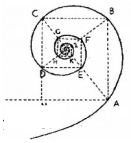
UNIVERSITÀ DEGLI STUDI DI MILANO



SCUOLA DI DOTTORATO IN MEDICINA MOLECOLARE

CICLO XXIX
Anno Accademico 2015/2016

TESI DI DOTTORATO DI RICERCA

BIO18

Evolutionary analyses provide insight into host-pathogen interactions and diet-related adaptations

Dottorando : Chiara PONTREMOLI
Matricola N° R10559

TUTORE: Prof. ssa Mara BIASIN
CO-TUTORE: Dott. ssa Manuela SIRONI

DIRETTORE DEL DOTTORATO: Ch.mo Mario CLERICI

SOMMARIO

INTRODUZIONE

La diversità genetica gioca un ruolo importante nella sopravvivenza e nei processi adattativi di tutte le specie. Quando l'ambiente cambia (es: condizioni climatiche, agenti patogeni, disponibilità di cibo), la popolazione è soggetta ad una pressione selettiva: col passare del tempo, i tratti variabili più vantaggiosi per l'adattamento vengono selezionati mediante processi di selezione naturale e trasmessi da una generazione alla successiva.

La diversità genetica è generata dalla combinazione di differenti processi evolutivi, che includono le mutazioni, la deriva genetica, le migrazioni e la selezione naturale.

La selezione naturale lascia una traccia molecolare caratteristica all'interno dei genomi, che può essere rilevata con l'applicazione di diversi metodi evolutivi. Questi metodi possono essere suddivisi in due categorie: test che ricercano segnali di selezione inter-specie (mediante il confronto tra individui di specie diverse), e test che, invece, si focalizzano su confronti intra-specie (cioè all'interno della stessa specie).

Durante il periodo di dottorato ho condotto studi evolutivi su geni soggetti a differenti pressioni selettive.

In un primo studio ho analizzato la storia evolutiva di geni coinvolti nell'adattamento alla dieta. I cambiamenti nella disponibilità di cibo e nella dieta avvenuti nel corso della storia, hanno creato forti pressioni selettive in diversi processi biologici. Ad esempio, nell'uomo, la rivoluzione agricola ha favorito il consumo di carboidrati.

Sfruttando la disponibilità di sequenze genomiche di organismi diversi, ho eseguito un'analisi evolutiva su geni che sono coinvolti in diverse fasi dell' assorbimento e della digestione dei carboidrati nell'orletto a spazzola della mucosa intestinale e ho usato dei dati di risequenziamento di campioni di DNA antichi per determinare quando si sono sviluppati gli alleli adattativi.

In una seconda serie di studi, invece, ho effettuato delle analisi evolutive per studiare l'interazione tra le proteine dell'ospite e quelle patogene.

Le molecole che partecipano alla risposta immunitaria sono coinvolte in una costante "corsa alle armi" contro i patogeni e contengono le tracce molecolari di questo conflitto. Ho quindi analizzato la storia evolutiva dei geni della difesa immunitaria (ad esempio geni di rilevamento antivirale, di induzione della risposta a interferone ed effettori antivirali), del sistema del complemento (che è un effettore dell'immunità innata) e di proteine batteriche che interagiscono con il complemento.

Tuttavia, i conflitti genetici con i patogeni non coinvolgono solo geni con una specifica funzione nella difesa immunitaria, ma anche molecole implicate in processi omeostatici cruciali.

Tra queste ci sono: (i) il trasportatore degli steroli, NPC1, utilizzato dai filovirus

come recettore per l'ingresso nella cellula dell'ospite; (ii) la basigina, una proteina multifunzionale con un ruolo nel funzionamento dei trofoblasti e nella spermatogenesi che viene utilizzata dal Plasmodium falciparum per l'invasione eritrocitaria.

Ho quindi studiato la storia evolutiva di questi geni e dei loro interattori virali e microbici.

SCOPO DEL LAVORO

Lo scopo del mio progetto è quello di utilizzare analisi evolutive per indagare e descrivere eventi adattativi in geni soggetti a diverse pressioni selettive, sia nei mammiferi sia in virus, protozoi e batteri. In particolare, ho condotto analisi inter-specie e analisi di genetica di popolazione-filogenetiche, con l'obiettivo di individuare la selezione positiva che ha agito per lunghi tempi evolutivi.

MATERIALI E METODI

Le sequenze di cDNA dei geni di interesse di numerose specie di mammiferi e patogeni sono state ottenute da banche dati pubbliche (NCBI, UCSC o ENSEMBL) oppure mediante sequenziamento diretto.

Le analisi condotte sui DNA antichi sono state possibili grazie a recenti pubblicazioni di dati relativi al sequenziamento del genoma di ominidi tra cui un uomo di Neanderthal, un uomo di Denisova, un cacciatore-raccoglitore europeo del Mesolitico proveniente dalla Spagna e un ominide siberiano del Paleolitico.

Le sequenze sono state allineate con software specifici (es: RevTrans 2.0, PRANK e GUIDANCE).

Dato che la ricombinazione genica può essere scambiata per selezione positiva, tutti gli allineamenti sono stati sottoposti a screening per la ricerca di breakpoints di ricombinazione utilizzando un software specifico (GARD).

I segnali di selezione positiva sono stati determinati con il software codeml; inoltre i siti selezionati positivamente sono stati identificati con due programmi differenti: BEB e MEME. Per valutare eventi di selezione episodica su specifici rami della filogenia sono stati utilizzati Bs-Rel e branch-site test del software codeml.

