAP1L1*, *PHLDA1*	35490; 1923	intron; intergenic	0.700; 0.700	PHLDA1 participates in regulating T-cell receptor/CD3-dependent induction of CD95/Fas	[78]
rs1369977; rs1369976	*PDHA2*	45648; 44682	intergenic	0.696; 0.687	Pyruvate dehydrogenase (lipoamide) alpha 2	
rs4684083; rs9681213; rs1516320	*CHL1*	49940; 175000; 162760	intergenic	0.690; 0.676; 0.675	Cell adhesion molecule with homology to L1CAM	
rs10014145	*SLC39A8*	genic	intron	0.685	*SLC39A8* encodes a zinc transporter which is up-regulated by different cytokines in lung epithelia and monocytes	[79,80]
rs4682429	*CD200R1L*	12582	intergenic	0.684	Engagement of CD200RL1 (aka CD200R2) results in the development of dendritic cells that preferentially induce Treg Cells	[33]
rs7130880	*PRMT3*	14431	intergenic	0.682	*PRMT3* encodes a protein arginine methyltransferase expressed in T and B cells; arginine methylation is important for T cell activation and is induced by CD28 engagements	[81]
rs504508	*KATNAL1*	111767	intergenic	0.682	Katanin p60 subunit A-like 1	
rs12371626	*GRIP1*	genic	intron	0.680	GRIP1 acts with Beta-catenin to enhance the activity of LEF1 (lymphoid enhancer-binding factor 1)	[82]
rs1441443	*PDZRN3*	252500	intergenic	0.678	PDZ domain containing ring finger 3	
rs236233	*IRAK1BP1*	452153	intergenic	0.677	IRAKBP1 is required for TNF-alpha activation of NF-kB dependent-gene expression	[83]
rs3807250	*DPP6*	genic	intron	0.677	Dipeptidyl-peptidase 6; a susceptibility gene for amyotrophic lateral sclerosis	[84]
rs11702528	*BTG3*	52819	intergenic	0.676	*BTG3* is a transcriptional target of p53 that inhibits E2F1	
rs985122	*AJ606331*	508	intergenic	0.674	Putative non-coding RNA	
rs4692241	*STIM2*	529000	intergenic, eQTL	0.673	STIM2 promotes store-operated $Ca2^+$ entry into T cells; STIM1 and STIM2 proteins are required for the development and function of regulatory T cells	[34]
rs10270302	*GIMAP7*	genic	intron	0.672	*GIMAP7* encodes a GTPase of the immunity-associated protein family; it is expressed at very high levels in $CD4^+$ and $CD8^+$ T cells, and in NK cells	SymAtlas
rs7258075	*RYR1*	genic	intron	0.672	Activation of RYR1 causes a rapid increase in the expression of MHCII molecules on the surface of dendritic cells	[57]
rs424138	*DPYSL2*	genic	intron	0.671	Dihydropyrimidinase-like 2	
rs9952350	*ZFP161*	genic	intron, eQTL	0.671	*ZFP161* encodes a transcriptional repressor expressed at maximum levels in $CD4^+$ T cells	[85], SymAtlas
rs1143683	*ITGAM*	genic	missense (Ala858Val)	0.671	ITGAM combines with ITGB2 to form a leukocyte-specific integrin. ITGAM is the target of an immunomodulatory molecue secreted by *Ancylostoma caninum*	[41]

SNP are ranked according to τ values.

a distance of the associated SNP from the candidate gene. In all cases the candidate gene was the known gene closest to the SNP.

Functional characterization of genes subjected to helminth-driven selective pressure

We investigated the role and functional relationship among helminth diversity-associated genes using the Ingenuity Pathway Analysis and the PANTHER classification system [20-22]. For these analyses, a SNP was ascribed to a given gene if it was located within the transcribed region or in the 25 kb upstream of the transcription start site.

Unsupervised IPA analysis retrieved five high-scoring networks ($p < 10^{-12}$) (Figure 2 and Additional file 4, Figure S1) and two additional networks with lower scores ($p < 10^{-9}$). The two highest scoring networks were merged, as well as networks 3 and 5. As shown in figure 2, 37 and 32 genes in merged networks 1 and 2 correlated with helminth diversity, respectively, corresponding to almost 60% of network nodes.

We next investigated the over-representation of PANTHER classification categories among genes significantly associated with helminth diversity. Table 3 shows the 5 significantly over-represented PANTHER pathways with the contributing genes, as well as the most significantly over-represented molecular functions and biological processes. Notably, genes involved in cytokine-mediated inflammation and integrin signaling accounted for two significantly over-represented pathways. While the multiple functions of integrins and cytokines on the immune system are established, the role of glutamate and α adrenergic receptor signaling in immune related processes are less understood. Recent evidence has indicated that

Fumagalli *et al. BMC Evolutionary Biology* 2010, **10**:264
http://www.biomedcentral.com/1471-2148/10/264

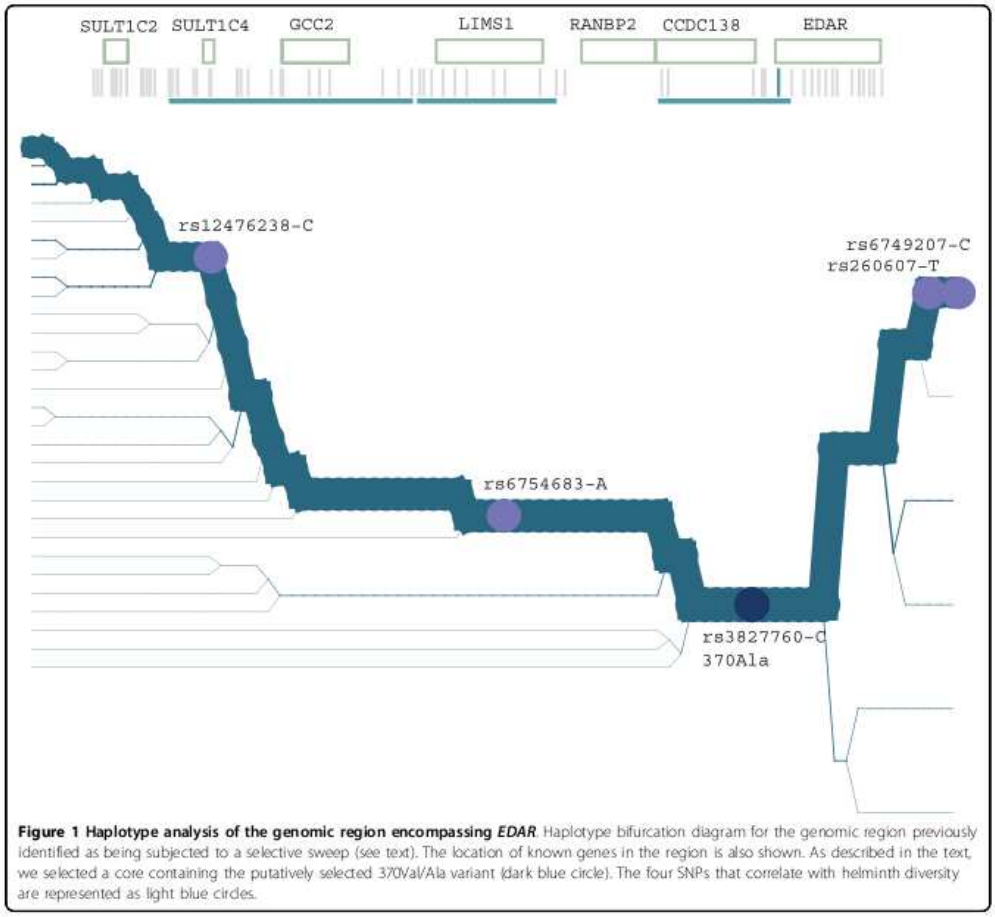


Figure 1 Haplotype analysis of the genomic region encompassing *EDAR*. Haplotype bifurcation diagram for the genomic region previously identified as being subjected to a selective sweep (see text). The location of known genes in the region is also shown. As described in the text, we selected a core containing the putatively selected 370Val/Ala variant (dark blue circle). The four SNPs that correlate with helminth diversity are represented as light blue circles.

ionotropic glutamate receptors are expressed by T lymphocytes (reviewed in [23]) and glutamate can exert different actions on these cells including triggering adhesion, proliferation and chemotaxis (reviewed in [23]). Similarly, α adrenergic receptors are expressed by different immunocompetent cells (reviewed in [24]); in particular, expression of *ADRA1A* is induced in human monocytes by inflammatory cytokines [25]. Finally, it is interesting to notice that components of the treleasing hormone (TRH) receptor signaling pathway are also over-represented among genes subjected to helminth driven selective pressure. Tacts on specific receptors within the pituitary gland to stimulate the release of thyroid stimulating hormone and prolactin. The immunomodulatory role of thyroid hormones on immune functions are clearlyestablished (reviewed in [26]). Moreover, experiments in mice have indicated that thyroxine plays a role in the establishment of *Schistosoma mansoni*infection [27,28] and animals treated with the hormone display increased parasite numbers and development of giant worms [28]. Notably, among variants correlating with helminth diversity at the genome-wide level, we also found one SNP relatively close to the gene encoding prolactin (rs13198653, $\tau = 0.64$, $p = 4.4 \times 10^{-9}$) and one variant within *PRLR* (prolactin receptor, rs4235652, $\tau = 0.62$, $p = 1.8 \times 10^{-9}$) (Additional file 3, Table S3). PRL was shown to be an immunomodulator and acts as a cytokine on many different immune cells; indeed, *PRLR* is expressed by B, T and NK cells, as well as macrophages (reviewed in [26]) and PRL expression in T cells is regulated by IL-2, IL-4 and IL-1β. These data

Fumagalli *et al. BMC Evolutionary Biology* 2010, **10**:264
http://www.biomedcentral.com/1471-2148/10/264

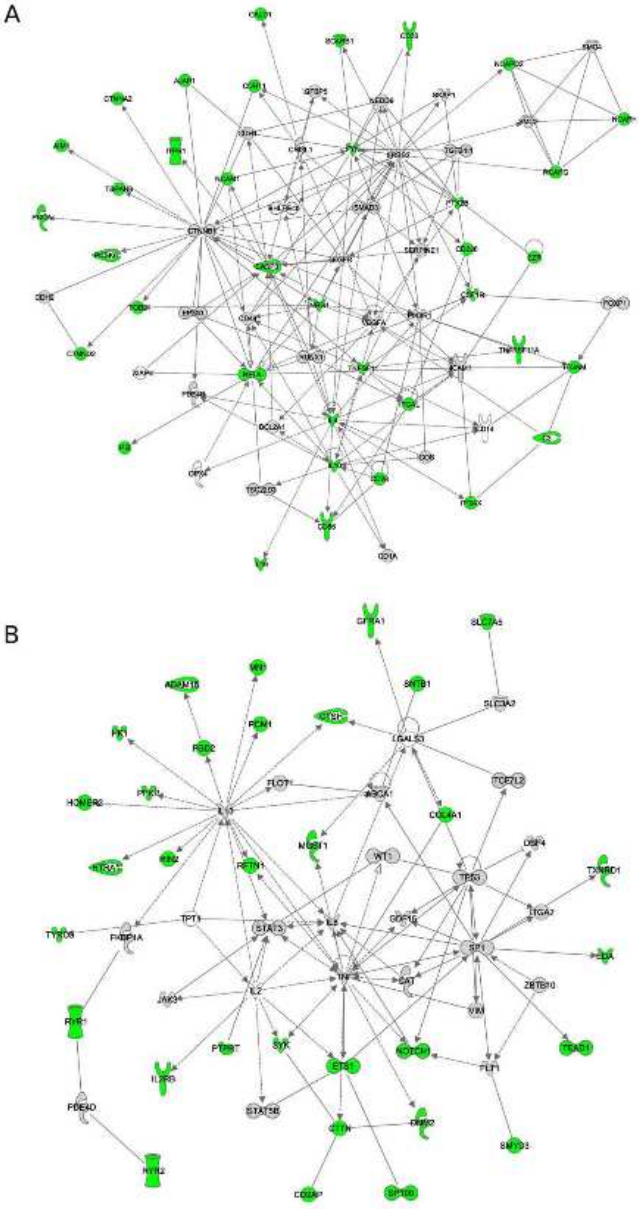


Figure 2 Network analysis of genes associated with helminth diversity. Genes are represented as nodes, edges indicate known interactions between proteins (solid lines depicts direct and dashed lines depict indirect interaction). Genes are color coded as follows: green, genes with at least one SNP significantly associated with helminth diversity; gray, genes covered by at least one SNP in the HGDP-CEPH panel; white, genes with no SNPs in the panel. Panel A and B represents the networks obtained by merging networks 1 with 2 and 2 with 5, respectively.

Fumagalli et al. BMC Evolutionary Biology 2010, 10:264
http://www.biomedcentral.com/1471-2148/10/264

Table 3 Panther over-represented categories among genes showing correlation with helminth diversity

PANTHER category	PANTHER description	Number of genes	p value[a]	Contributing genes
Pathway	Integrin signalling pathway	26	0.0067	*ITGAV, COL4A1, ITGA8, FLNB, ITGA9, ITGB8, ITGAM, ITGBL1, COL4A2, COL1A2, DOCK2, PTK2B, ITGAX, FYN, MAPK13, LAMA2, ITGAL, LIMS1, COL24A1, ELMO1, COL15A1, DOCK1, GRB2, ITGB7, MAPK14, COL9A3*
	Alpha adrenergic receptor signaling pathway	8	0.0138	*PLCB1, ADRA1A, ITPR1, GNAQ, SNAP25, PLCE1, VAMP3, ITPR2*
	Inflammation mediated by chemokine and cytokine signaling pathway	31	0.0206	*SOCS6, PLCH2, PLCB1, GRB2, GNG10, ITPR1, CXCR6, MYH14, ITGA9, MYO3B, PLA2G4B, GNAQ, CAMK2A, PLCL1, CAMK2 D, ITGAM, CCR9, COL1A2, ITGB7, PAK7, VAV2, PLCD3, ADCY2, PTK2B, PLCE1, RELA, ITPR2, CCL20, PLA2G4A, ITGAL, COL23A1*
	Ionotropic glutamate receptor pathway	11	0.0263	*SLC17A8, GRIK2, GRIA1, CACNG5, GRIN3A, SHANK2, CAMK2A, SNAP25, CAMK2 D, VAMP3, CACNG8*
	Thyrotropin-releasing hormone receptor signaling pathway	11	0.0303	*PLCH2, PLCB1, CGA, CACNB2, GNG10, GNAQ, PLCD3, CHGA, PLCE1, VAMP3, SNAP25*
Biological process	Signal transduction	270	4.05×10^{-17}	n.r.
	Cell adhesion	73	7.18×10^{-11}	n.r.
	Cell communication	113	9.11×10^{-10}	n.r.
	Cell structure and motility	103	1.35×10^{-8}	n.r.
	Neuronal activities	63	2.25×10^{-8}	n.r.
	Developmental processes	163	2.87×10^{-8}	n.r.
	Ion transport	64	8.48×10^{-7}	n.r.
	Cation transport	54	1.61×10^{-6}	n.r.
Molecular function	Ion channel	44	4.06×10^{-7}	n.r.
	Receptor	117	2.79×10^{-6}	n.r.
	Hydrolase	66	1.92×10^{-5}	n.r.
	G-protein modulator	45	2.84×10^{-5}	n.r.
	Cell adhesion molecule	42	3.92×10^{-5}	n.r.
	Signaling molecule	68	6.44×10^{-5}	n.r.
	Membrane-bound signaling molecule	20	8.95×10^{-4}	n.r.

[a] *p* values are Bonferroni corrected
n.r.: not reported

therefore suggest a role for both thyroid hormones and prolactin in the resistance to parasitic worms.