Le analisi evolutive nell'uomo, nello scimpanzé e nel gorilla sono state eseguite utilizzando un approccio di genetica di popolazione-filogenetico. Per valutare il significato funzionale dei residui selezionati positivamente, sono state condotte analisi di strutture 3D.

Grazie a modelli statistici di regressione logistica, sono stati condotti studi di associazione genetica tra le varianti alleliche selezionate ed il fenotipo di resistenza all'infezione da HIV ed è stata valutata l'espressione genica in risposta alla stimolazione con interferone alpha (IFN-α). In particolare sono state selezionate tre coorti europee di individui sieronegativi esposti a HIV-1 (HESN) che differiscono per origine geografica ed esposizione al virus. Le varianti di interesse sono state genotipizzate mediante sequenziamento diretto del DNA di questi soggetti.

I risultati ottenuti per ciascuna delle tre casistiche analizzate, sono stati combinati

mediante tecniche di meta-analisi con PLINK.

Il saggio di infezione da HIV è stato condotto separando in coltura le cellule mononucleari del sangue periferico (PBMC) dei soggetti HESN. Dopo averne valutato la vitalità, è stato fornito l'input virale HIV-1_{Ba-L} p24. La titolazione virale è stata effettuata saggiando l'antigene virale p24 nel terreno di coltura, rilasciato dopo 7 giorni di incubazione, mediante saggio ELISA.

Per la stimolazione con IFN-α, le PBMC isolate da controlli sani sono state piastrate in terreno di coltura con o senza l'aggiunta di IFN-α. Dopodichè, l'RNA è stato estratto dalle cellule e retro-trascritto in cDNA. La quantificazione del cDNA è stata eseguita tramite Real-Time PCR e i risultati sono stati espressi come rapporto tra il gene di interesse e un gene housekeeping (GAPDH).

RISULTATI

Le analisi evolutive possono fornire informazioni estremamente rilevanti non solo sulla storia evolutiva del nostro genoma, ma anche riguardanti la presenza e la localizzazione di varianti genetiche funzionali legate alla diversità fenotipica e alla salute umana.

Nello studio sui geni dell'orletto a spazzola della mucosa intestinale, i risultati indicano che la pressione evolutiva data dall'adattamento a diete specializzate ha portato ad una selezione positiva diffusa nei mammiferi e nelle popolazioni umane. Inoltre, ho osservato che gli alleli moderni che sono stati selezionati positivamente, sono insorti prima dell'avvento dell'agricoltura.

Nella seconda serie di analisi ho dimostrato che anche la risposta immunitaria ha esercitato una forte pressione evolutiva: infatti ho trovato tracce di selezione positiva pervasiva in residui e regioni proteiche che esercitano un ruolo funzionale (ad esempio porzioni regolatorie o che conferiscono attività antivirale). È interessante notare che spesso la selezione naturale ha come target dei residui che si trovano nella stessa posizione spaziale in proteine diverse. Questo vale ad esempio per i siti trovati in OAS1, OAS2 e MB21D1, che hanno mostrato un'evoluzione in parallelo.

Nelle analisi di proteine virali e microbiche e dei loro interattori nelle cellule ospite, la maggior parte dei siti selezionati positivamente si trova all'interno della regione all'interfaccia tra ospite e patogeno. Questo rispecchia l'aspettativa di uno scenario di conflitto genetico, per cui i geni dell'ospite e del patogeno si evolvono per tentare di evitare l'interazione nel primo caso, e di rafforzarla nel secondo. Alcuni dei siti selezionati positivamente all'interfaccia di interazione hanno un ruolo nell'affinità di legame tra ospite e patogeno, altri molto probabilmente hanno contribuito al cambiamento di tropismo dell'ospite. Alcuni residui virali, invece, si trovano in determinanti antigenici. Dato che le strategie di trattamento farmacologico più promettenti comprendono quelle basate sulle combinazioni di anticorpi, questi risultati dovrebbero far riflettere sulle problematiche relative alla loro efficacia a lungo termine.

CONCLUSIONI

Questi lavori evidenziano l'importanza delle analisi evolutive. In particolare, ho

mostrato che gli studi evolutivi possono: (i) fornire informazioni riguardo agli eventi adattativi che hanno plasmato i tratti specifici umani, (ii) identificare regioni funzionali e siti che evolvono sotto selezione positiva, (iii) prevedere superfici di interazione ospite-patogeno a livello del singolo aminoacido, (iv) fornire informazioni sulle determinanti molecolari alla base della specie-specificità della suscettibilità alle infezioni e chiarire la risposta differenziale a molecole naturali o sintetiche, (v) contribuire a individuare i serbatoi più probabili per i patogeni zoonotici, (vi) spiegare le variazioni di tropismo del patogeno e (vii) fornire informazioni su possibili bersagli terapeutici (che riguardano ad esempio l'efficacia dei trattamenti a lungo termine basati sulle combinazioni di anticorpi).

SUMMARY

INTRODUCTION

Genetic diversity plays an important role in the survival and adaptability of all species. When a population environment (meaning, for instance, climatic conditions, pathogens, and food availability) changes, the population is subject to a selective pressure. Variation in the population gene pool provides variable traits which can be selected for, via natural selection, leading to an adaptive change to survive. Genetic diversity is generated by a combination of different evolutionary processes such as mutation, genetic drift, migration and natural selection.

Natural selection leaves a distinctive molecular signature in genomes. Such molecular signatures can be detected with evolutionary tests that can be divided into those that search for selection at the inter-species level (e.g., human versus primates and mammals) and those that focus on within-species data (e.g., among human populations). In my work, I used evolutionary studies to analyze genes under different selective pressures.

In a first study, I investigate the evolutionary history of genes possibly involved in diet adaptation. Changes in food availability and diet likely created strong selective pressures on multiple biological processes. In humans, the agricultural revolution favored carbohydrate consumption. I exploit the availability of genome sequences from different organisms, together with resequencing data of ancient DNA samples to perform a comprehensive evolutionary analysis of genes involved in sugar absorption/digestion at the brush-border and to test when adaptive alleles arose.

In a second set of studies, I use evolutionary analyses to investigate the interaction between host proteins and viral/protozoan/bacterial protein products. Molecules that participate in immune response are expected to be engaged in a constant arms-race with pathogens and to harbour the molecular signatures of such a conflict. I thus investigate the evolutionary history of genes involved in immune defense, such as antiviral sensing proteins, genes with IFN-inducible properties and antiviral effectors. I also investigate the evolutionary history of the complement system, an innate immunity effector, and of bacterial-encoded complement-interacting proteins.

Nevertheless, not only genes with specific defense function, but also molecules involved in central homeostatic processes may be engaged in genetic conflicts with pathogens. This is exemplified by (i) the sterol transporter NPC1, used as receptor for filoviruses entry and (ii) basigin, a multifunctional protein with a role in trophoblast function and in spermatogenesis which is used for erythrocyte invasion by Plasmodium falciparum. I therefore study the evolutionary history of these genes and of their viral/microbial interactors.

AIM OF THE WORK

The purpose of my project is to use evolutionary analyses to investigate and describe adaptive events at candidate genes subject to different selective pressures, in species ranging from mammals to viruses/protozoa/bacteria. In

particular, I focus on inter-species and population genetics-phylogenetics analyses, with the aim to detect positive selection acting over long evolutionary timescales.

MATERIALS AND METHODS

Mammalian and pathogen coding gene sequences were retrieved from public databases (Ensembl, UCSC and NCBI) or obtained by direct sequencing.
Information from the Neandertal and Denisova high-coverage genomes, as well as from a hunter-gatherer Mesolithic European from Spain and a Paleolithic Siberian was derived from previous works.
Sequences were aligned using RevTrans 2.0 utility, PRANK, and unreliably aligned codons were then filtered using GUIDANCE. Since recombination can be mistaken as positive selection, all alignments were screened for the presence of recombination breakpoints using a specific software (GARD). To detect selection, codeml models were fitted to the data using different models of equilibrium codon frequencies. Sites under selection were identified using BEB and MEME. Branches and sites subject to episodic positive selection were identified using Bs-Rel or the branch-site tests from the codeml software.
Evolutionary analysis in the human, chimpanzee and gorilla lineages was performed using a population genetics-phylogenetics approach.
Analysis of 3D structures was used to infer the functional significance of positively selected sites.
In order to assess the role of selected variants in HIV-1 susceptibility, and to evaluate gene expression in response to interferon alpha (IFN-α), genetic association analyses, in vitro HIV-1 infection and IFN-α stimulation assays were also performed.
Genotyping was carried out on three independent European cohorts of HIV-1 exposed seronegative individuals (HESN) with different geographic origin and distinct exposure route. Variants were genotyped through direct sequencing. Genetic association analyses were performed by logistic regression and results from the three cohorts were combined using a random-effect meta-analysis; all analyses were performed using PLINK.
For the HIV infection assay, PBMC from HESN subjects were separated on lymphocyte separation medium. After viability assessment, they were resuspended in a medium containing HIV-$1_{Ba\text{-}L}$ p24 viral input. After 7 days, supernatants were collected for p24 antigen ELISA analyses and absolute levels of p24 were measured.
For IFN-α stimulation, freshly isolated PBMC from healthy controls were incubated with medium alone or with IFN-α. Cultured PBMC RNA was extracted and then reverse transcribed into first-strand cDNA. cDNA quantification was performed by a Real-Time PCR strategy. Results were presented as ratios between the target gene and the GAPDH housekeeping mRNA.

RESULTS

Evolutionary analysis can provide extremely relevant information not only on the evolutionary history of our genome, but also on the presence and location of functional genetic variants especially in relation to phenotypic diversity and,

ultimately, human health.

In the study of the brush border genes, results indicated pervasive selection in mammals and human populations, reflecting specific adaptation to specialized diets. Furthermore, I found that positively selected modern alleles predate the emergence of agriculture.

In the second set of analyses I found that pervasive positive selection is driven by pressure related to immune response. Selection acted on functionally relevant protein regions (e.g.: regulatory regions, regions which confer antiviral activity) and residues. Interestingly, natural selection often targeted residues located in the same spatial position in different proteins. This is for example the case of sites detected in OAS1, OAS2 and MB21D1, which revealed parallel evolution.

In the analyses of viral/protozoan/bacterial proteins and their host interactors, I found most of the positively selected sites within the region at the host-pathogen interface. These results epitomize the expectation under a genetic conflict scenario, whereby the host and the pathogen genes evolve within binding avoidance-binding seeking dynamics. Part of these sites had a role in host-pathogen binding affinity, some other adaptive changes most likely contributed to the shift to human hosts, and still other residues found to evolve under positive selection in viruses reside in antigenic determinants. Because antibody combinations are the most promising treatment strategies, these findings should pose a serious concern to their effectiveness in the long-term.