Helminth-driven selection and susceptibility to allergy and asthma

We next wished to analyze the relationship between variants/genes associated with helminth diversity and the genetic susceptibility to asthma and allergy. We searched among published genome-wide association studies (GWAS) for SNPs that have been associated with allergy, asthma or related traits (serum IgE levels and plasma eosinophil count). Only 12 SNPs were retrieved, 9 of them genotyped in the HGDP-CEPH panel. One of these SNPs (rs12619285) displayed a significant correlation ($p = 5.8 \times 10^{-8}$) with helminth diversity (Table 4) and the allele associated with high eosinophil counts [29] positively correlated with the diversity of parasitic worms. In order to gain further insight into this issue, we focused on genes rather than variants, in line with the view that the gene rather than the allele should be regarded as the replication unit. Given the low consistency that often plagues association studies, only robust allergy/asthma susceptibility genes were considered (see methods). As shown in Table 4, we observed that 12 allergy/atopy genes displayed at least one SNP (either genic or intergenic) significantly associated, at the genome-wide level, with helminth diversity. Among these genes, one SNP in *IL4* that has been associated with asthma (rs2070874, +33C/T) has also been genotyped in the HGDP-CEPH panel; again, the allele that positively correlates with helminth diversity (T) is associated with asthma [30].

It is worth noting that we found several SNPs located upstream the transcription start site of *CLTA4* and subjected to helminth driven selective pressure. *CLTA4* is located on the long arm of chromosome 2, telomeric to *CD28*. This raises the possibility that the observed allele associations at these two genes (see Figure 3) derive

Table 4 SNPs that correlate with helminth diversity in asthma/allergy genes

SNP	Gene	Distance	τ	Reference
rs2243268 rs2070874	*IL4*	genic genic	0.61 0.60	[9]
rs17316177 rs4368333	*KCNS3*	11291 29279	0.63 0.59	OMIM
rs231735 rs231804 rs11571291	*CTLA4*	38632 23862 11376	0.63 0.60 0.58	[9]
rs4353658 rs7579207	*DPP10*	genic genic	0.62 0.58	[9]
rs1930713 rs2245960 rs7849955	*TLR4*	253946 277604 85563	0.62 0.61 0.59	[9]
rs708491	*PTGER2*	16071	0.58	[9]
rs10905349	*GATA3*	276916	0.57	[9]
rs7329078	*PHF11*	genic	0.57	[9], OMIM
rs1554286	*IL10*	genic	0.57	[9]
rs10237930	*NPSR1*	16676	0.56	[9]
rs877741	*ADRB2*	9418	0.56	[9], OMIM
rs12619285	*IKZF2*	40365	0.56	[29]

from linkage to a single selected allele. Yet, analysis of linkage disequilibrium (LD) (Additional file 5, Figure S2) indicates that LD is not extensive across the genomic region (also due to the presence of a recombination hot-spot in between the two genes), suggesting that *CD28* and the promoter region of *CTLA4* are independent selection targets.

Discussion

Here we propose that candidate susceptibility genes for parasitic worm infections can be identified by searching for SNPs that display a strong correlation with the diversity of helminth species/genera transmitted in different geographic areas. Our approach relies on the assumption that helminth-driven selective pressure has affected the spatial distribution of these variants and it may suffer from a few limits and caveats. First, we used the diversity of parasitic worms as a measure of helminth-driven selective pressure for the reasons reported above. Yet, helminth diversity may be affected by report biases (e.g. under-reporting in developing countries due to limited research and clinical facilities) and it weights equally all helminth species irrespective of their prevalence and disease burden. As an example, rare helminth species that are reported in Gideon and included in our diversity measure may have exerted a very limited selective force (although prevalence may have changed over time and it is difficult to infer selective pressure for single species). In order to evaluate the impact of rare species, we recalculated helminth diversity by taking into account only parasites that are common in at least one country (see methods and Additional file 1, Table S1). This diversity measure strongly correlated ($\tau = 0.75$, $p < 10^{-5}$) with parasite diversity calculated over all species, suggesting that the inclusion of rare helminths should not largely affect our results. Another possible confounding factor is accounted for by the co-variation of helminth diversity with other environmental variables (e.g. climate and other infectious agents). We verified that variants associated with helminth diversity are significantly enriched within immune response genes and that the SNPs we identified do not correlate with climatic factors, suggesting that climate does not act as an important confounding factor. Yet, we cannot rule out the possibility that other infectious agents have affected the spatial distribution of the variants we identified as the diversity of helminths correlates with that of other human pathogens [4,8] across geographic locations.

Finally, although we searched for SNPs strongly associated with helminth diversity (uncorrected p value < 7.6×10^{-8}) and we applied a correction based on the MAF-matched distribution of Kendall's rank correlation coefficients, we cannot exclude that the spatial distribution of a fraction of the variants we identified is due to population demography, migration history and drift, therefore representing false positives.

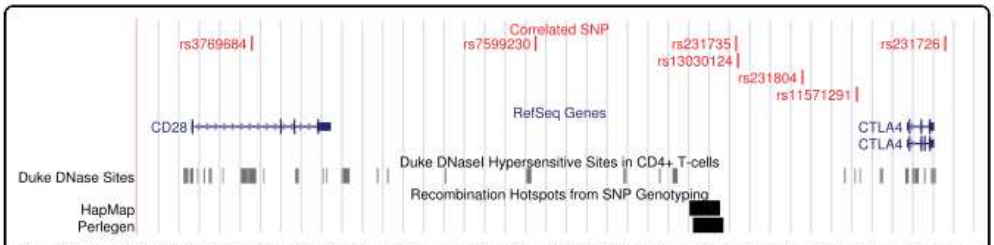


Figure 3 Analysis of SNP located in the genomic region encompassing *CD28* and *CTLA4*. SNPs significantly associated with helminth diversity are shown in red while the region covered by *CD28* and *CTLA4* are shown in blue. The location of DNAse hypersensitive sites in $CD4^+$ T cells is shown in gray while recombination hotspots are in black. The image was generated by using the "add custom track" utility available through the UCSC Genome Browser.

Fumagalli et al. BMC Evolutionary Biology 2010, 10:264
http://www.biomedcentral.com/1471-2148/10/264

Despite these limitations we were able to identify several genes that can be regarded as good candidates as modulators of susceptibility to helminth infection as testified by our interaction network and PANTHER analyses. Mammalian hosts respond to parasitic worms in a relatively uniform manner by producing specific cytokines (mainly IL-4, IL-10, IL-5 and IL-13) and IgE, as well as through the activation of effector cells such as eosinophils, basophils and mast cells [31]. Overall, the response to helminth infection is Th2-dominated and serves to both oppose the parasite and to contain tissue-damage. In line with this concept, *IL4* and *IL10*, as well as *IL13* are hub genes in two interaction networks we identified (Figure 2). Nonetheless, the role of other immune components during helminth infections is becoming increasingly clear. In addition to Th1 cells that mediate host response in some stages of *Schistosoma* and *Brugia malayi* infection (reviewed in [31,32]), the role of regulatory T cells (Treg) is now recognized (reviewed in [31,32]). An increase in Tregs has been observed in different experimental mouse models of helminthiases and in infected humans (reviewed in [31]). Notably, among the 20 genes more strongly associated with helminth diversity (Table 2) we identified two loci, namely *CD200R1L* [33] and *STIM2* [34] that are involved in the development and function of Treg cells. Additionally, we found several SNPs located upstream the transcription start site of *CTLA4* to strongly correlate with helminth diversity (Table 4). The gene encodes a co-inhibitory lymphocyte molecule that is preferentially expressed by Treg cells [35] and is thought to be at least partially responsible for the hypo-responsive phenotype of Th2 effector cells (referred to as "conditioned Th2") which is often observed in helminth infections (reviewed in [31]). As an example, blockade of CTLA-4 during *Nippostrongylus brasiliensis* infection results in higher Th2 cytokine production and decreased parasite numbers [36]. CTLA-4 competes with CD28 for binding to CD86 (and CD80) (reviewed in [37]). Both *CD28* and *CD86* carry variants significantly associated with helminth diversity (Figure 3 and Additional file 3, Table S3) and their binding provides a co-stimulatory signal for naive T cells; however, binding of CTLA-4 to CD86/CD80 on dendritic and T cells leads to functional inhibition (reviewed in [37]). Similarly to *CTLA4*, both *CD86* and *CD28* have been implicated in the immune response against helminths. Specifically, previous studies have indicated that anti-CD86 treatment blocks immune response to *Schistosoma mansoni* and *Heligmosomoides polygyrus* and *cd28-/-* mice display increased susceptibility to *S. mansoni* [38]. Interestingly, we also found *CD247*, another co-inhibitory molecule to correlate with helminth diversity (Additional file 3, Table S3). A recent report [39] has shown that *Schistosoma* induces anergy of $CD4^+$ and $CD8^+$ T cells by up-regulation of *CD247* expression on macrophages. These latter cells are considered an important component of anti-helminth response and an alternative form of macrophages has been described in subjects infected by parasitic worms. These cells up-regulate arginase instead of iNOS and express specific molecules including RETNLB (in humans the entry corresponding to *FIZZ1/retnla* has been discontinued and replaced with *RETNLB*) and CHIA (acidic chitinase or AMCase) (reviewed in [31,32]). Notably, we found one SNP in *RETNLB* to significantly correlate with helminth diversity (Additional file 3, Table S3); with respect to *CHIA*, we noticed that one variant (rs10494133) displayed a strong correlation with helminth diversity although it did not withstand Bonferroni correction at the genome-wide level (τ = 0.49, $p = 3.3 \times 10^{-6}$).

We found several integrins and adhesion molecules to correlate with helminth diversity (Table 3). Among these, *ITGAM*, *ITGAL* and *ITGAX* (Figure 2) encode integrin α chains that combine with the β 2 chain (encoded by *ITGB2*) to form leukocyte-specific heterodimeric integrins. These molecules regulate lymphocyte adhesion and transendothelial migration, playing therefore a central role in inflammatory processes. ITGAL and ITGAM are bound by neutrophil inhibitory factor (NIF), an antiadhesive glycoprotein isolated from the canine hookworm *Ancylostoma caninum* [40,41], suggesting that leukocyte integrins are relevant to the immune response to helminth infections. Interestingly, more recent evidence [42] has revealed that the genome of the human parasite *Necator americanus* encodes at least 9 genes with similarity to NIF, suggesting that their products might play a similar role in establishing an immunocompromised niche for the parasite. The ITGAM/ITGB2 and ITGAX/ITGB2 integrins bind iC3b (Figure 2), a cleavage product of complement component 3 (C3). Previous studies have shown that iC3b is deposited on *N. brasiliensis* larvae [43,44] and *c3* deficient mice carry high lung larval burdens [43]. In line with these findings, *C3* is essential for killing *Strongyloides stercoralis* larvae in mice [45] and *c3-/-* mice do not develop an effective Th2 response after infection with *S.mansoni* and cannot clear the parasite after chemotherapy [46].

As far as the second network is concerned (Figure 2), it is worth mentioning that *NOTCH1* and *ETS1* have been implicated in multiple immune functions. The NOTCH signaling pathway is involved in the intrathymic differentiation of T cells, as well as in Th cell development in the periphery [47]. In line with the role of Treg cells in helminth infection, *NOTCH1* has recently been shown to be involved in Treg function [48] and to cooperate with TGFb for regulation of the *FOXP3*

promoter [49]. With respect to *ETS1*, it functions as a transcriptional regulator of several cytokine genes including *IL5*, *IL2* and *GMCSF* [50-52]. It also regulates expression of *CD226* [53] and it is known to induce of Th1 mediated inflammation [54].

Also, network B (Figure 2) contains two genes coding for ryanodine receptors (*RYR1* and *RYR2*); we also found *RYR3* to correlate significantly with helminth diversity (Additional file 3, Table S3). The function of these molecules in the immune system is poorly understood yet, both *RYR1* and *RYR3* have been involved in calcium signaling in T cells [55,56]. Moreover, recent evidences have indicated that dendritic cells express *RYR1* and activation of the receptor causes a rapid increase in the expression of MHCII molecules on the surface of these cells [57].

Finally, it is interesting to notice that among the genes subjected to helminth-driven selective pressure in network B we found *SYK*, encoding a tyrosine kinase that interacts with the high affinity IgE receptor and mediates IgE signaling in mast cells and basophils [58,59]. Similarly, *CD226* and *FYN* (both in network A, Figure 2) have been involved in mast cell activation mediated by the high affinity IgE receptor [60], suggesting a role for these genes in allergic inflammation. Indeed, *syk* has been shown to mediate airway hyper-responsiveness in an experimental mouse model [61] and most genes discussed above have been involved in the elicitation of allergic phenomena. In addition genes reported in Table 4, the interaction of CD28 with CD86 is central to induction of allergic airway inflammation in mice [62] and CD86 antisense oligonucleotides suppress airway hyper-responsiveness in allergic animals [63]. Variants in C3 have been associated with asthma [64] and mice deficient in C3 exhibit diminished airway hyper-responsiveness and lung eosinophilia when challenged with allergen [65]; also, NOTCH1 is involved $CD8^+$ T cell-mediated development of airway hyper-responsiveness and inflammation [66], while ITGAL/ITGB2 mediates altered responsiveness of atopic asthmatic airway smooth muscle in rabbits [67]. Finally, Ets-1 induces tenascin expression in bronchial fibroblasts [68]. In this respect it is worth mentioning that, although subjects genotyped in the HGDP-CEPH panel are supposed to be healthy, a proportion of them may suffer from relatively mild diseases including asthma, atopy and related disorders; this may be especially true in some areas such as Latin America, for example, where urban centres have the highest reported prevalence of asthma worldwide (reviewed in [69]). While this possibility does not affect the results we reported herein, it highlights the fact that the epidemiology of these disorders is rapidly changing, and several reports have revealed a general increase in prevalence with urbanization, leading to the suggestion that environmental factors (including helminth infections) may play a central role in modulating the susceptibility to these diseases (reviewed in [69]). The relationship between asthma/allergy susceptibility and parasitic worms is though to be complex (reviewed in [70]). On one hand helminth-driven selective pressure is expected to favor individuals carrying alleles that allow a strong Th2 response and, therefore to promote the transmission and spread of asthma-susceptibility variants. On the other hand, lack of parasites in developed countries has likely removed the immunomodulatory role of these organisms, eventually leading to the increased incidence of atopic conditions. The current knowledge of asthma/allergy susceptibility alleles (12 alleles identified by GWAS) is too limited to warrant extensive speculation on the first issue. Still, our data indicate that many genes we identified carry variants associated with asthma/allergy or have been involved in the elicitation of airway hyper-responsiveness. Therefore, our results expand the previously noticed parallelism between genes involved in the development of asthma/allergy and those responsible for responding to parasitic worms, suggesting that the evolutionary scenario underlying the increase in asthma, allergy and related phenotype envisages a relevant role for these long-standing parasites.

Among the genes subjected to helminth-driven selective pressure we identified *EDAR* and *EDA*, its ligand. Binding of ectodysplasin to EDAR activates the NF-kB pathway through the NEMO protein. The EDA/EDAR pair mediates signals needed for the development of ectodermal appendages and mutations in both genes result in hypohidrotic ectodermal dysplasia. Many studies [15,17,71,72] have indicated that *EDAR* has been subjected to a strong selective pressure resulting in the rapid spread of the putatively selected 370Ala allele in Asian populations. This allele is responsible for the hair phenotype of these populations but the selective pressure underlying the selective sweep is unknown. Hypotheses have been proposed that increased hair thickness might be protective against cold climates or be favored through sexual selection [18,73]. We found that in Asian populations, most chromosomes carrying the selected allele also carry four SNPs subjected to helminth-driven selective pressure (Figure 1). Both *EDA* and *EDAR* are expressed in human lymphocytes and dendritic cells (see methods), suggesting that they may function as NF-kB activators in these cell types, as well. It is therefore tempting to speculate that helminths represent the selective pressure underlying the spread of a selected allele in Asia. This idea is consistent with the concept whereby infectious agents have represented one of the major selective forces for human populations.

Fumagalli et al. BMC Evolutionary Biology 2010, 10:264
http://www.biomedcentral.com/1471-2148/10/264

Conclusions

In summary, our data are consistent with the notion whereby parasitic worms have acted as a powerful selective force on human populations and have contributed to shape nucleotide variability at a number of genes involved in immune responses. We also show that several genes associated with helminth diversity are involved in the pathogenesis of atopic conditions or in airway hyper-responsiveness.

Methods

Data retrieval and statistical analysis

Helminth absence/presence matrices for the 21 countries where HGDP-CEPH populations are located were derived from the Gideon database. Information in Gideon is weekly updated and derives from World Health Organization reports, National Health Ministries, PubMed searches and epidemiology meetings. The Gideon Epidemiology module follows the status of known infectious diseases globally, as well as in individual countries, with specific notes indicating the disease's history, incidence and distribution per country. We manually curated helminth absence/presence matrices by extracting information from single Gideon entries. Following previous suggestions [4-6], we recorded only helminths that are transmitted in the 21 countries, meaning that cases of transmission due to tourism and immigration were not taken into account. A total of 60 helminth species were identified in at least one country (Additional file 6, Table S4). Prevalence data for single helminth infections were similarly obtained from Gideon, as described in the text. In order to calculate parasite diversity for species that are common in at least one country, we inspected Gideon entries for survey data or prevalence notes; helminth infections reported as "rare in humans" were discarded; similarly, parasites with no prevalence estimates or notes were considered as rare; therefore, this diversity measure should be regarded as an approximate estimate.

The annual minimum and maximum temperature were retrieved from the NCEP/NCAR database (Legates and Willmott Average, re-gridded dataset) using the geographic coordinates reported by HGDP-CEPH website for each population. Similarly, net short wave radiation flux data were obtained from NCEP/NCAR (Reanalysis 1: Surface Flux); these data were read using Grid Analysis and Display System (GrADS).

Since helminth diversity, due to data organization in Gideon, can only be calculated per country (rather than per population), the same procedure was applied to climatic variables. Therefore the values of annual temperature, radiation flux and precipitation rate were averaged for populations located in the same country. This assures that a similar number of ties is maintained in all correlation analyses.

Data concerning the HGDP-CEPH panel derive from a previous work [11]. Atypical or duplicated samples and pairs of close relatives were removed [74]. Following previous indications [4,5], Bantu individuals (South Africa) were considered as one population.

A SNP was ascribed to a specific gene if it was located within the transcribed region or no farther than 500 bp upstream the transcription start site. MAF for any single SNP was calculated as the average over all populations. The list of immune response genes was derived from the Immunology Database and Analysis Portal (ImmPort). Expression data were obtained from SymAtlas. SNPs identified in GWAS and associated with allergy, asthma or related traits (serum IgE levels and plasma eosinophil count) were derived from the A Catalog of Published Genome-Wide Association Studies. The list of allergy/asthma susceptibility genes was obtained from a previous review [9] or from the Online Medelian Inheritance in Man website (MIM: 600807).

All correlations were calculated by Kendall's rank correlation coefficient (τ), a non-parametric statistic used to measure the degree of correspondence between two rankings. The reason for using this test is that even in the presence of ties, the sampling distribution of τ satisfactorily converges to a normal distribution for values of n larger than 10 [75].

In order to estimate the probability of obtaining 246 genes carrying at least one significantly associated SNP out of a group of 2,287 genes (the number of ImmPort genes), we applied a re-sampling approach after dividing genes on the basis of the number of SNPs typed in the HGDP-CEPH panel. In particular, all genes covered by at least one SNP in the HGDP-CEPH panel (number of genes = 15,280) were divided in 24 intervals based on the distribution of typed SNPs per gene (Additional file 7, Table S5). Samples of 2,287 genes were randomly extracted from a list of all genes covered by at least one SNP in the HGDP-CEPH panel by applying the criterion that for each ImmPort gene, a control gene was selected from the same interval. For each sample the number of genes with at least one significant SNP were counted. The empirical probability of obtaining 246 genes was then calculated from the distribution of counts deriving from 10,000 random samples. Similarly, the number of SNPs in ImmPort genes was compared to the distribution of SNPs in the 10,000 re-samplings.

Analysis of PANTHER over-represented functional categories and pathways was performed using the "Compare Classifications of Lists" tool available at the PANTHER classification system website. Briefly, gene

lists are compared to the reference list using the binomial test [22] for each molecular function, biological process, or pathway term in PANTHER. All p values were Bonferroni corrected. All calculation were performed in the R environment [76]. For PANTHER analysis we widened the inclusion criteria in that SNPs located within the transcribed region or in the 25 kb upstream the transcription start site were ascribed to the gene.

eQTL data were derived from the eQTL Resource web site held at the University of Chicago (Prtichard Lab).

Network construction

Biological network analysis was performed with Ingenuity Pathways Analysis (IPA) software using an unsupervised analysis. IPA builds networks by querying the Ingenuity Pathways Knowledge Base for interactions between the identified genes and all other gene objects stored in the knowledge base; it then generates networks with a maximum network size of 35 genes/proteins. We used all genes showing at least one significantly associated SNP as the input set; in this case a SNP was ascribed to a gene if it was located within the transcribed region or in the 25 kb upstream. All network edges are supported by at least one published reference or from canonical information stored in the Ingenuity Pathways Knowledge Base. To determine the probability of the analyzed genes to be found together in a network from Ingenuity Pathways Knowledge Base due to random chance alone, IPA applies a Fisher's exact test. The network score represents the -log (p value).

LINKS

The Immunology Database and Analysis Portal, https://www.immport.org

Ingenuity Pathway Analysis, Ingenuity Systems, http://www.ingenuity.com

NCEP/NCAR, Surface flux, http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.surfaceflux.html

GrADS, http://www.iges.org/grads/

SymAtlas,: http://symatlas.gnf.org

Catalog of published GWAS, http://www.genome.gov

Panther,: http://www.pantherdb.org

eQTL resources @ the pritchard lab, http://eqtl.uchicago.edu/Home.html

UCSC Genome Browser, http://genome.ucsc.edu

HGDP-CEPH Panel, http://hagsc.org/hgdp/

Sweep software, http://www.broadinstitute.org/mpg/sweep/

Additional material

Additional file 1: Table S1. Populations in the HGDP-CEPH panel and helminth diversity estimates.

Additional file 2: Table S2. Genes in the ImmPort list that display at least one SNP significantly associated with helminth diversity. For each gene the SNP showing the strongest correlation is reported. SNPs are ranked according to the value of τ.

Additional file 3: Table S3. SNPs significantly associated with helminth diversity. The table reports all SNPs that withstood Bonferroni correction at the genome-wide level (with alpha = 0.5) and displayed a tau percentile rank higher than the 95th among MAF-matched SNPs, as described in the main text. SNPs are ranked according to the value of τ. If the SNP is located within a genic region (or in the 500 upstream nucleotides) the gene symbol is reported. Alternatively, the gene closest to the SNP and its distance (in bp) are indicated.

Additional file 4: Figure S1. In addition to the two merged networks in the main text, IPA identified three additional networks (A-C) with $p < 10^{-9}$. Genes are represented as nodes, edges indicate known interactions between proteins (sold lines depicts direct and dashed lines depict indirect interaction). Genes are color coded as follows: green, genes with at least one SNP significantly associated with helminth diversity; gray, genes covered by at least one SNP in the HGDP-CEPH panel; white, genes with no SNPs in the panel.

Additional file 5: Figure S2. Analysis of LD in the genomic region encompassing *CD28* and *CTLA4*. SNPs significantly associated with helminth diversity are shown in red while the region covered by *CD28* and *CTLA4* are shown in blue. The location of DNAse hypersensitive sites in $CD4^+$ T cells is shown in gray while recombination hot-spots are in black. LD plots ($r2$) are shown for Yoruba (YRI), Europeans (CEU) and Asians (JPT+CHB). The image was generated by using the "add custom track" utility available through the UCSC Genome Browser.

Additional file 6: Table S4. Helminth species/genera transmitted in at least one country and that are common in at least one country.

Additional file 7: Table S5. Gene subdivision on the basis of SNP number. Genes were divided in 24 intervals according to the number of SNPs typed in the HGDP-CEPH panel.

Abbreviations

SNP: single nucleotide polymorphism; Treg: regulatory T cell; LD: linkage disequilibrium; MAF: minor allele frequency; NIF: neutrophil inhibitory factor; TRH: releasing hormone; PRL: prolactin.

Acknowledgements

M.S. is part of the Doctorate School in Molecular Medicine, University of Milan.

This study was supported by grants from Istituto Superiore di Sanita' "Programma Nazionale di Ricerca sull' AIDS", the EMPRO and AVIP EC WP6 Projects, the nGIN EC WP7 Project, the Japan Health Science Foundation, 2008 Ricerca Finalizzata [Italian Ministry of Health], 2008 Ricerca Corrente [Italian Ministry of Health], Progetto FIRB RETI: Rete Italiana Chimica Farmaceutica CHEM-PROFARMA-NET [RBPR05NWWC], and Fondazione CARIPLO.

Author details

[1]Scientific Institute IRCCS E. Medea, Bioinformatic Lab, Via don L. Monza 20, 23842 Bosisio Parini (LC), Italy. [2]Bioengineering Department, Politecnico di Milano, P.zza L. da Vinci, 32, 20133 Milan, Italy. [3]Dino Ferrari Centre, Department of Neurological Sciences, University of Milan, IRCCS Ospedale Maggiore Policlinico, Mangiagalli and Regina Elena Foundation, Via F. Sforza 35, 20100 Milan, Italy. [4]Department of Biomedical Sciences and Technologies LITA Segrate, University of Milan, Via F.lli Cervi 93, 20090 Milan, Italy. [5]Don C. Gnocchi ONLUS Foundation IRCCS, Via Capecelatro 66, 20148 Milan, Italy.

Authors' contributions

MF, UP, RC and MS performed the analyzes, analyzed and interpreted the data; GPC and NB participated in the study coordination; MC and MS conceived the study and wrote the paper.

All authors read and approved the final manuscript.

Received: 12 February 2010 Accepted: 31 August 2010
Published: 31 August 2010

References

1. Hotez PJ, Brindley PJ, Bethony JM, King CH, Pearce EJ, Jacobson J: **Helminth infections: the great neglected tropical diseases.** *J Clin Invest* 2008, **118(4)**:1311-1321.
2. Quinnell RJ: **Genetics of susceptibility to human helminth infection.** *Int J Parasitol* 2003, **33(11)**:1219-1231.
3. Dunne DW, Cooke A: **A worm's eye view of the immune system: consequences for evolution of human autoimmune disease.** *Nat Rev Immunol* 2005, **5(5)**:420-426.
4. Fumagalli M, Pozzoli U, Cagliani R, Comi GP, 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* 2009, **206(6)**:1395-1408.
5. Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi GP, Menozzi G, Bresolin N, Sironi M: **Widespread balancing selection and pathogen-driven selection at blood group antigen genes.** *Genome Res* 2009, **19(2)**:199-212.
6. Prugnolle F, Manica A, Charpentier M, Guegan JF, Guernier V, Balloux F: **Pathogen-driven selection and worldwide HLA class I diversity.** *Curr Biol* 2005, **15(11)**:1022-1027.
7. Pullan R, Brooker S: **The health impact of polyparasitism in humans: are we under-estimating the burden of parasitic diseases?** *Parasitology* 2008, **135(7)**:783-794.
8. Guernier V, Hochberg ME, Guegan JF: **Ecology drives the worldwide distribution of human diseases.** *PLoS Biol* 2004, **2(6)**:e141.
9. Vercelli D: **Discovering susceptibility genes for asthma and allergy.** *Nat Rev Immunol* 2008, **8(3)**:169-182.
10. Clerici M, Butto S, Lukwiya M, Saresella M, Declich S, Trabattoni D, Pastori C, Piconi S, Fracasso C, Fabiani M, Ferrante P, Rizzardini G, Lopalco L: **Immune activation in africa is environmentally-driven and is associated with upregulation of CCR5. Italian-Ugandan AIDS Project.** *AIDS* 2000, **14(14)**:2083-2092.
11. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM: **Worldwide human relationships inferred from genome-wide patterns of variation.** *Science* 2008, **319(5866)**:1100-1104.
12. Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, Myers RM, Cavalli-Sforza LL, Feldman MW, Pritchard JK: **The role of geography in human adaptation.** *PLoS Genet* 2009, **5(6)**:e1000500.
13. Handley LJ, Manica A, Goudet J, Balloux F: **Going the distance: human population genetics in a clinal world.** *Trends Genet* 2007, **23(9)**:432-439.
14. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L: **Natural selection has driven population differentiation in modern humans.** *Nat Genet* 2008, **40(3)**:340-345.
15. Voight BF, Kudaravalli S, Wen X, Pritchard JK: **A map of recent positive selection in the human genome.** *PLoS Biol* 2006, **4(3)**:e72.
16. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, *et al*: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449(7164)**:851-861.
17. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES, International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, *et al*: **Genome-wide detection and characterization of positive selection in human populations.** *Nature* 2007, **449(7164)**:913-918.
18. Fujimoto A, Kimura R, Ohashi J, Omi K, Yuliwulandari R, Batubara L, Mustofa MS, Samakkarn U, Settheetham-Ishida W, Ishida T, Morishita Y, Furusawa T, Nakazawa M, Ohtsuka R, Tokunaga K: **A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness.** *Hum Mol Genet* 2008, **17(6)**:835-843.
19. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES: **Detecting recent positive selection in the human genome from haplotype structure.** *Nature* 2002, **419(6909)**:832-837.
20. Thomas PD, Kejariwal A, Guo N, Mi H, Campbell MJ, Muruganujan A, Lazareva-Ulitsky B: **Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools.** *Nucleic Acids Res* 2006, **34(Web Server)**:W645-50.
21. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A: **PANTHER: a library of protein families and subfamilies indexed by function.** *Genome Res* 2003, **13(9)**:2129-2141.
22. Cho RJ, Campbell MJ: **Transcription, genomes, function.** *Trends Genet* 2000, **16(9)**:409-415.
23. Levite M: **Neurotransmitters activate T-cells and elicit crucial functions via neurotransmitter receptors.** *Curr Opin Pharmacol* 2008, **8(4)**:460-471.
24. Kavelaars A: **Regulated expression of alpha-1 adrenergic receptors in the immune system.** *Brain Behav Immun* 2002, **16(6)**:799-807.
25. Heijnen CJ, Rouppe van der Voort C, van de Pol M, Kavelaars A: **Cytokines regulate alpha(1)-adrenergic receptor mRNA expression in human monocytic cells and endothelial cells.** *J Neuroimmunol* 2002, **125(1-2)**:66-72.
26. Kelley KW, Weigent DA, Kooijman R: **Protein hormones and immunity.** *Brain Behav Immun* 2007, **21(4)**:384-392.
27. Wahab MF, Warren KS, Levy RP: **Function of the thyroid and the host-parasite relation in murine schistosomiasis mansoni.** *J Infect Dis* 1971, **124(2)**:161-171.
28. Saule P, Adriaenssens E, Delacre M, Chassande O, Bossu M, Auriault C, Wolowczuk I: **Early variations of host thyroxine and interleukin-7 favor Schistosoma mansoni development.** *J Parasitol* 2002, **88(5)**:849-855.
29. Gudbjartsson DF, Bjornsdottir US, Halapi E, Helgadottir A, Sulem P, Jonsdottir GM, Thorleifsson G, Helgadottir H, Steinthorsdottir V, Stefansson H, Williams C, Hui J, Beilby J, Warrington NM, James A, Palmer LJ, Koppelman GH, Heinzmann A, Krueger M, Boezen HM, Wheatley A, Altmuller J, Shin HD, Uh ST, Cheong HS, Jonsdottir B, Gislason D, Park CS, Rasmussen LM, Porsbjerg C, *et al*: **Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction.** *Nat Genet* 2009, **41(3)**:342-347.
30. Kabesch M, Tzotcheva I, Carr D, Hoffer C, Weiland SK, Fritzsch C, von Mutius E, Martinez FD: **A complete screening of the IL4 gene: novel polymorphisms and their association with asthma and IgE in childhood.** *J Allergy Clin Immunol* 2003, **112(5)**:898-898.
31. Maizels RM, Balic A, Gomez-Escobar N, Nair M, Taylor MD, Allen JE: **Helminth parasites–masters of regulation.** *Immunol Rev* 2004, **201**:89-116.
32. Maizels RM, Yazdanbakhsh M: **Immune regulation by helminth parasites: cellular and molecular mechanisms.** *Nat Rev Immunol* 2003, **3(9)**:733-744.
33. Gorczynski R, Khatri I, Lee L, Boudakov I: **An interaction between CD200 and monoclonal antibody agonists to CD200R2 in development of dendritic cells that preferentially induce populations of CD4+CD25+ T regulatory cells.** *J Immunol* 2008, **180(9)**:5946-5955.
34. Oh-Hora M, Yamashita M, Hogan PG, Sharma S, Lamperti E, Chung W, Prakriya M, Feske S, Rao A: **Dual functions for the endoplasmic reticulum calcium sensors STIM1 and STIM2 in T cell activation and tolerance.** *Nat Immunol* 2008, **9(4)**:432-443.
35. Takahashi T, Tagami T, Yamazaki S, Uede T, Shimizu J, Sakaguchi N, Mak TW, Sakaguchi S: **Immunologic self-tolerance maintained by CD25(+)CD4(+) regulatory T cells constitutively expressing cytotoxic T lymphocyte-associated antigen 4.** *J Exp Med* 2000, **192(2)**:303-310.
36. McCoy K, Camberis M, Gros GL: **Protective immunity to nematode infection is induced by CTLA-4 blockade.** *J Exp Med* 1997, **186(2)**:183-187.
37. Chen L: **Co-inhibitory molecules of the B7-CD28 family in the control of T-cell immunity.** *Nat Rev Immunol* 2004, **4(5)**:336-347.
38. King CL, Xianli J, June CH, Abe R, Lee KP: **CD28-deficient mice generate an impaired Th2 response to Schistosoma mansoni infection.** *Eur J Immunol* 1996, **26(10)**:2448-2455.
39. Smith P, Walsh CM, Mangan NE, Fallon RE, Sayers JR, McKenzie AN, Fallon PG: **Schistosoma mansoni worms induce anergy of T cells via selective up-regulation of programmed death ligand 1 on macrophages.** *J Immunol* 2004, **173(2)**:1240-1248.
40. Lo SK, Rahman A, Xu N, Zhou MY, Nagpala P, Jaffe HA, Malik AB: **Neutrophil inhibitory factor abrogates neutrophil adhesion by blockade of CD11a and CD11b beta(2) integrins.** *Mol Pharmacol* 1999, **56(5)**:926-932.
41. Moyle M, Foster DL, McGrath DE, Brown SM, Laroche Y, De Meutter J, Stanssens P, Bogowitz CA, Fried VA, Ely JA: **A hookworm glycoprotein that**