CONCLUSIONS

These works highlight the importance of evolutionary analysis. Specifically, I show that evolutionary studies can (i) provide information on the past adaptive events that shaped human-specific traits, (ii) identify functional regions and sites evolving under positive selection, (iii) predict host-pathogen interaction surfaces at the single amino acid resolution, (iv)provide valuable information on the molecular determinant underlying species-specific infection susceptibility and clarify the differential response to natural or synthetic molecules, (v) help identify the most likely reservoirs for zoonotic pathogens, (vi) explain changes in pathogen tropism and (vii) provide information on possible therapeutic targets (e.g. effectiveness of long-term treatment based on antibody combinations).

I N D E X

SYMBOL LIST

dN: the observed number of nonsynonymous substitutions per nonsynonymous site

dS: the observed number of synonymous substitutions per synonymous site

ω: dN/dS

LD: Linkage Disequilibrium

LRT: Likelihood Ratio Test

SNP: Single Nucleotide Polymorphism

MAF: Minor Allele Frequency

DAF: Derived Allele Frequency

F_{ST}: Fixation Index

DIND: Derived Intra-allelic Nucleotide Diversity

1. INTRODUCTION

1.1 Genetic variability and evolution

As humans and other organisms moved to inhabit every part of the world, they were exposed to new environments, including different climatic conditions, pathogen species and food sources, to which they were forced to adapt [1, 2]. Variation is essential to species survival and adaptation during evolution: if the environment changes, species that have a higher genetic variability will be better able to evolve to adapt. Variants that confer an advantage for a species will be selected for by natural selection.

Selection operates at the level of the phenotype. Within and between species, there are multiple potential sources of phenotypic variation, and each of these sources reflects a different underlying cause. The particular source of phenotypic variation determines whether that trait has the ability to respond to environmental changes. Researchers are therefore highly interested in determining the relative importance of both the genetic and the environmental factors that lead to particular traits for several reasons: 1) current genetic diversity contains information about the size and movements of past populations (referred to as groups of organisms of one species), and on the history of adaptation to environmental changes; 2) information can be used to predict the evolutionary dynamics of an entire population with respect to the specific trait [3-5]; 3) this knowledge has biomedical and therapeutic implications, since it may be translated into novel treatment strategies in term of molecular targets or drug development.

Genetic diversity is shaped by a number of different processes which include mutation, recombination, genetic drift, migration and natural selection.

1.1.1 Mutation and recombination

Mutation is the only process generating new alleles, providing the raw material on which evolution can act. A mutation is a heritable change in the DNA sequence. If a mutation occurs in reproductive cells, it may also be passed from parent to offspring. Mutations are essential factors in evolution, as they provide the variation that enables organisms to change and adapt to their environment when selective pressure is applied. Based on their effects on the ability of the individual to survive and reproduce, mutations can be divided into three broad categories: deleterious, that decrease this ability, neutral, that do not have an effect and are fixed or lost by drift, and advantageous, that increase the ability of the organism to survive and reproduce [6].

New combinations of mutations may then arise through recombination in meiosis: they occur as a part of sexual reproduction and enhance the ability of organisms to adapt to their environment by combining advantageous alleles at different loci. Nevertheless, recombination can also generate new combinations of alleles on the same DNA molecule, known as haplotype, and in this way increase haplotype diversity. Consequently, recombination is also capable of breaking up advantageous allelic combinations.

1.1.2 Genetic drift and migrations

Genetic drift is a process in which allele frequencies within a species change by chance alone as a result of different possible scenarios. These changes are not driven by environmental or adaptive pressures and can either increase or decrease over time. Typically, genetic drift occurs in small populations, where infrequently occurring alleles face a greater chance of being lost. Common examples of genetic drift are bottlenecks and the founder effect. A population bottleneck arises when a significant number of individuals in a population dies or is otherwise prevented from breeding,

resulting in a drastic decrease in the size of the population, whereby the level of genetic variation is extremely limited. The founder effect, instead, involves migration, where a small group of organisms may migrate from a large continental population and establish a colony in a new location. In any case, the final effect of genetic drift is to reduce genetic variation within populations and increase divergence among populations [7].

Migration is the movement of individuals of the same population from one geographic location to another. Gene flow is the outcome of a migration, which is the exchange of genes between populations. Genetic variation is added to new populations and increases genetic differences within the recipient population. At the same time, migration prevents subpopulation from becoming too different: in the absence of migration there will be differing frequencies of common alleles among local populations and some local populations will posses certain rare alleles not found in others; migration leads to mixing of the gene pool. Hence, migration increases genetic variation within population and reduces divergence between populations.

1.1.2.1 Modern human evolutionary history

Anatomically modern humans appeared in East Africa about 200k years ago, spread out from sub-Saharan Africa approximately 100k years ago, and subsequently colonized Europe and East Asia.

A study of Cann and colleagues [8] indicated that there is a strong correlation between geographic distance among populations and genetic variability within population; indeed populations closer to Africa show more variability than population further from Africa, reflecting the route of migrations of ancient humans. The allele pool of populations reflects this geographic scenario: in fact, most of the alleles/haplotypes of non-African

populations are a subset of the African ones. This observation is crucial in order to identify genetic adaptation, because it must be considered when evaluating whether a specific variant is influenced by the action of natural selection. Demographic events influence variability in all genes in the same way, whereas selection specifically targets defined regions; so deviation from the general behavior of genetic variation can be an indication of the action of natural selection. It should also be considered that, during their evolutionary history, humans have adapted to the environment, and this adaptation is driven by different forces, with major effects generally ascribed to pathogens, climate and diet.