inhibits neutrophil function is a ligand of the integrin CD11b/CD18. *J Biol Chem* 1994, **269(13)**:10008-10015.
42. Daub J, Loukas A, Pritchard DI, Blaxter M: **A survey of genes expressed in adults of the human hookworm, Necator americanus.** *Parasitology* 2000, **120(Pt 2)**:171-184.
43. Giacomin PR, Wang H, Gordon DL, Botto M, Dent LA: **Loss of complement activation and leukocyte adherence as Nippostrongylus brasiliensis develops within the murine host.** *Infect Immun* 2005, **73(11)**:7442-7449.
44. Giacomin PR, Gordon DL, Botto M, Daha MR, Sanderson SD, Taylor SM, Dent LA: **The role of complement in innate, adaptive and eosinophil-dependent immunity to the nematode Nippostrongylus brasiliensis.** *Mol Immunol* 2008, **45(2)**:446-455.
45. Kerepesi LA, Hess JA, Nolan TJ, Schad GA, Abraham D: **Complement component C3 is required for protective innate and adaptive immunity to larval strongyloides stercoralis in mice.** *J Immunol* 2006, **176(7)**:4315-4322.
46. La Flamme AC, MacDonald AS, Huxtable CR, Carroll M, Pearce EJ: **Lack of C3 affects Th2 response development and the sequelae of chemotherapy in schistosomiasis.** *J Immunol* 2003, **170(1)**:470-476.
47. Osborne BA, Minter LM: **Notch signalling during peripheral T-cell activation and differentiation.** *Nat Rev Immunol* 2007, **7(1)**:64-75.
48. Asano N, Watanabe T, Kitani A, Fuss IJ, Strober W: **Notch1 signaling and regulatory T cell function.** *J Immunol* 2008, **180(5)**:2796-2804.
49. Samon JB, Champhekar A, Minter LM, Telfer JC, Miele L, Fauq A, Das P, Golde TE, Osborne BA: **Notch1 and TGFbeta1 cooperatively regulate Foxp3 expression and the maintenance of peripheral regulatory T cells.** *Blood* 2008, **112(5)**:1813-1821.
50. Blumenthal SG, Aichele G, Wirth T, Czernilofsky AP, Nordheim A, Dittmer J: **Regulation of the human interleukin-5 promoter by Ets transcription factors. Ets1 and Ets2, but not Elf-1, cooperate with GATA3 and HTLV-I Tax1.** *J Biol Chem* 1999, **274(18)**:12910-12916.
51. Romano-Spica V, Georgiou P, Suzuki H, Papas TS, Bhat NK: **Role of ETS1 in IL-2 gene expression.** *J Immunol* 1995, **154(6)**:2724-2732.
52. Thomas RS, Tymms MJ, McKinlay LH, Shannon MF, Seth A, Kola I: **ETS1, NFkappaB and AP1 synergistically transactivate the human GM-CSF promoter.** *Oncogene* 1997, **14(23)**:2845-2855.
53. Jian JL, Zhu CS, Xu ZW, Ouyang WM, Ma DC, Zhang Y, Chen LJ, Yang AG, Jin BQ: **Identification and characterization of the CD226 gene promoter.** *J Biol Chem* 2006, **281(39)**:28731-28736.
54. Grenningloh R, Kang BY, Ho IC: **Ets-1, a functional cofactor of T-bet, is essential for Th1 inflammatory responses.** *J Exp Med* 2005, **201(4)**:615-626.
55. Schwarzmann N, Kunerth S, Weber K, Mayr GW, Guse AH: **Knock-down of the type 3 ryanodine receptor impairs sustained Ca^{2+} signaling via the T cell receptor/CD3 complex.** *J Biol Chem* 2002, **277(52)**:50636-50642.
56. Dammermann W, Zhang B, Nebel M, Cordiglieri C, Odoardi F, Kirchberger T, Kawakami N, Dowden J, Schmid F, Dornmair K, Hohenegger M, Flügel A, Guse AH, Potter BV: **NAADP-mediated Ca^{2+} signaling via type 1 ryanodine receptor in T cells revealed by a synthetic NAADP antagonist.** *Proc Natl Acad Sci USA* 2009, **106(26)**:10678-10683.
57. Vukcevic M, Spagnoli GC, Iezzi G, Zorzato F, Treves S: **Ryanodine receptor activation by Ca v 1.2 is involved in dendritic cell major histocompatibility complex class II surface expression.** *J Biol Chem* 2008, **283(50)**:34913-34922.
58. Shiue L, Green J, Green OM, Karas JL, Morgenstern JP, Ram MK, Taylor MK, Zoller MJ, Zydowsky LD, Bolen JB: **Interaction of p72syk with the gamma and beta subunits of the high-affinity receptor for immunoglobulin E, Fc epsilon RI.** *Mol Cell Biol* 1995, **15(1)**:272-281.
59. Zhang J, Berenstein EH, Evans RL, Siraganian RP: **Transfection of Syk protein tyrosine kinase reconstitutes high affinity IgE receptor-mediated degranulation in a Syk-negative variant of rat basophilic leukemia RBL-2H3 cells.** *J Exp Med* 1996, **184(1)**:71-79.
60. Bachelet I, Munitz A, Mankutad D, Levi-Schaffer F: **Mast cell costimulation by CD226/CD112 (DNAM-1/Nectin-2): a novel interface in the allergic process.** *J Biol Chem* 2006, **281(37)**:27190-27196.
61. Matsubara S, Li G, Takeda K, Loader JE, Pine P, Masuda ES, Miyahara N, Miyahara S, Lucas JJ, Dakhama A, Gelfand EW: **Inhibition of spleen tyrosine kinase prevents mast cell activation and airway hyperresponsiveness.** *Am J Respir Crit Care Med* 2006, **173(1)**:56-63.
62. Mathur M, Herrmann K, Qin Y, Gulmen F, Li X, Krimins R, Weinstock J, Elliott D, Bluestone JA, Padrid P: **CD28 interactions with either CD80 or CD86 are sufficient to induce allergic airway inflammation in mice.** *Am J Respir Cell Mol Biol* 1999, **21(4)**:498-509.
63. Crosby JR, Guha M, Tung D, Miller DA, Bender B, Condon TP, York-DeFalco C, Geary RS, Monia BP, Karras JG, Gregory SA: **Inhaled CD86 antisense oligonucleotide suppresses pulmonary inflammation and airway hyper-responsiveness in allergic mice.** *J Pharmacol Exp Ther* 2007, **321(3)**:938-946.
64. Barnes KC, Grant AV, Baltadzhieva D, Zhang S, Berg T, Shao L, Zambelli-Weiner A, Anderson W, Nelsen A, Pillai S, Yarnall DP, Dienger K, Ingersoll RG, Scott AF, Fallin MD, Mathias RA, Beaty TH, Garcia JG, Wills-Karp M: **Variants in the gene encoding C3 are associated with asthma and related phenotypes among African Caribbean families.** *Genes Immun* 2006, **7(1)**:27-35.
65. Drouin SM, Corry DB, Kildsgaard J, Wetsel RA: **Cutting edge: the absence of C3 demonstrates a role for complement in Th2 effector functions in a murine model of pulmonary allergy.** *J Immunol* 2001, **167(8)**:4141-4145.
66. Guseh JS, Bores SA, Stanger BZ, Zhou Q, Anderson WJ, Melton DA, Rajagopal J: **Notch signaling promotes airway mucous metaplasia and inhibits alveolar development.** *Development* 2009, **136(10)**:1751-1759.
67. Grunstein MM, Hakonarson H, Maskeri N, Kim C, Chuang S: **Intrinsic ICAM-1/LFA-1 activation mediates altered responsiveness of atopic asthmatic airway smooth muscle.** *Am J Physiol Lung Cell Mol Physiol* 2000, **278(6)**:L1154-63.
68. Nakamura Y, Esnault S, Maeda T, Kelly EA, Malter JS, Jarjour NN: **Ets-1 regulates TNF-alpha-induced matrix metalloproteinase-9 and tenascin expression in primary bronchial fibroblasts.** *J Immunol* 2004, **172(3)**:1945-1952.
69. Cooper PJ: **Interactions between helminth parasites and allergy.** *Curr Opin Allergy Clin Immunol* 2009, **9(1)**:29-37.
70. Hopkin J: **Immune and genetic aspects of asthma, allergy and parasitic worm infections: evolutionary links.** *Parasite Immunol* 2009, **31(5)**:267-273.
71. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R: **Localizing recent adaptive evolution in the human genome.** *PLoS Genet* 2007, **3(6)**:e90.
72. Tang K, Thornton KR, Stoneking M: **A new approach for using genome scans to detect recent positive selection in the human genome.** *PLoS Biol* 2007, **5(7)**:e171.
73. Bryk J, Hardouin E, Pugach I, Hughes D, Strotmann R, Stoneking M, Myles S: **Positive selection in East Asians for an EDAR allele that enhances NF-kappaB activation.** *PLoS One* 2008, **3(5)**:e2209.
74. Rosenberg NA: **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* 2006, **70(Pt 6)**:841-847.
75. Salkind NJ: *Encyclopedia of measurement and statistics* Thousand Oaks, CA: Sage Publications 2007.
76. R Development Core Team: *R: A Language and Environment for Statistical Computing. Vienna, Austria* 2008.
77. Kraus DM, Elliott GS, Chute H, Horan T, Pfenninger KH, Sanford SD, Foster S, Scully S, Welcher AA, Holers VM: **CSMD1 is a novel multiple domain complement-regulatory protein highly expressed in the central nervous system and epithelial tissues.** *J Immunol* 2006, **176**:4419-4430.
78. Park CG, Lee SY, Kandala G, Lee SY, Choi Y: **A novel gene product that couples TCR signaling to Fas(CD95) expression in activation-induced cell death.** *Immunity* 1996, **4**:583-591.
79. Besecker B, Bao S, Bohacova B, Papp A, Sadee W, Knoell DL: **The human zinc transporter SLC39A8 (Zip8) is critical in zinc-mediated cytoprotection in lung epithelia.** *Am J Physiol Lung Cell Mol Physiol* 2008, **294**:1127-1136.
80. Begum NA, Kobayashi M, Moriwaki Y, Matsumoto M, Toyoshima K, Seya T: **Mycobacterium bovis BCG cell wall and lipopolysaccharide induce a novel gene, BIGM103, encoding a 7-TM protein: identification of a new protein family having Zn-transporter and Zn-metalloprotease signatures.** *Genomics* 2002, **80**:630-645.
81. Blanchet F, Cardona A, Letimier FA, Hershfield MS, Acuto O: **CD28 costimulatory signal induces protein arginine methylation in T cells.** *J Exp Med* 2005, **202**:371-377.
82. Li H, Kim JH, Koh SS, Stallcup MR: **Synergistic effects of coactivators GRIP1 and beta-catenin on gene activation: cross-talk between androgen receptor and Wnt signaling pathways.** *J Biol Chem* 2003, **279**:4212-4220.

83. Kwon HJ, Bittner EH, Vig-Varga E, Luo Y, Lee Y, Goebl MG, Harrington MA: **Tumor necrosis factor alpha induction of NF-kappaB requires the novel coactivator SIMPL.** *Mol Cell Biol* 2004, **24**:9317-9326.
84. van Es MA, van Vught PW, Blauw HM, Franke L, Saris CG, Van den Bosch L, de Jong SW, de Jong V, Baas F, van't Slot R, Lemmens R, Schelhaas HJ, Birve A, Sleegers K, Van Broeckhoven C, Schymick JC, Traynor BJ, Wokke JH, Wijmenga C, Robberecht W, Andersen PM, Veldink JH, Ophoff RA, van den Berg LH: **Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis.** *Nat Genet* 2008, **40**:29-31.
85. Wang J, Kudoh J, Takayanagi A, Shimizu N: **Novel human BTB/POZ domain-containing zinc finger protein ZNF295 is directly associated with ZFP161.** *Biochem Biophys Res Commun* 2005, **327**:615-627.

doi:10.1186/1471-2148-10-264
Cite this article as: Fumagalli *et al.*: The landscape of human genes involved in the immune response to parasitic worms. *BMC Evolutionary Biology* 2010 **10**:264.

2.7 Genetic diversity at endoplasmic reticulum aminopeptidases is maintained by balancing selection and is associated with natural resistance to HIV-1 infection

HMG Advance Access published September 29, 2010

Human Molecular Genetics, 2010 **1–10**
doi:10.1093/hmg/ddq401

Genetic diversity at endoplasmic reticulum aminopeptidases is maintained by balancing selection and is associated with natural resistance to HIV-1 infection

Rachele Cagliani[1], Stefania Riva[1], Mara Biasin[2], Matteo Fumagalli[1,4], Uberto Pozzoli[1], Sergio Lo Caputo[5], Francesco Mazzotta[5], Luca Piacentini[2], Nereo Bresolin[1,6], Mario Clerici[3,7,†] and Manuela Sironi[1,*,†]

[1]Bioinformatic Laboratory, Scientific Institute IRCCS E. Medea, Via don L. Monza 20, 23842 Bosisio Parini (LC), Italy, [2]DISC LITA Vialba and [3]Department of Biomedical Sciences and Technologies LITA Segrate, University of Milan, Milan, Italy, [4]Bioengineering Department, Politecnico di Milano, P.zza L. da Vinci, 32, 20133 Milan, Italy, [5]Divisione Malattie Infettive, Ospedale S.M. Annunziata, Florence, Italy, [6]Department of Neurological Sciences, Dino Ferrari Centre,University of Milan, IRCCS Ospedale Maggiore Policlinico, Mangiagalli and Regina Elena Foundation, Via F. Sforza 35, 20100 Milan, Italy and [7]Fondazione Don C. Gnocchi, IRCCS, Milan, Italy

Received August 2, 2010; Revised and Accepted September 10, 2010

Human *ERAP1* and *ERAP2* encode two endoplasmic reticulum aminopeptidases. These enzymes trim peptides to optimal size for loading onto major histocompatibility complex class I molecules and shape the antigenic repertoire presented to CD8$^+$ T cells. Therefore, *ERAP1* and *ERAP2* may be considered potential selection targets and modulators of infection susceptibility. We resequenced two genic regions in *ERAP1* and *ERAP2* in three HapMap populations. In both cases, we observed high levels of nucleotide variation, an excess of intermediate-frequency alleles, and reduced population genetic differentiation. The genealogy of *ERAP1* and *ERAP2* haplotypes was split into two major branches with deep coalescence times. These features suggest that long-standing balancing selection has acted on these genes. Analysis of the Lys528Arg (rs30187 in *ERAP1*) and Asn392Lys (rs2549782 in *ERAP2*) variants in an Italian population of HIV-1-exposed seronegative (ESN) individuals and a larger number of Italian controls indicated that rs2549782 significantly deviates from Hardy–Weinberg equilibrium (HWE) in ESN but not in controls. Technical errors were excluded and a goodness-of-fit test indicated that a recessive model with only genetic effects adequately explains HWE deviation. The genotype distribution of rs2549782 is significantly different in the two cohorts ($P = 0.004$), mainly as the result of an over-representation of Lys/Lys genotypes in the ESN sample (*P*-value for a recessive model: 0.00097). Our data suggest that genetic diversity in *ERAP1* and *ERAP2* has been maintained by balancing selection and that variants in *ERAP2* confer resistance to HIV-1 infection possibly via the presentation of a distinctive peptide repertoire to CD8$^+$ T cells.

INTRODUCTION

Cells expressing foreign proteins in association with a major histocompatibility complex (MHC) class I molecule are recognized and eliminated by CD8$^+$ T cells; this pathway is specialized in the recognition of intracellular pathogens including viruses. The composition of the peptide repertoire displayed by the MHC I largely depends on the

*To whom correspondence should be addressed. Tel: +39 031877915; Fax: +39 031877499; Email: manuela.sironi@bp.lnf.it
†These authors contributed equally to this work.

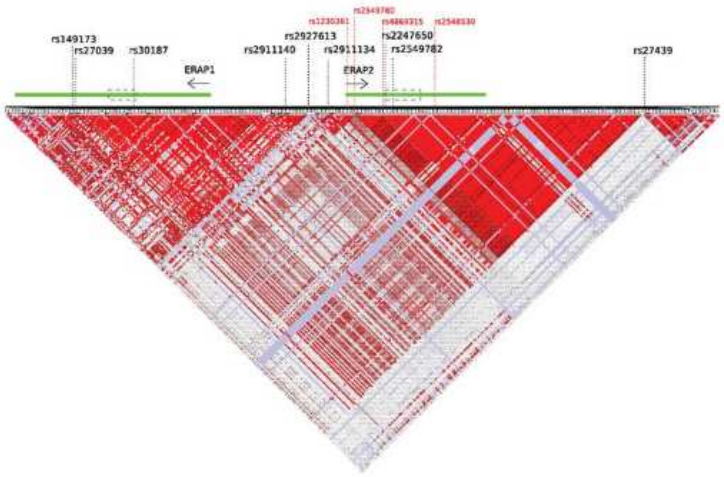


Figure 1. Schematic diagram of the genomic region encompassing *ERAP1* and *ERAP2*. Green lines represent the *ERAP1* and *ERAP2* gene regions. The direction of transcription is marked by the arrow. The position of SNPs mentioned in the text is shown. Tag SNPs are shown in red. The two regions we resequenced are denoted by the hatched boxes. LD (D'/Lod) refers to CEU, and data were derived from HapMap.

specificity of the MHC peptide-binding groove, but also on the availability of suitable peptides generated by the antigen-processing pathway. Therefore, antigen processing is essential for assuring immune surveillance and for establishing immunodominance (i.e. the phenomenon whereby only a minority of epitopes are functionally immunogenic).

The processing of intracellular proteins is a multi-step event initiated by proteolysis in the cytosol that generates peptides between 2 and 25 amino acids in length (reviewed in 1). A fraction of these peptides is transported to the endoplasmic reticulum (ER) by the transporter associated with antigen-processing protein and trimmed at the N-terminus so as to generate optimal size fragments for MHC I to be loaded into MHC I clefts (reviewed in 1).

In humans, at least two aminopeptidases located in the ER, encoded by *ERAP1* (MIM *606832) and *ERAP2* (MIM *609497), act in concert to trim peptides at their N-terminus (2), whereas mice have only one ER-aminopeptidase (*Erap1*). Experiments in *Erap1*$^{-/-}$ mice indicated that the enzyme shapes the peptide repertoire displayed by MHC I molecules in both normal and virus-infected cells (3–5). In the absence of *Erap1*, some viral or endogenous peptides are presented at lower levels, whereas others display enhanced presentation (4–6). In human cells, RNA interference experiments have produced contrasting results concerning the effect of *ERAP1* and *ERAP2* on MHC class I molecule expression and peptide presentation (2,7), suggesting that factors such as peptide sequence, cell-type and MHC class I alleles may determine whether the two aminopeptidases enhance or inhibit antigen presentation.

These observations suggest that *ERAP1* and *ERAP2*, along with the other components of the antigen-processing and presentation pathway, play a role both in protecting from infectious diseases and in maintaining immunotolerance to self-peptides. In line with this view, variants in both genes have been associated with an increased risk of developing ankylosing spondylitis (8,9). Notably, the role of *ERAP1* and *ERAP2* polymorphisms in modulating the susceptibility to infection has never been studied.

MHC class I genes are extremely polymorphic in humans, and genetic variability is maintained by balancing selection which is, at least in part, pathogen-driven (10,11). Since the repertoire of peptides presented by the MHC I molecules ultimately depends on *ERAP1* and *ERAP2*, these genes may be considered potential selection targets, as diverse alleles may display differential activity for specific peptides.

Here we show that long-standing balancing selection has maintained genetic variability at the human *ERAP1* and *ERAP2* genes and that variants in *ERAP2* are associated with natural resistance against HIV-1 infection.

RESULTS

Nucleotide diversity and neutrality tests

ERAP1 and *ERAP2* are located in a head-to-head orientation on the long arm of chromosome 5 (Fig. 1). Analysis of linkage disequilibrium (LD) in three HapMap populations, namely Yoruba (YRI), Europeans (CEU) and East Asians (EAS), indicated that the two genes lie in distinct LD blocks (Fig. 1 and Supplementary Material, Fig. S1).

Throughout the manuscript, single-nucleotide polymorphisms (SNPs) are indicated using the NCBI notation for reference SNP cluster IDs (rs#) (http://www.ncbi.nlm.nih.gov/projects/SNP/), and derived alleles are defined through comparison with the chimpanzee reference sequence. In order to

Table 1. Nucleotide diversity and neutrality tests for the *ERAP1* and *ERAP2* gene regions

Gene	Pop.[a]	N[b]	S[c]	θ_W ($\times 10^{-4}$) Value	Rank[d]	π ($\times 10^{-4}$) Value	Rank[d]	Tajima's D Value (P)[e]	Rank[d]	Fu and Li's D^* Value (P)[e]	Rank[d]	Fu and Li's F^* Value (P)[e]	Rank[d]
ERAP1	YRI	40	36	21.33	0.98	26.32	0.99	0.82 (0.0456)	0.95	−0.28 (0.414)	0.51	0.12 (0.221)	0.68
	CEU	40	37	21.92	0.99	26.13	0.99	0.67 (0.240)	0.73	−0.43 (0.650)	0.41	−0.06 (0.502)	0.53
	EAS	40	25	14.81	0.97	26.90	0.99	2.76 (0.004)	0.99	1.45 (0.029)	0.98	2.24 (0.003)	0.99
ERAP2	YRI	40	37	13.11	0.86	21.91	0.99	2.35 (<0.001)	>0.99	1.02 (0.024)	0.97	1.75 (0.001)	>0.99
	CEU	40	42	14.88	0.98	23.95	0.99	2.15 (0.013)	0.97	0.58 (0.226)	0.75	1.33 (0.055)	0.93
	EAS	40	33	11.69	0.95	23.06	0.98	3.37 (<0.001)	>0.99	1.59 (0.005)	>0.99	2.61 (<0.001)	>0.99

[a]Population.
[b]Sample size (chromosomes).
[c]Number of segregating sites.
[d]Percentile rank relative to a distribution of 238 2 kb windows from NIEHS genes.
[e]*P*-value calculated by coalescent simulations.

test the hypothesis whereby balancing selection has maintained nucleotide variability at *ERAP1* and *ERAP2*, we resequenced two genomic regions internal to these genes in the same three populations. Specifically, an ~3.9 kb region encompassing the Lys528Arg variant (rs30187) was analysed for *ERAP1* (Fig. 1). This choice was motivated by the fact that, although multiple variants in the gene have been associated with ankylosing spondylitis (12), the 528Arg allele was also shown to decrease enzymatic activity (13), suggesting that rs30187 may represent a functional variant. The derived 528Arg allele has also been associated with essential hypertension (14) and haemolytic uraemic syndrome (15), and a previous analysis (16) indicated that rs30187 defines an expression QTL (eQTL) for *ERAP1*. This same observation applies to rs2247650, an SNP located within *ERAP2* (16). This SNP lies relatively close to rs2549782 (Fig. 1), which determines the substitution of the highly conserved asparagine residue at codon 392 with a lysine; the polymorphism has been associated with both ankylosing spondylitis and pre-eclampsia (17,18). Thus, an ~6.6 kb region within *ERAP2* and covering both variants was resequenced (Fig. 1).

Forty-nine and 48 SNPs were detected in the *ERAP1* and *ERAP2* regions, respectively, none of which represented a novel nonsynonymous variant.

Two major effects of balancing selection that can be detected through resequencing data are (i) a distortion of the site frequency spectrum (SFS) towards intermediate-frequency alleles and (ii) an excess of diversity due to the maintenance of polymorphisms linked to the selected variant(s).

Common population genetic tests based on the SFS include Tajima's D (D_T) (19) and Fu and Li's D^* and F^* (20). D_T tests the departure from neutrality by comparing two nucleotide diversity indexes: θ_W (21), an estimate of the expected per-site heterozygosity, and π (22), the average number of pairwise sequence nucleotide differences. Positive values of D_T indicate an excess of intermediate-frequency variants. Fu and Li's F^* and D^* are also based on SNP frequency spectra and differ from D_T in that they also take into account whether mutations occur in external or internal branches of a genealogy (20). As an empirical comparison, θ_W, π, as well as D_T, F^* and D^* were calculated for 5 kb windows (thereafter referred to as reference windows), deriving from 238 genes resequenced by the NIEHS program in CEU, YRI and EAS. Additionally, the statistical significance of neutrality tests was evaluated by performing coalescent simulations with a population genetic model that incorporates demographic scenarios (see Materials and Methods).

As shown in Table 1, the regions we analysed in both *ERAP1* and *ERAP2* display extreme nucleotide diversity, with both θ_W and π ranking above the 95th percentile in the distribution of 5 kb reference windows in all populations, with the exception of θ_W in YRI.

All tests in Table 1 rejected neutral evolution at *ERAP1* in EAS, and D_T was significantly high in YRI. Conversely, no significant deviation from neutrality was observed, using these tests, in CEU; this is partially due to the presence of a single divergent haplotype in this population (see what follows and Fig. 2) that introduces several singletons and affects SFS-based statistics.

As for *ERAP2*, most tests yielded significantly high results, using both simulations and empirical comparisons in all populations.

A hallmark of balancing selection is an excess of polymorphism compared with neutral expectations. Indeed, our data (Table 1) indicate that nucleotide diversity indexes are extremely high for both *ERAP1* and *ERAP2*. Yet, polymorphism levels also depend on local mutation rates, and under neutral evolution, the amount of within- and between-species diversity is expected to be similar at all loci in the genome (23). The multi-locus HKA test was developed to verify this expectation (24). We applied a multi-locus MLHKA (maximum-likelihood HKA) test by comparing polymorphism and divergence levels at the *ERAP1* and *ERAP2* genomic regions with 16 NIEHS genes resequenced in YRI, CEU and EAS. The results are shown in Table 2 and indicate that a significant excess of nucleotide diversity versus divergence is detectable in all populations for both *ERAP1* and *ERAP2*.

F_{ST} (25) measures variations in allele frequencies among populations (genetic differentiation) and largely depends on demographic history (which affects all loci equally). Yet, natural selection may drive allele frequencies to differ more or less than expected on the basis of demography alone. In particular, balancing selection may lead to a decrease in population differentiation compared with neutrally evolving loci (26). F_{ST} among YRI, CEU and EAS calculated for the

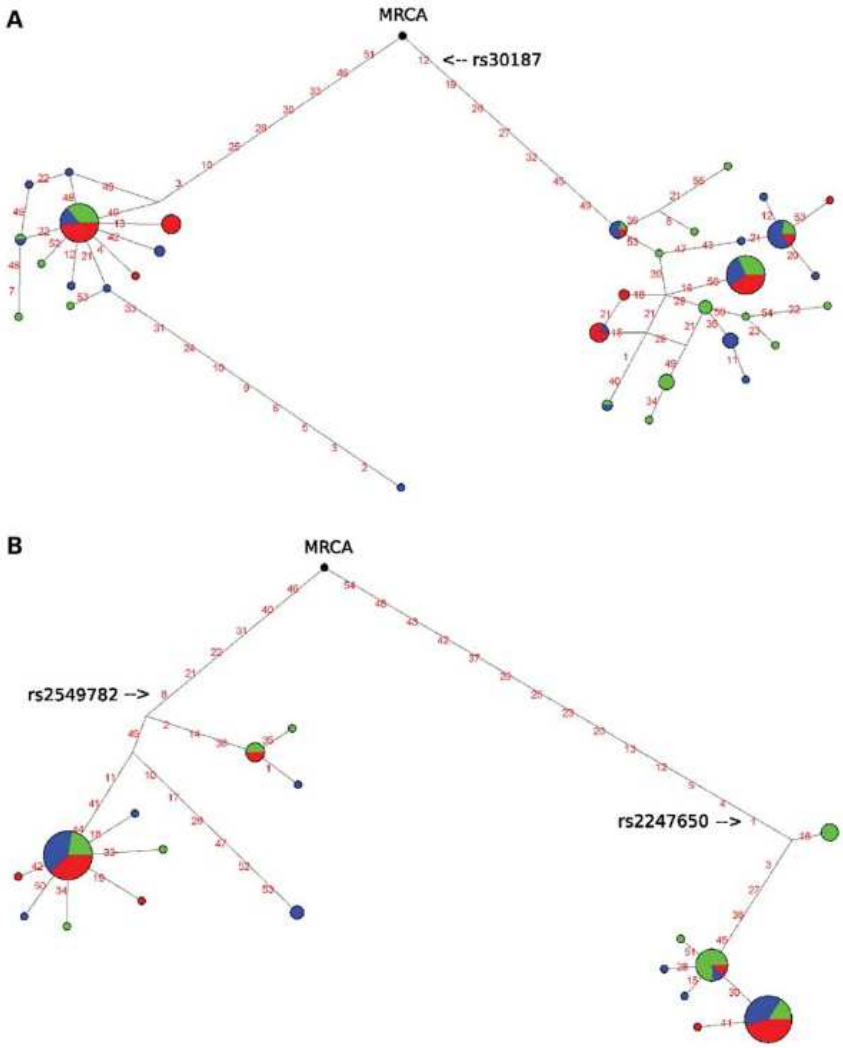


Figure 2. Genealogy of *ERAP1* and *ERAP2* haplotypes reconstructed through a median-joining network. The *ERAP1* and *ERAP2* networks are shown in (**A**) and (**B**), respectively. Each node represents a different haplotype, with the size of the circle proportional to frequency. Nucleotide differences between haplotypes are indicated on the branches of the network. Circles are colour-coded according to population (green: YRI; blue: CEU; red: EAS). The most recent common ancestor (MRCA) is also shown (black circle). The relative position of mutations along a branch is arbitrary.