1.1.3 Natural selection

Another process that generates variability is natural selection, which is the tendency for traits to increase or decrease in frequency in a population, depending on the reproductive success (fitness) they confer [9]. Traits with a greater level of fitness contribute more to the gene pool of subsequent generations. Selection represents the action of environmental factors on a particular phenotype and genotype through selective pressures and can occur at any stage, from the formation of the genotype at fertilization to the end of the reproductive period.

There are three major forms of natural selection: purifying or negative selection, positive or adaptive Darwinian selection and balancing selection. Purifying selection is the most pervasive form of natural selection acting on genomes. It reduces the frequency of deleterious alleles in a population. New mutations often have detrimental effect on biological fitness and purifying selection reduces the number of new mutation in the gene pool. The main consequence of negative selection is a local reduction of diversity

and an increase of rare alleles when selection is not strong enough to completely eliminate deleterious variants from the population (that is, weak negative selection) [10] (Fig 1A).

Balancing selection is a form of natural selection whereby multiple alleles are maintained at an appreciable frequency within the gene pool. It principally acts by two mechanisms: heterozygote advantage (or overdominance) and frequency-dependent selection. The first refers to a situation in which heterozygotes show a higher level of fitness than homozygotes. This leads to the maintenance of two or multiple alleles in a population at a given locus (Fig 1B). Frequency-dependent selection occurs when the fitness of a genetically-determined phenotype is dependent on its frequency relative to other phenotypes in a given population. For example, in negative frequency-dependent selection, the fitness of a phenotype decreases as it becomes more common. Whatever the mechanisms, balancing selection leads to an excess of intermediate-frequency variants, which will results in increased levels of diversity. Finally, positive Darwinian selection refers to selection acting upon newly advantageous mutations. Under positive selection, advantageous alleles rapidly achieve high frequencies within the population. This occurs at a rate much faster than that of a neutral allele. When an advantageous mutation increases in frequency in the population as a result of positive selection, linked neutral variation will be dragged along with it in a process known as genetic hitchhiking. As a consequence, variation that is not associated with the selected allele is eliminated, resulting in a selective sweep that leads to an overall reduction of genetic diversity around the selected site (Fig 1C). Additional features include a skew in the distribution of allele frequencies towards an excess of rare and high-frequency derived alleles, and a transitory increase in the strength of linkage disequilibrium associated with

the selected allele.

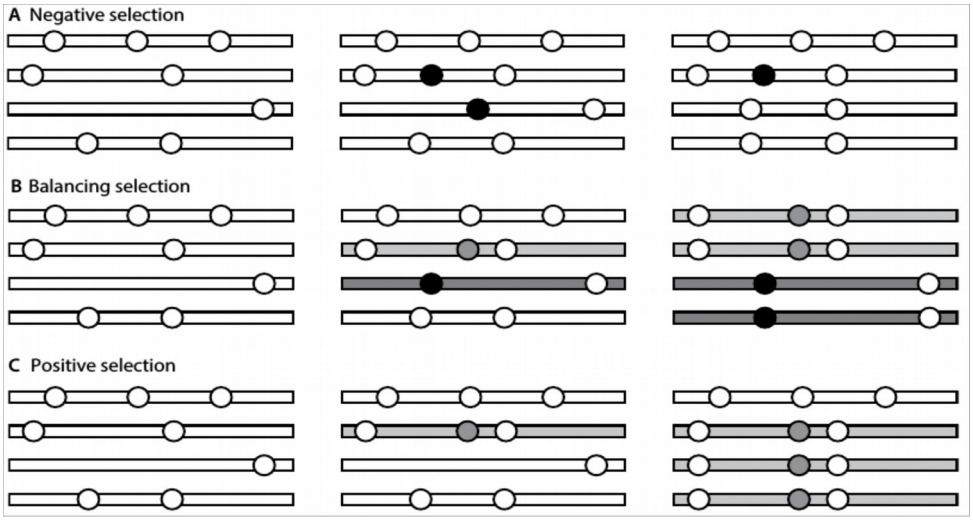


Figure 1: Rows A, B and C illustrate three patterns of natural selection. The columns represent three generations: the first column shows the starting group of four individuals looking at the same chromosome in each; the second column shows the first generation with mutations, and the third column shows the final outcome for the chromosomes in three different patterns of natural selection. Each circle represent a polymorphism, within a haplotype. White circles represent mutations under neutrality, black circle deleterious mutations and gray circles indicate advantageous ones. Pattern A illustrates genetic polymorphisms under negative selection. Deleterious mutations arise (black dot) and they can be removed immediately (if severely deleterious, e.g. line 3 in column 3) or kept at low frequencies (if weakly deleterious, e.g., line 2 in column 3). Linked neutral polymorphism will also disappear (or be kept at low frequencies, e.g. line 3 in column 3). Pattern B illustrates balancing selection. Two new alleles are shown (shaded and black circles) and, if they confer advantage in the heterozygous state, they will increase to intermediate frequencies. Linked neutral polymorphism will also increase to intermediate frequencies. Pattern C illustrates genetic polymorphism under positive selection. When a new advantageous mutation arises (shaded circle in line 2, column 2), the allele increases in frequency (in the population) along with linked neutral polymorphisms (lines 3 and 4 in column 3, which now resemble line 2, column 2) (Image adapted from [11]).