ERAP1 and *ERAP2* regions amounted to 0.022 and 0.0088, respectively. Both values are lower than the genome average of 0.123 (27), and their percentile rank in the distribution of F_{ST} calculated for 5 kb reference windows were 0.05 and 0.022, respectively, indicating that both regions display unusually low population genetic differentiation.

Haplotype analysis and time to the most recent common ancestor estimation

Another feature of balancing selection is the maintenance of two or more highly divergent haplotype clades; in cases of long-standing balancing selection, the coalescence time of

Table 2. MLHKA test for the two regions in *ERAP1* and *ERAP2*

Gene	Fixed substitutions	MLHKA YRI k^a	*P*-value	CEU k^a	*P*-value	EAS k^a	*P*-value
ERAP1	30	4.35	2.42×10^{-4}	6.46	5.19×10^{-6}	4.65	4.34×10^{-4}
ERAP2	34	4.4	6.58×10^{-4}	6.09	2.86×10^{-6}	5.32	3.6×10^{-5}

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

these genealogies is deeper than expected under neutrality (28). In order to analyse the haplotype structure of the two regions in *ERAP1* and *ERAP2*, we constructed median-joining networks (Fig. 2). Both haplotype networks display two major clades separated by deep branches. In line with the extended LD pattern (Fig. 1 and Supplementary Material, Fig. S1), the haplotype network for the *ERAP2* gene region presented no reticulations and few recurrent mutations. Both the Asn392Lys variant (rs2549782, variant 8 in the network) and rs2247650 (variant 1) are located on the two major branches (Fig. 2). As for *ERAP1*, the network displays some reticulations and recurrent mutations/gene conversions. The rs30187 variant (Lys528Arg, variant 12 in the network) separates the two major clades. In order to estimate the time to the most recent common ancestor (TMRCA) of the *ERAP1* and *ERAP2* haplotype genealogies, we applied a phylogeny-based method (29) based on the average pairwise difference between the haplotype clusters. Using mutation rates based on the number of fixed differences with chimpanzee and a separation time of 6 million years (MY) (30), we estimated TMRCAs of 4.12 MY (SD: 0.860 MY) and 5.08 MY (SD: 0.918 MY) for *ERAP1* and *ERAP2*, respectively.

In order to obtain more robust TMRCA estimates, we used GENETREE, which is based on a maximum-likelihood coalescent analysis (31,32). The method assumes an infinite-site model without recombination; therefore, haplotypes and sites that violate these assumptions need to be removed: we removed 10 and 6 variants for *ERAP1* and *ERAP2*, respectively. The resulting trees, rooted using the chimpanzee sequence, were partitioned into two major branches (Fig. 3). Using this method, TMRCA estimates of 4.26 MY (SD: 0.493 MY; $N_e = 21200$) and 4.65 MY (SD: 0.355 MY; $N_e = 14470$) were obtained for *ERAP1* and *ERAP2*, respectively (Fig. 3). Recombination rate over the *ERAP1* and *ERAP2* regions we analysed is relatively low, and exclusion of recombinant haplotypes, when they represent a minority of the data set, is accepted practice in tree construction (33); nonetheless, estimation of TMRCA in recombining regions may not be robust if no recombination is assumed (34). Thus, we applied an additional method that reconstructs haplotype genealogies through the coalescent with recombination (ancestral recombination graph) (35,36). Using this method, we obtained an estimate of recombination rate and TMRCA for each marker along our entire gene regions (Supplementary Material, Fig. S2). Although recombination rate is extremely low for *ERAP2*, it increases in the middle of the *ERAP1* region. The posterior means of TMRCA estimates resulted to be around 2.8 and 4.2 MY for *ERAP1* and *ERAP2*, respectively, assuming that N_e for humans equals 10400 (37). Therefore, all TMRCA estimates we obtained are highly unlikely under neutrality, as estimates for neutrally evolving autosomal loci range between 0.8 and 1.5 MY (38).

ERAP1/*ERAP2* genotypes and resistance to HIV-1 infection

Given the role of *ERAP1* and *ERAP2* in antigen presentation, we wished to test whether the two nonsynonymous variants may modulate the susceptibility to HIV-1 infection. Most humans are susceptible to the virus, but a minority of individuals do not seroconvert despite multiple exposures. We genotyped rs30187 and rs2549782 in a cohort of well-characterized Italian heterosexual HIV-exposed seronegative (ESN) individuals who have a history of unprotected sex with their seropositive partners and in a larger sample of randomly selected Italian subjects (controls). No ESN was homozygous for the CCR5Δ32 variant, which confers resistance to R5 HIV-1 strains (39). rs30187 and rs2549782 display no LD in these samples ($r^2 = 0.017$, $D' = 0.19$), in line with the analysis presented in Figure 1. The Asn392Lys variant (rs2549782) significantly deviated from Hardy–Weinberg equilibrium (HWE) with an excess of homozygotes in the ESN sample; similarly, we observed a higher proportion of homozygote genotypes for rs30187 (in *ERAP1*) in the ESN sample only, with marginal statistical significance after Bonferroni correction (Table 3).

Deviations from HWE in cases (here represented by ESN subjects) may indicate a real association between SNP genotypes and the trait being analysed or may result from different effects such as population structure, sampling biases, unrecognized copy number variants or genotyping errors. As for the latter, the genomic DNA of all ESN subjects was PCR-amplified and directly resequenced twice, and no uncertainty in electropherograms was observed, suggesting that technical errors are not the cause for deviation from HWE. In order to verify whether the excess of homozygotes in ESN may be due to the presence of a large deletion segregating in this sample or to population structure, we genotyped six more SNPs in the *ERAP1*/*ERAP2* region. In particular, these variants were selected among those genotyped in the HapMap project to have a minor allele frequency similar to rs30187 and rs2549782 in CEU. Three of these SNPs are located in the intergenic region separating *ERAP1* and *ERAP2*, two are centromeric to *ERAP1* and one telomeric to *ERAP2* (Fig. 1). All these polymorphisms are located in LD blocks distinct from those where the two nonsynonymous variants lie and none of them deviated significantly from HWE in ESN

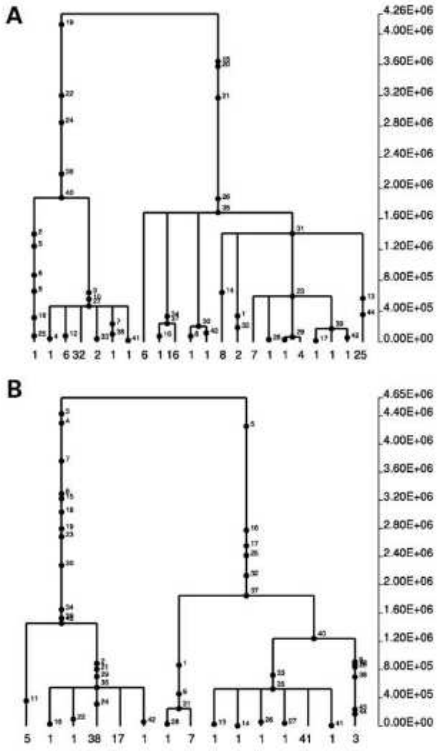


Figure 3. Estimated haplotype tree for the *ERAP1* and *ERAP2* gene regions we resequenced. The *ERAP1* and *ERAP2* trees are shown in (**A**) and (**B**), respectively. Mutations are represented as black dots and named for their physical position along the regions. The absolute frequency of each haplotype is also reported. Note that mutation numbering does not correspond to that reported in Figure 2.

(Table 3). Thus, unrecognized copy number variants or population structure is unlikely to be the cause of HWE deviation.

Wittke-Thompson *et al.* (40) have developed a test to verify whether deviations from HWE can be explained by an underlying genetic model for the trait being analysed rather than by other effects. Using a K_p (prevalence of ESN phenotype in the general population) of 0.20 (41,42), the best model fitting the genotypic proportions in cases and controls was a recessive model with q (susceptibility allele frequency) = 0.442, α (risk in non-susceptible homozygotes) = 0.162, β (heterozygote relative risk) = 1 and γ (homozygote relative risk) = 2.216. For this model, the goodness-of-fit test was not significant ($\chi^2 = 0.586$, $P = 0.75$, 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 0.05 to 0.30) values and similar results were obtained (not shown).

Comparison of genotype frequencies in the ESN and control samples indicated no significant difference for rs30187 in *ERAP1* (Table 3). Conversely, the genotype distribution of the *ERAP2* variant (rs2549782) was significantly different in the two cohorts (Bonferroni-corrected $P = 0.004$, Table 3), with the difference being mainly accounted for by an over-representation of GG (Lys/Lys) genotypes in the ESN sample (Bonferroni-corrected P-value for a recessive model = 0.00097, Table 3).

In order to test whether the association between rs2549782 and the ESN phenotype might be secondary to LD with other variants along the gene, we analysed four more markers (Fig. 1) selected to be tag SNPs (see Materials and Methods for details). No significant difference was observed in the genotype distribution of these variants between ESN and controls (Supplementary Material, Table S1). As for HWE, a tendency towards deviation with an excess of homozygotes in ESN was observed for rs4869315 (Supplementary Material, Table S1), located 1.7 kb upstream the Asn392Lys variant. This is likely due to partial LD between the two variants ($r^2 = 0.42$, $D' = 0.78$).

DISCUSSION

Pathogen-driven balancing selection is partially responsible for shaping nucleotide diversity at MHC class I molecules, with viruses acting as the strongest selective pressure (10). Since *ERAP1* and *ERAP2* perform the final and crucial step in the generation of MHC class I-binding peptides, we studied their selective pattern and tested whether variants in these genes may modulate the susceptibility to viral infections. Our population genetic analysis of the two aminopeptidases suggests that they have been subjected to long-standing balancing selection, and the low LD between the two genes suggests that they represent independent selection targets. Indeed, Andres *et al.* (43) have previously identified *ERAP2* in a genome-wide screen for genes subjected to balancing selection; in their analyses, these authors applied population genetic tests to resequencing data derived from exonic regions only. Therefore, the approach they applied is quite different from the one we describe herein where a 6.6 kb continuous region was resequenced and analysed. Thus, our analysis and that of Andres *et al.* can be regarded as largely independent demonstrations that balancing selection has been acting on *ERAP2*.

The action of natural selection obviously implies the presence of a functional variant with an effect on fitness and a selective pressure to act on it. Owing to the action of recombination and mutation, balancing selection signatures are expected to extend over relatively short genomic intervals (28), indicating that a selection target must be located within or in close proximity to the gene region we analysed. Specifically, the two nonsynonymous variants in the *ERAP1* and *ERAP2* regions are located on the branches separating the major haplotype clades (Fig. 2), suggesting that they may represent (or be in linkage with) the selected variants. In the case of *ERAP2*, this same observation applies to rs2247650, a possible eQTL (16). As for the selective pressure acting on these genes, given their role in antigen processing and presentation, it is conceivable that, in analogy to MHC class I genes, the two aminopeptidases may have mainly evolved in response to viral threats. Following these observations, Lys528Arg (in *ERAP1*)

Table 3. Genotype counts, HWE proportions and association analysis in ESN and controls

Phenotype	SNP ID	Genotype counts	Genotype counts (recessive model)	*P*-value[a] (HWE)	χ^2 (*P*-value)[a] (genotype)	χ^2 (*P*-value)[a] (recessive)
ESN	rs31087 (A/G)	14/24/31	14/55	0.062	5.175 (0.150)	3.47 (0.124)
Control	rs31087 (A/G)	25/104/89	25/193	>0.05		
ESN	rs2549782 (G/T)	25/25/19	25/44	0.05	12.34 (0.004)	12.18 (0.001)
Control	rs2549782 (G/T)	36/110/72	36/182	>0.05		
ESN	rs149173 (G/C)	11/28/30	—	>0.05	—	—
ESN	rs27039 (A/G)	9/29/31	—	>0.05	—	—
ESN	rs2911140 (C/T)	6/23/40	—	>0.05	—	—
ESN	rs2927613 (G/A)	9/39/21	—	>0.05	—	—
ESN	rs2911134 (C/T)	10/26/33	—	>0.05	—	—
ESN	rs27439 (G/T)	7/30/32	—	>0.05	—	

[a]*P*-values are Bonferroni-corrected.

and Asn392Lys (in *ERAP2*) may be regarded as possible genetic modifiers of viral infection susceptibility and we tested this hypothesis by analysing the genotype distribution of these variants in a cohort of Italian subjects who remained seronegative despite multiple exposures to HIV-1.

The role of *ERAP1* and *ERAP2* in antiviral response in humans has never been directly addressed, although previous data have indicated that the Ala148Pro escape mutation in the Gag protein of HIV abolishes its ability to be cleaved by ERAP1, resulting in decreased cytotoxic T-cell responses in chronically infected individuals (44). Interestingly, previous studies also reported that, in the presence of the same MHC class I alleles, the repertoire of peptides recognized by $CD8^+$ T cells differs in HIV-1-infected compared with ESN individuals (45). This observation suggests that variations in epitope immunodominance generated by the antigen-processing pathway may influence the susceptibility to HIV-1 infection.

The data herein indicate that variants in *ERAP2* are associated with resistance to HIV-1 infection. Specifically, a recessive model with GG (i.e. Lys/Lys) homozygotes protected from HIV-1 infection yielded the most significant result. Therefore, the question is whether or not a recessive model of resistance to HIV-1 afforded by the *ERAP2* variant is biologically plausible. The 392Lys allele changes an asparagine residue which is highly conserved in vertebrate aminopeptidases (Supplementary Material, Fig. S3). The functional effect of this variant has never been experimentally tested and, although we found an association with this polymorphism, we cannot exclude that the causal variant is located somewhere else in or outside the gene and in LD with rs2549782. In any case, starting from the hypothesis whereby $CD8^+$ T cells from ESN subjects recognize 'unconventional' peptides that may confer resistance to HIV-1 (45), we can imagine at least two possible reasons as to why a recessive model applies. First, these peptides may be destroyed by ERAP2 molecules carrying the non-protective variant so that they are available at no or low frequency in Lys/Asn and Asn/Asn cells. Second, these peptides may have lower affinity for MHC class I molecules and may be out-competed by peptides generated by ERAP2 molecules carrying the non-protective allele. Indeed, peptides generated by Erap1 cleavage compete for MHC-binding with those normally expressed by *Erap1*-deficient mouse cells (46).