Much research in recent years has focused on the development of genomic methods to identify positive selection: this is because positive selection is considered to be the primary mechanism of adaptation, so understanding the traits (and genes underlying them) that have undergone positive selection during evolution can provide insight into the events that have shaped the genomic variability of the species, as well as into the diseases

that continue to plague today.

1.2 Detection of natural selection

Based on types of data they utilize, the timescale they focused on and the type of selective signature they identify, methods to detect natural selection in the genome may be distinguished into inter- and intra-species analyses.

Inter-species analysis are used to identify selective events that took place within the deep past and that reflect macroevolutionary trends that occur as a result of selection between species. They are based on the comparison of homologous traits of DNA sequences among related species. Inter-specific analysis are based on few important intuitions: (i) closely related organisms share a very similar gene pool, which they have inherited from a common ancestor; conserved sequences between species indicate that a particular sequence may have been maintained by evolution, more likely for their basic role in proteins function and stability. Mutations in a highly conserved region usually lead to a non-viable life form of the protein and are therefore quickly removed from the gene pool. Mutations that confer an advantage for organism survival in a particular environment are instead maintained in the genome. Indeed, inter-specific analysis search for rare variable regions among huge conserved ones (ii) traits that are conserved across many clades of a phylogeny but that show extreme differentiation in one or a few lineages are thought to be candidates for selection.

Intra-species analysis, instead, are used to identify more recent selective events (e.g. lineage specific selection in human populations) comparing genetic variation within sequences of the same species. These analyses search for genes/gene regions subject to different form of selection: positive selection, which refers to the case in which DNA variants has a selective advantage over others, and consequently rises in frequency; and balancing selection, which refers to selective regimes that increase genetic

variation within a species. These two kind of selection leave different signatures on genomic regions. These signatures can be searched for and exploited to identify selected variants: (i) in a selective sweep, a genetic variant reaches high frequencies together with nearby linked variants (the result is a high frequency derived allele). From this homogenous background, new alleles arise but are initially at low frequency (surplus of rare alleles); (ii) selective sweeps bring a genetic region to high frequencies in a population, including the causal variant and its neighbors. The associations between these alleles define a haplotype, which persists in the population until recombination breaks these associations (the result is a long unbreak haplotype); (iii) selection acting on an allele in one population but not in another creates a marked difference in the frequency of that allele between the two populations. This effect of differentiation stands out against the differentiation between populations with respect to neutral (i.e., non selected) alleles [1].

1.2.1 Detection of natural selection at the inter-species level

Each form of natural selection leaves distinctive “signatures” or patterns of genetic variation in DNA sequences. Methods to detect selection at the inter-species level typically rely on the comparison of homologous traits or sequence among related species. These approaches take as an input the alignment and the site-by-site analysis of orthologous coding sequences. For each site it will be determined which among all substitution would lead to an amino-acid replacement (non-synonymous substitution) or not (synonymous substitution). The number of nonsynonymous differences per nonsynonymous site (dN) and the number of synonymous differences per synonymous site (dS) are then calculated. Under neutral evolution, the rate at which aminoacid replacements accumulate is expected to be comparable to the rate of synonymous changes and, therefore, dN/dS (ω)

should be equal to 1. Nonetheless, in most species the majority of the aminoacid replacements are deleterious and eliminated by negative selection: in this case dN/dS<1. Conversely, a specific selective pressure may favor aminoacid replacements: in this case dN/dS will be higher than 1, a hallmark of positive selection (Fig 2).

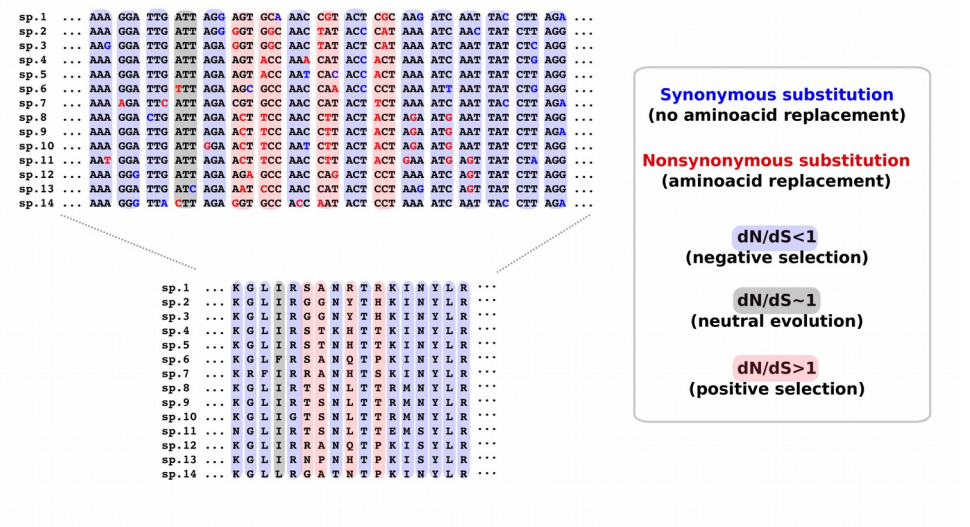


Figure 2: Example of a site-by-site analysis of 14 orthologous coding sequences (sp = species). The first one is a nucleotide alignment, the second one represents its amino acid translation. Image adapted from Sironi and colleagues review [12]..

1.2.2 Detection of selection at a specific site

The most widely used models to detect selection at specific sites are the sites models implemented in the PAML (phylogenetic Analysis by Maximum Likelihood) package [13], which are used to infer positive selection and to identify sites under selection. These models allow dN/dS to vary from site to site, assuming a constant rate at synonymous sites. The analysis is based on fitting the data (multi-species alignment and phylogenetic tree) to different models that allow (positive models) or do not allow (neutral models) a class of codons to evolve with dN/dS>1. To determine whether the neutral model can be rejected in favor of the positive selection model,

likelihood ratio tests (LTR) are then applied. If neutral models are rejected in favor of positive selection ones, the gene can be declared to be positively selected. Only in this case, a Bayes Empirical Bayes approach (BEB) can be used to detect specific sites targeted by selection [14]. In particular, BEB calculates the posterior probability that each site belongs to the class of codons with dN/dS >1) [15].

The PAML approach assumes that the strength and direction of natural selection is uniform across all lineages. Because this is often not the case, different approaches can be applied. For instance, MEME (the Mixed Effects Model of Evolution), developed by Murrell and co-workers [16], allows the distribution of dN/dS to vary from site to site and from branch to branch. The method has therefore greater power to detect episodic selection, a selective pressure acting on a limited number of lineages, especially if it is confined to a small subset of branches in the phylogeny.

1.2.3 Detection of selection of lineages under positive selection and lineages-specific sites

Often positive selection is focused on few sites with specific roles in the function or stability in an otherwise selective constrained protein.

To detect selection along specific branches, PAML also implements methods called “branch-site” models [17]. These models require the phylogeny to be divided into two classes: “foreground” and “background” branches. An LTR is then applied to compare a model that allows positive selection on a class of codons only for the foreground branches with a model that does not allow such selection [17]. Designation of the foreground branches needs *a priori* information, possibly based on biological evidence. If this information is not available, it is still possible to run the analysis by designating each branch of the tree as “foreground”.

Different methods can then detect selection at specific lineages without *a priori* branch partition. One of these is the branch site-random effects likelihood (Bs-Rel) method. Bs-Rel simulates three different evolutionary scenarios (purifying, neutral, and diversifying selection) for all branches in a given tree, and each branch is considered independently from the others; in this way the algorithm applies sequential likelihood ratio tests to identify branches that are influenced by positive selection [18].

1.2.4 Detection of positive selection using a population genetics-phylogenetics approach

Rather than comparing the amount of change at synonymous and nonsynonymous sites between two or more species, there are methods that exploit the availability of an ever increasing number of sequences for the same species to improve power and resolution in detecting natural selection. An example is the population genetic-phylogenetic method developed by Wilson and colleagues [19], which jointly compares intra-species variation and inter-specific divergence to estimate the distribution of selection coefficients (γ) along coding regions. gammaMap estimates the frequency of codons at which non-synonymous mutations fall into one of twelve categories, that range from strongly beneficial ($\gamma = 100$) to inviable ($\gamma = -500$), with γ equal to 0 indicating neutrality. This method allow also to identify specific codons evolving under positive selection in each lineage.

1.2.5 Detection of positive selection in human populations

The complex patterns of genetic diversity in human populations are the product of many levels of demographic and evolutionary events acting on different timescales, 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, and they strongly support an African origin for modern humans [20, 21]. Therefore, population genetic analyses must take into account demographic history. This is achieved mainly by using an outlier approach: genes/variants are considered candidate selection targets if they represent outliers in an empirical distribution.

To search for polymorphisms that have been targeted by natural selection in human populations, a set of tests are applied. Briefly, the following tests/parameters are calculated:

- nucleotide diversity: both θ_W [22], an estimate of the expected per site heterozigosity, and π [23], defined as the average number of nucleotide differences per site between two DNA sequences, are extimated; these are increased under balancing selection, decreased under purifying selection or selective sweeps.

-Tajima's D [24]: this statistic takes into account both SNP frequency and allele spectra. It is close to 0 under the standard hypothesis of neutral evolution; positive values indicate balancing selection while negative values are suggestive of purifying selection or selective sweep.

-F_{st}: it is a measure of population genetic differentiation. Lower F_{st} values are expected at loci under balancing selection compared to neutrally evolving ones [25]; high values are suggestive of positive selection.

-DIND test [26]: the rationale behind the DIND test is that a derived allele under positive selection will display lower levels of nucleotide diversity at linked sites than expected from its frequency in the population. Thus, the ratio of intra-allelic diversity associated with the ancestral and derived alleles ($i\pi_A/i\pi_D$) is analyzed against the frequency of the derived allele (DAF); given a DAF interval, a high value of $i\pi_A/i\pi_D$ indicates that the

neutral diversity associated with the derived allele is limited, suggesting positive selection.

- Fay and Wu's H (DH): the test is based on the idea that directional selection at one site may drive linked mutations to high frequency; this is also applied to derived alleles (which usually display lower frequency). Negative values indicate an excess of high frequency derived alleles and represent a signature of selective sweeps.

1.3 Drivers of natural selection

Migration and colonization of new habitats are a major component of the species history. Such processes lead to sharp changes in environmental conditions (water and food availability, climate, pathogens richness and diversity) experienced by organisms. Changes in environmental conditions can lead to shifts in selective pressures which in turn alter the phenotypic composition of species [27].

Detecting how natural selection has affected genome variability has proven to be a powerful tool to delineate genes and biological functions having played a key role in species adaptation [28]. This holds true for both mammal and pathogen evolutionary history.