As for *ERAP1* and its coding variant, both HWE proportions and distribution in ESN subjects versus controls were similar to those observed at *ERAP2*, but the statistical significance did not withstand Bonferroni correction. Therefore, larger cohorts will be required to address the role of this gene in infection susceptibility. An effort to exclude possible confounding effects and the role of other variants along the gene was made in the case of *ERAP2*; the data we report are based on a relatively small sample of ESN subjects and will therefore require an independent validation in a larger cohort. If these results will be confirmed, it is possible to speculate that specific alleles in *ERAP2* (and maybe in *ERAP1*) confer differential susceptibility to distinct pathogen species (one of these being HIV-1), and, therefore, balancing selection might be acting to maintain diversity in these genes under pathogen-driven selection. Yet, the recessive model for HIV-1 protection we describe suggests a situation different from that observed at MHC class I genes, which evolve under balancing selection that probably involves an element of heterozygote advantage (47–49). Yet, the molecular function of aminopeptidases is extremely different from that of MHC class I molecules, and the two possible scenarios we envisaged to explain the recessive model of HIV-1 resistance may well apply to other viral infections. Classic explanations for the action of balancing selection include, in addition to heterozygote advantage, adaptation to variable environmental conditions and frequency-dependent selection (reviewed in 28). Both these explanations may apply to the selection regime we described for *ERAP1* and *ERAP2* and they often denote host–pathogen interaction dynamics. Thus, distinct alleles in the two aminopeptidases might result in the differential processing of some peptides deriving from intracellular pathogens, resulting in a distinctive repertoire of antigens presented to $CD8^+$ T cells and in altered susceptibility to specific infections. An alternative and not mutually exclusive possibility is that variants in *ERAP1* and *ERAP2* are maintained by selection due to their modulation of phenotypic traits not directly related to pathogen resistance. In addition to their being associated with ankylosing spondylitis, rs30187 has been identified as a susceptibility variant for essential hypertension (14), whereas the Asn392Lys polymorphism predisposes to pre-eclampsia (18). Therefore, additional selective pressures targeting genes involved in

blood pressure homeostasis and reproduction might have contributed to shaping the genetic variability at *ERAP1* and *ERAP2*.

MATERIALS AND METHODS

HapMap samples and sequencing

Human genomic DNA from HapMap subjects (20 individuals for each population) was obtained from the Coriell Institute for Medical Research. All analysed regions were PCR-amplified and directly sequenced; primer sequences are available upon request. PCR products were treated with ExoSAP-IT (USB Corporation, Cleveland, OH, 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.

Human subjects and genotyping

Blood samples were collected from 69 Italian ESN subjects. Inclusion criteria were a history of multiple unprotected sexual episodes for more than 4 years at the time of the enrolment, with at least three 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 ESN subjects are part of a well-characterized cohort of serodiscordant heterosexual couples that has been followed since 1997 (reviewed in 50). As for controls, 218 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, Italy. Written informed consent was obtained from all subjects.

All variants in the *ERAP1*/*ERAP2* genomic region were genotyped in the ESN and control samples through direct sequencing (primer sequences are available upon request). The polymorphic 32 bp deletion at the *CCR5* locus was typed using a PCR-based method. Specifically, PCR amplifications were performed with JumpStart AccuTaq LA DNA polymerase (Sigma-Aldrich) and primers flanking the 32 bp deletion (forward: 5′-TGGTGGCTGTGTTTGCGTCT-3′ and reverse: 5′-ATGACAAGCAGCGGCAGGAC-3′). The PCR products were electrophoretically separated on 3% agarose gels; the expected sizes for the deleted and non-deleted alleles are 137 and 169 bp, respectively.

Data retrieval and haplotype construction

Genotype data for 5 kb regions from 238 resequenced human genes were derived from the NIEHS SNPs program website (http://egp.gs.washington.edu). In particular, we selected genes that had been resequenced in populations of defined ethnicity including CEU, YRI and EAS (NIEHS panel 2).

Haplotypes were inferred using PHASE version 2.1 (51,52), a program for reconstructing haplotypes from unrelated genotype data through a Bayesian statistical method. Haplotypes for individuals resequenced in this study are available as supplemental material (Supplementary Material, Table S2).

LD analyses were performed using the Haploview (v. 4.1) (53), and blocks were identified through an algorithm implemented in the software. Data for LD analysis were derived from HapMap.

Statistical analysis

Tajima's D (19), Fu and Li's D^* and F^* (20) statistics, as well as diversity parameters θ_W (21) and π (22) were calculated using libsequence (54), a C++ class library providing an object-oriented framework for the analysis of molecular population genetic data. Calibrated coalescent simulations were performed using the cosi package (55) and its best-fit parameters for YRI, EU and EAS populations with 10 000 iterations. Coalescent simulations were conditioned on mutation rate and recombination rate. Estimates of the population recombination rate parameter ρ were obtained from diploid data by a composite likelihood method (56), with the use of the Web application MAXDIP (http://genapps.uchicago.edu/maxdip/). The maximum-likelihood-ratio HKA test was performed using the MLHKA software (24), as proposed previously (57). Briefly, 16 reference loci were randomly selected among NIEHS loci shorter than 20 kb that have been resequenced in the three populations; the only criterion was that Tajima's D did not suggest the action of natural selection (i.e. Tajima's D is higher than the 5th and lower than the 95th percentiles in the distribution of NIEHS genes). The reference set was accounted for by the following genes: *VNN3, PLA2G2D, MB, MAD2L2, HRAS, CYP17A1, ATOX1, BNIP3, CDC20, NGB, TUBA1, MT3, NUDT1, PRDX5, RETN* and *JUND*.

In all analyses, the chimpanzee sequence was used as the out-group.

Wittke-Thompson *et al*. (40) derived genotype frequencies for biallelic loci in cases and controls, assuming HWE in the general population. 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 40) 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 γ [dominant: $\beta = \gamma$; recessive: $\beta = 1$, $\gamma > 1$; additive: $\beta = (\gamma + 1)/2$, $\gamma > 1$ and multiplicative: $\beta = \mathrm{sqrt}(\gamma)$, $\gamma > 1$]. Given the different number of parameters in the general model, the Akaike Information Criterion 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.

All calculations were carried out in the R environment (58).

Tag SNPs along *ERAP2* were selected with Tagger (59) (http://www.broadinstitute.org/mpg/tagger/) using multimarker predictors to capture alleles with minor allele frequency >0.20 and an r^2 threshold of 0.8. As allowed by the software, we specified to include rs2549782 among tag

SNPs. The input region cover the whole gene 5 kb upstream the transcription start site. Association analyses were performed using PLINK (60).

Haplotype analysis and TMRCA calculation

Median-joining networks to infer haplotype genealogy were constructed using NETWORK 4.5 (29). The estimate of the TMRCA was obtained using a phylogeny-based approach implemented in NETWORK 4.5 using a mutation rate based on the number of fixed differences between chimpanzee and humans. Additional TMRCA estimates derived from the application of a maximum-likelihood coalescent method implemented in GENETREE (31,32). 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 MY ago (30) and with a generation time of 25 years. The migration matrix was derived from previous estimated migration rates (55). Using this μ and θ maximum-likelihood (θ_{ML}), we estimated the effective population size parameter (N_e). With these assumptions, the coalescence time, scaled in $2N_e$ units, was converted into years. For the coalescence process, 10^6 simulations were performed.

In order to obtain TMRCA estimates that take recombination into account, we used the InterRho program (35,36), kindly provided by Ying Wang. Genealogies are related through an ancestral recombination graph, and an algorithm based on a full-likelihood Bayesian Markov Chain Monte Carlo method estimates background mutation rates and hotspots. The times, in $4N_e$ units, were scaled into years by considering $N_e = 10400$ (37).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

We wish to thank Dr Ying Wang (University of Chicago), who provided the InterRho program and helped in the calculation of recombination rates and TMRCA estimates. Also, we are grateful to Dr Franca Rosa Guerini (Fondazione Don C. Gnocchi, IRCCS, Milan) for helpful advice on genotyping.

Conflict of Interest statement. None declared.

FUNDING

M.C. is supported by grants from Istituto Superiore di Sanita' 'Programma Nazionale di Ricerca sull' AIDS', the EMPRO and AVIP EC WP6 Projects, the nGIN EC WP7 Project, the Japan Health Science Foundation, 2008 Ricerca Finalizzata (Italian Ministry of Health), 2008 Ricerca Corrente (Italian Ministry of Health), Progetto FIRB RETI: Rete Italiana Chimica Farmaceutica CHEM-PROFARMA-NET (RBPR05NWWC) and Fondazione CARIPLO.

M.S. is a member of the Doctorate School in Molecular Medicine, University of Milan.

REFERENCES

1. Jensen, P.E. (2007) Recent advances in antigen processing and presentation. *Nat. Immunol.*, **8**, 1041–1048.
2. Saveanu, L., Carroll, O., Lindo, V., Del Val, M., Lopez, D., Lepelletier, Y., Greer, F., Schomburg, L., Fruci, D., Niedermann, G. *et al.* (2005) Concerted peptide trimming by human ERAP1 and ERAP2 aminopeptidase complexes in the endoplasmic reticulum. *Nat. Immunol.*, **6**, 689–697.
3. Hammer, G.E., Gonzalez, F., James, E., Nolla, H. and Shastri, N. (2007) In the absence of aminopeptidase ERAAP, MHC class I molecules present many unstable and highly immunogenic peptides. *Nat. Immunol.*, **8**, 101–108.
4. Hammer, G.E., Gonzalez, F., Champsaur, M., Cado, D. and Shastri, N. (2006) The aminopeptidase ERAAP shapes the peptide repertoire displayed by major histocompatibility complex class I molecules. *Nat. Immunol.*, **7**, 103–112.
5. Blanchard, N., Kanaseki, T., Escobar, H., Delebecque, F., Nagarajan, N.A., Reyes-Vargas, E., Crockett, D.K., Raulet, D.H., Delgado, J.C. and Shastri, N. (2010) Endoplasmic reticulum aminopeptidase associated with antigen processing defines the composition and structure of MHC class I peptide repertoire in normal and virus-infected cells. *J. Immunol.*, **184**, 3033–3042.
6. York, I.A., Brehm, M.A., Zendzian, S., Towne, C.F. and Rock, K.L. (2006) Endoplasmic reticulum aminopeptidase 1 (ERAP1) trims MHC class I-presented peptides *in vivo* and plays an important role in immunodominance. *Proc. Natl Acad. Sci. USA*, **103**, 9202–9207.
7. York, I.A., Chang, S.C., Saric, T., Keys, J.A., Favreau, J.M., Goldberg, A.L. and Rock, K.L. (2002) The ER aminopeptidase ERAP1 enhances or limits antigen presentation by trimming epitopes to 8–9 residues. *Nat. Immunol.*, **3**, 1177–1184.
8. Australo-Anglo-American Spondyloarthritis Consortium (TASC)Reveille, J.D., Sims, A.M., Danoy, P., Evans, D.M., Leo, P., Pointon, J.J., Jin, R., Zhou, X., Bradbury, L.A. *et al.* (2010) Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci. *Nat. Genet.*, **42**, 123–127.
9. Wellcome Trust Case Control Consortium, Australo-Anglo-American Spondylitis Consortium (TASC)Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I. *et al.* (2007) Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat. Genet.*, **39**, 1329–1337.
10. Prugnolle, F., Manica, A., Charpentier, M., Guegan, J.F., Guernier, V. and Balloux, F. (2005) Pathogen-driven selection and worldwide HLA class I diversity. *Curr. Biol.*, **15**, 1022–1027.
11. Apanius, V., Penn, D., Slev, P.R., Ruff, L.R. and Potts, W.K. (1997) The nature of selection on the major histocompatibility complex. *Crit. Rev. Immunol.*, **17**, 179–224.
12. Harvey, D., Pointon, J.J., Evans, D.M., Karaderi, T., Farrar, C., Appleton, L.H., Sturrock, R.D., Stone, M.A., Oppermann, U., Brown, M.A. *et al.* (2009) Investigating the genetic association between ERAP1 and ankylosing spondylitis. *Hum. Mol. Genet.*, **18**, 4204–4212.
13. Goto, Y., Hattori, A., Ishii, Y. and Tsujimoto, M. (2006) Reduced activity of the hypertension-associated Lys528Arg mutant of human adipocyte-derived leucine aminopeptidase (A-LAP)/ER-aminopeptidase-1. *FEBS Lett.*, **580**, 1833–1838.
14. Yamamoto, N., Nakayama, J., Yamakawa-Kobayashi, K., Hamaguchi, H., Miyazaki, R. and Arinami, T. (2002) Identification of 33 polymorphisms in the adipocyte-derived leucine aminopeptidase (ALAP) gene and possible association with hypertension. *Hum. Mutat.*, **19**, 251–257.
15. Taranta, A., Gianviti, A., Palma, A., De Luca, V., Mannucci, L., Procaccino, M.A., Ghiggeri, G.M., Caridi, G., Fruci, D., Ferracuti, S. *et al.* (2009) Genetic risk factors in typical haemolytic uraemic syndrome. *Nephrol. Dial. Transplant.*, **24**, 1851–1857.
16. Ouyang, C., Smith, D.D. and Krontiris, T.G. (2008) Evolutionary signatures of common human *cis*-regulatory haplotypes. *PLoS One*, **3**, e3362.
17. Tsui, F.W., Haroon, N., Reveille, J.D., Rahman, P., Chiu, B., Tsui, H.W. and Inman, R.D. (2010) Association of an ERAP1 ERAP2 haplotype with familial ankylosing spondylitis. *Ann. Rheum. Dis.*, **69**, 733–736.
18. Johnson, M.P., Roten, L.T., Dyer, T.D., East, C.E., Forsmo, S., Blangero, J., Brennecke, S.P., Austgulen, R. and Moses, E.K. (2009) The ERAP2 gene is associated with preeclampsia in Australian and Norwegian populations. *Hum. Genet.*, **126**, 655–666.

19. Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.
20. Fu, Y.X. and Li, W.H. (1993) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.
21. Watterson, G.A. (1975) On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.*, **7**, 256–276.
22. 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.
23. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.
24. Wright, S.I. and Charlesworth, B. (2004) The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. *Genetics*, **168**, 1071–1076.
25. Wright, S. (1950) Genetical structure of populations. *Nature*, **166**, 247–249.
26. Boweock, A.M., Kidd, J.R., Mountain, J.L., Hebert, J.M., Carotenuto, L., Kidd, K.K. and Cavalli-Sforza, L.L. (1991) Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. *Proc. Natl Acad. Sci. USA*, **88**, 839–843.
27. Akey, J.M., Zhang, G., Zhang, K., Jin, L. and Shriver, M.D. (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.*, **12**, 1805–1814.
28. Charlesworth, D. (2006) Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet.*, **2**, e64.
29. Bandelt, H.J., Forster, P. and Rohl, A. (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.*, **16**, 37–48.
30. Glazko, G.V. and Nei, M. (2003) Estimation of divergence times for major lineages of primate species. *Mol. Biol. Evol.*, **20**, 424–434.
31. Griffiths, R.C. and Tavare, S. (1995) Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Math. Biosci.*, **127**, 77–98.
32. Griffiths, R.C. and Tavare, S. (1994) Sampling theory for neutral alleles in a varying environment. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **344**, 403–410.
33. Templeton, A.R. (2005) Haplotype trees and modern human origins. *Am. J. Phys. Anthropol.*, **41** (suppl.), 33–59.
34. Schierup, M.H. and Hein, J. (2000) Recombination and the molecular clock. *Mol. Biol. Evol.*, **17**, 1578–1579.
35. Wang, Y. and Rannala, B. (2009) Population genomic inference of recombination rates and hotspots. *Proc. Natl Acad. Sci. USA*, **106**, 6215–6219.
36. Wang, Y. and Rannala, B. (2008) Bayesian inference of fine-scale recombination rates using population genomic data. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **363**, 3921–3930.
37. Charlesworth, B. (2009) Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.*, **10**, 195–205.
38. Tishkoff, S.A. and Verrelli, B.C. (2003) Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu. Rev. Genomics Hum. Genet.*, **4**, 293–340.
39. Samson, M., Libert, F., Doranz, B.J., Rucker, J., Liesnard, C., Farber, C.M., Saragosti, S., Lapoumeroulie, C., Cognaux, J., Forceille, C. *et al.* (1996) Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature*, **382**, 722–725.
40. Wittke-Thompson, J.K., Pluzhnikov, A. and Cox, N.J. (2005) Rational inferences about departures from Hardy–Weinberg equilibrium. *Am. J. Hum. Genet.*, **76**, 967–986.
41. Plummer, F.A., Ball, T.B., Kimani, J. and Fowke, K.R. (1999) Resistance to HIV-1 infection among highly exposed sex workers in Nairobi: what mediates protection and why does it develop? *Immunol. Lett.*, **66**, 27–34.
42. Fowke, K.R., Nagelkerke, N.J., Kimani, J., Simonsen, J.N., Anzala, A.O., Bwayo, J.J., MacDonald, K.S., Ngugi, E.N. and Plummer, F.A. (1996) Resistance to HIV-1 infection among persistently seronegative prostitutes in Nairobi, Kenya. *Lancet*, **348**, 1347–1351.
43. Andres, A.M., Hubisz, M.J., Indap, A., Torgerson, D.G., Degenhardt, J.D., Boyko, A.R., Gutenkunst, R.N., White, T.J., Green, E.D., Bustamante, C.D. *et al.* (2009) Targets of balancing selection in the human genome. *Mol. Biol. Evol.*, **26**, 2755–2764.
44. Draenert, R., Le Gall, S., Pfafferott, K.J., Leslie, A.J., Chetty, P., Brander, C., Holmes, E.C., Chang, S.C., Feeney, M.E., Addo, M.M. *et al.* (2004) Immune selection for altered antigen processing leads to cytotoxic T lymphocyte escape in chronic HIV-1 infection. *J. Exp. Med.*, **199**, 905–915.
45. Kaul, R., Dong, T., Plummer, F.A., Kimani, J., Rostron, T., Kiama, P., Njagi, E., Irungu, E., Farah, B., Oyugi, J. *et al.* (2001) CD8(+) lymphocytes respond to different HIV epitopes in seronegative and infected subjects. *J. Clin. Invest.*, **107**, 1303–1310.
46. Kanaseki, T. and Shastri, N. (2008) Endoplasmic reticulum aminopeptidase associated with antigen processing regulates quality of processed peptides presented by MHC class I molecules. *J. Immunol.*, **181**, 6275–6282.
47. Hughes, A.L. and Yeager, M. (1998) Natural selection at major histocompatibility complex loci of vertebrates. *Annu. Rev. Genet.*, **32**, 415–435.
48. Hughes, A.L. and Nei, M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, **335**, 167–170.
49. Takahata, N. and Nei, M. (1990) Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics*, **124**, 967–978.
50. Miyazawa, M., Lopalco, L., Mazzotta, F., Lo Caputo, S., Veas, F. and Clerici, M. and ESN Study Group (2009) The 'immunologic advantage' of HIV-exposed seronegative individuals. *AIDS*, **23**, 161–175.
51. Stephens, M., Smith, N.J. and Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, **68**, 978–989.
52. Stephens, M. and Scheet, P. (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.*, **76**, 449–462.
53. Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.
54. Thornton, K. (2003) Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics*, **19**, 2325–2327.
55. Schaffner, S.F., Foo, C., Gabriel, S., Reich, D., Daly, M.J. and Altshuler, D. (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.*, **15**, 1576–1583.
56. Hudson, R.R. (2001) Two-locus sampling distributions and their application. *Genetics*, **159**, 1805–1817.
57. Fumagalli, M., Cagliani, R., Pozzoli, U., Riva, S., Comi, G.P., Menozzi, G., Bresolin, N. and Sironi, M. (2009) Widespread balancing selection and pathogen-driven selection at blood group antigen genes. *Genome Res.*, **19**, 199–212.
58. R Development Core Team (2008) *R: A Language and Environment for Statistical Computing*. Vienna, Austria.
59. de Bakker, P.I., Yelensky, R., Pe'er, I., Gabriel, S.B., Daly, M.J. and Altshuler, D. (2005) Efficiency and power in genetic association studies. *Nat. Genet.*, **37**, 1217–1223.
60. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

3. CONCLUSIONS

There are several possible reasons for studying and identifying signatures of natural selection in the human genome. One event motivation is to gain a deeper understanding into the evolutionary history of our species. Another important driver for evolutionary biologists is based on the straightforward observation whereby natural selection acts on phenotypes and these, in turn, derive from functional variants. Therefore, the identification of natural selection signatures implies the identification of genes/gene regions carrying polymorphic functional variants in human populations. This is particularly relevant when the genes being analysed have been involved in human diseases or phenotypic traits of medical relevance. Moreover, the selective pressures underlying human adaptation are often environment-driven; therefore evolutionary biology approaches can provide information on how humans have adapted to their environment and how environmental shifts (as those carried along by agriculture first, industrialization and modernization next) may have resulted in disease susceptibility. Consequently, these approaches are particularly well-suited to the study of complex diseases/traits. The genetic architecture of complex traits envisages a situation whereby multiple loci, each with a small overall effect, contribute to the phenotype. Moreover, several complex diseases result from a combination of genetic and environmental effects.

Therefore, the questions addressed in this work can be summarized as follows:

1) Can population genetic approaches be applied to identify novel susceptibility alleles for complex traits (specifically, susceptibility to infection)?
2) Can population genetic approaches be used to infer the evolutionary history of disease alleles for human autoimmune diseases?

As for the first question, we have addressed the ability of geographic-explicit population genetic approaches to identify susceptibility alleles for virus-, protozoan- and helminth-borne infections at the genome-wide level [34-36]. These approaches have resulted in the retrieval of a number o putative susceptibility genes and SNPs. These latter should be regarded as causal or in linkage with the causal variant. Therefore, all these variants should now be analysed in detail and by means of classic case/control association studies with the aim of identifying the underlying selective pattern and the role in infection susceptibility. The case of *IFIH1* is emblematic in this respect: we identified a variant as being subjected to virus-driven selective pressure [34] and a subsequent analysis of the gene revealed a complex selective pattern with locally exerted selective pressures [37]. Similarly, the case of *ERAP2* indicates that the identification of gene regions subjected to natural selection can provide information on the location of functional variants and these, in turn, may be regarded as strong candidates to prioritize on case/control association studies. In the case of *ERAP2* we carried out one such study and verified that a nonsynonymous variant subjected to naural selection affects the natural resistance to HIV-1 infection [38].

With respect to the second question, we have shown that, while the evolutionary history of several interleukin/interleukin receptor genes has been dominated by a helminth-driven selective pressure, a subset of disease alleles for inflammatory bowel disease and celiac disease have evolved in response to non-helminthic

pathogens (i.e. viruses and bacteria), the underlying selective regime for some of them being balancing selection [39]. Therefore, our data for interleukin genes seem to support a conundrum of the hygiene hypothesis whereby disease alleles for autoimmune diseases result from adaptation to a pathogen-rich environment. Nonetheless, analysis of type 1 diabetes susceptibility alleles in *IFIH1* did not yield the same conclusion but rather suggested that these disease alleles have been neutrally evolving in human populations. Clearly, analysis of larger allele samples will be required to gain a comprehensive scenario of the evolutionary forces shaping the distribution of autoimmune susceptibility alleles. Even more so after the identification of a subset of alleles with an opposite risk profile for distinct autoimmune diseases (reviewed in [40]). In conclusion, our data suggest that the evolutionary patterns underlying the maintenance of autoimmune alleles in human populations are manifold, and possibly depend on either gene function, or diseases pathophysiology or other unknown factors. Analysis of many more variants and identification of causal polymorphisms (rather than linked genetic markers) will be required to gain a comprehensive scenario both from medical genetic perspective and from an evolutionary viewpoint.

4. 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. 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.
4. International HapMap Consortium, "A haplotype map of the human genome", *Nature,* Vol. 437, no. 7063, 2005, pp. 1299-1320.
5. International HapMap Consortium, "A second generation human haplotype map of over 3.1 million SNPs", *Nature,* Vol. 449, no. 7164, 2007, pp. 851-861.
6. Jablonski N. G., Chaplin G., "The evolution of human skin coloration", *J Hum Evol,* Vol. 39, no. 1, 2000, pp. 57-106.
7. Chaplin G., Jablonski N. G., "Vitamin D and the evolution of human depigmentation", *Am J Phys Anthropol,* Vol. 139, no. 4, 2009, pp. 451-461.
8. Di Rienzo A., Hudson R. R., "An evolutionary framework for common diseases: The ancestral-susceptibility model", *Trends Genet,* Vol. 21, no. 11, 2005, pp. 596-601.
9. Swallow D. M., "Genetics of lactase persistence and lactose intolerance", *Annu Rev Genet,* Vol. 37, 2003, pp. 197-219.
10. 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.
11. Tishkoff S. A., Reed F. A., Ranciaro A., Voight B. F., Babbitt C. C., Silverman J. S., Powell K., Mortensen H. M., Hirbo J. B., Osman M., et al, "Convergent adaptation of human lactase persistence in africa and europe", *Nat Genet,* Vol. 39, no. 1, 2007, pp. 31-40.
12. Kimura M. **The neutral theory of molecular evolution**, Cambridge University Press, 1983
13. Kimura M., "The rate of molecular evolution considered from the standpoint of population genetics", *Proc Natl Acad Sci U S A,* Vol. 63, no. 4, 1969, pp. 1181-1188.
14. Ewens W. J., "The sampling theory of selectively neutral alleles", *Theor Popul Biol,* Vol. 3, no. 1, 1972, pp. 87-112.
15. Smith J. M., Haigh J., "The hitch-hiking effect of a favourable gene", *Genet Res,* Vol. 23, no. 1, 1974, pp. 23-35.
16. 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.
17. Charlesworth D., "Balancing selection and its effects on sequences in nearby genome regions", *PLoS Genet,* Vol. 2, no. 4, 2006, pp. e64.
18. Tajima F., "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism", *Genetics,* Vol. 123, no. 3, 1989, pp. 585-595.
19. 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.
20. 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.
21. Fay J. C., Wu C. I., "Hitchhiking under positive darwinian selection", *Genetics,* Vol. 155, no. 3, 2000, pp. 1405-1413.
22. Fu Y. X., Li W. H., "Statistical tests of neutrality of mutations", *Genetics,* Vol. 133, no. 3, 1993, pp. 693-709.
23. 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.
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. 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.
26. Wright S., "Genetical structure of populations", *Nature,* Vol. 166, no. 4215, 1950, pp. 247-249.
27. 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.
28. Hancock A. M., Witonsky D. B., Gordon A. S., Eshel G., Pritchard J. K., Coop G., Di Rienzo A., "Adaptations to climate in candidate genes for common metabolic disorders", *PLoS Genet,* Vol. 4, no. 2, 2008, pp. e32.
29. Young J. H., Chang Y. P., Kim J. D., Chretien J. P., Klag M. J., Levine M. A., Ruff C. B., Wang N. Y., Chakravarti A., "Differential susceptibility to hypertension is due to selection during the out-of-africa expansion", *PLoS Genet,* Vol. 1, no. 6, 2005, pp. e82.
30. Hancock A. M., Witonsky D. B., Ehler E., Alkorta-Aranburu G., Beall C., Gebremedhin A., Sukernik R., Utermann G., Pritchard J., Coop G., et al, "Colloquium paper: Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency", *Proc Natl Acad Sci U S A,* Vol. 107 Suppl 2, 2010, pp. 8924-8930.
31. Prugnolle F., Manica A., Charpentier M., Guegan J. F., Guernier V., Balloux F., "Pathogen-driven selection and worldwide HLA class I diversity", *Curr Biol,* Vol. 15, no. 11, 2005, pp. 1022-1027.
32. Neel J. V., "Diabetes mellitus: A "thrifty" genotype rendered detrimental by "progress"?", *Am J Hum Genet,* Vol. 14, 1962, pp. 353-362.
33. Vander Molen J., Frisse L. M., Fullerton S. M., Qian Y., Del Bosque-Plata L., Hudson R. R., Di Rienzo A., "Population genetics of CAPN10 and GPR35: Implications for the evolution of type 2 diabetes variants", *Am J Hum Genet,* Vol. 76, no. 4, 2005, pp. 548-560.
34. 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.
35. 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.

36. 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,.
37. Fumagalli M., Cagliani R., Riva S., Pozzoli U., Biasin M., Piacentini L., Comi G. P., Bresolin N., Clerici M., Sironi M., "Population genetics of IFIH1: Ancient population structure, local selection and implications for susceptibility to type 1 diabetes", *Mol Biol Evol,* 2010,.
38. Cagliani R., Riva S., Biasin M., Fumagalli M., Pozzoli U., Lo Caputo S., Mazzotta F., Piacentini L., Bresolin N., Clerici M., et al, "Genetic diversity at endoplasmic reticulum aminopeptidases is maintained by balancing selection and is associated with natural resistance to HIV-1 infection", *Hum Mol Genet,* 2010,.
39. 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.
40. Sironi M., Clerici M., "The hygiene hypothesis: An evolutionary perspective", *Microbes Infect,* Vol. 12, no. 6, 2010, pp. 421-427.

APPENDIX

Additional manuscripts published during the PhD.

Cagliani R, Fumagalli M, Biasin M, Piacentini L, Riva S, Pozzoli U, Bonaglia MC, Bresolin N, Clerici M, Sironi M. Long-term balancing selection maintains trans-specific polymorphisms in the human *TRIM5* gene. *Hum Genet*. 2010 Sep 2.

Torri F, Akelai A, Lupoli S, Sironi M, Amann-Zalcenstein D, Fumagalli M, Dal Fiume C, Ben-Asher E, Kanyas K, Cagliani R, Cozzi P, Trombetti G, Strik Lievers L, Salvi E, Orro A, Beckmann JS, Lancet D, Kohn Y, Milanesi L, Ebstein RB, Lerer B, Macciardi F. Fine mapping of AHI1 as a schizophrenia susceptibility gene: from association to evolutionary evidence. *FASEB J*. 2010 Aug;24(8):3066-82.

Cagliani R, Fumagalli M, Riva S, Pozzoli U, Fracassetti M, Bresolin N, Comi GP, Sironi M. Polymorphisms in the *CPB2* gene are maintained by balancing selection and result in haplotype-preferential splicing of exon 7. *Mol Biol Evol*. 2010 Aug;27(8):1945-54.

Cagliani R, Fumagalli M, Riva S, Pozzoli U, Comi GP, Bresolin N, Sironi M. Genetic variability in the ACE gene region surrounding the Alu I/D polymorphism is maintained by balancing selection in human populations. *Pharmacogenet Genomics*. 2010 Feb;20(2):131-4.

Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi GP, Menozzi G, Bresolin N, Sironi M. A population genetics study of the familial Mediterranean fever gene: evidence of balancing selection under an overdominance regime. *Genes Immun.* 2009 Dec;10(8):678-86.

Cagliani R, Fumagalli M, Pozzoli U, Riva S, Cereda M, Comi GP, Pattini L, Bresolin N, Sironi M. A complex selection signature at the human AVPR1B gene. *BMC Evol Biol*. 2009 Jun 1;9:123.

Cagliani R, Fumagalli M, Pozzoli U, Riva S, Comi GP, Torri F, Macciardi F, Bresolin N, Sironi M. Diverse evolutionary histories for beta-adrenoreceptor genes in humans. *Am J Hum Genet*. 2009 Jul;85(1):64-75. Epub 2009 Jul 2.

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

Crimella C, Arnoldi A, Crippa F, Mostacciuolo ML, Boaretto F, Sironi M, D'Angelo MG, Manzoni S, Piccinini L, Turconi AC, Toscano A, Musumeci O, Benedetti S, Fazio R, Bresolin N, Daga A, Martinuzzi A, Bassi MT. Point mutations and a large intragenic deletion in SPG11 in complicated spastic paraplegia without thin corpus callosum. *J Med Genet*. 2009