1.3.1 Adaptation to dietary pressure

Selective pressures related to diet have varied over geographic space, as species adapted to different ecological niches and to use dietary components available in those environments. In addition, human diets have changed over time including the recent past (the last 100000 years) with the agricultural revolution and the consequent domestication of plants and animals [29]. Changes in food availability and diet composition during evolution, likely created strong selective pressures on multiple biological processes. The identification of the genetic loci that were targeted by these

diet-related selected pressures may provide insight into the evolutionary history of the species as well as into the biological pathways that mediate the effect of dietary risk factors for common diseases, such as diabetes, hypertension and cancer [30].

1.3.1.1 Candidate genes of natural selection under dietary-driven selective pressures

Among the biological processes that may have adapted to dietary-driven selective pressures I mention:

-Metabolism: the process of building molecular structures from nutrients and breaking them down for energy production. Although many metabolic pathways play an important role under all dietary regimes, specific pathways and reactions may become critical when species specialize on particular dietary components. For this reason, initial efforts to identify genetic adaptation to dietary specializations have focused on enzymes that have a well-characterized and a highly specific functional role in nutrient metabolism [29].

-Sensory perception: feeding is a multisensorial experience; before being tested, food is seen, touched and smelled. Processing of food in the mouth leads to the release of additional molecules that stimulate taste and olfaction, as well as to the production of sounds that stimulate auditions. The five senses are developed to different degrees across species, reflecting their feeding habits and needs, as well as to the whole complex of interactions with the surrounding environment. Many studies have sought to understand the evolution of olfaction, vision and taste in human and non-human primates, with important implications for understanding adaptations to dietary changes. A special focus of evolutionary analyses has been on

the taste receptors (in particular bitter taste receptors because of their role in preventing the ingestion of poisonous foods) and olfactory receptors, and on the broader category of genes involved in chemosensory perception [26, 31-33].

-Appetite control: appetite includes various aspect of eating patterns, such as frequency and size of eating episodes, choice high-fat or low-fat foods, energy content and diversity of food consumed. The regulation of appetite is a biological process tightly connected to the environmental availability of food and to lifestyle. Therefore genes involved in appetite control are potential candidates as target selective pressure when species faced different dietary challenges [33, 34].

-Morphological development of the digestive system: is likely to have adapted to dietary shifts. An example is dental enamel thickness, a phenotypic trait of particular interest, as it may reflect adaptation to variable fracture-resistence properties of food items. Enamelin peptides are thought to be involved in the formation and elongation of enamel crystallites during tooth development genes and for these processes were identified as targets of recent positive selection [35-37].

1.3.2 Adaptation to pathogens

Infectious diseases and epidemics have always accompanied and characterized the history of most species, representing a major cause of death.

Migration and cultural changes during the recent human evolutionary history (the past 10000 years) exposed populations to dangerous pathogens as they colonized new environments, increased in population density and had closer contact with animal disease vectors, including both domesticated animals (dog, cattle and sheep) and those exploiting permanent human settlements (rodents and sparrows) [9].

Even today, despite progress in sanitation and medical research, infection are estimated to account for about 15% of deaths in the world's population, reaching about 41% in Africa (WHO 2008, http://www.who.int/en). In a recent report (September 2016), the WHO estimated that in 2015 about 5.9 million children under the age of five died of infection (preterm birth complications, pneumonia, birth asphyxia, diarrhoea and malaria).

Data on human forager-farmers traditionalists with limited access to modern medicine also show infections as a main cause of death (72%) [38] and longitudinal studies of the Gombe chimpanzees [39] identify infection in the majority of deaths (67%) for all ages [40].

It is not surprising so the idea whereby pathogens have been acting as a powerful selective pressure.

1.3.2.1 Zoonotic diseases

Mammals display different susceptibility to distinct pathogens, and, nearly two-thirds (61%) of human diseases originate through a zoonotic transmission from a reservoir animal host [41]. For instance, bats have long been known to harbour and disseminate a wide range of viruses that are likely pathogenic for humans, among which Ebola and Middle East Respiratory Syndrome (MERS-CoV) viruses [42]; the same for rodents, which are known to harbour Lassa virus [43], responsible for a severe hemorrhagic fever. The etiological agent of AIDS, human immunodeficiency virus type 1 and 2, are retroviruses that must likely originated in chimpanzees and in sooty mangabey monkeys, respectively; *Plasmodium falciparum*, the causative agent of malaria, originated from gorillas; high concentrated poultry and pig farming provided optimal condition for increased mutation, reassortment and recombination of influenza viruses [41].

Thus, domestic and wild mammalian (and non-mammalian) species

represent natural reservoirs of human pathogens and/or may provide the adaptive environment for pathogen spillover. Because host reservoir species and their pathogens often co-evolved for millions of years, evolutionary analyses may help to explain host adaptive events associated with low susceptibility and mild disease outcome, as well as to provide valuable information on the differential susceptibility to infection within and among species [12].

1.3.2.2 “Red Queen” scenario

Interactions of pathogens with their hosts result in a situation that is defined as an “arms-race”, in which continuous selective pressure on the host occurs to generate resistance against pathogens, and, at the same time, on pathogens, which try to develop new strategies to evade host defenses for a successful infection. Evolutionary fit pathogens, which are able to survive, replicate, and spread effectively within the host, have an improved chance of passing their genes on to the next generation. Similarly, host genotypes are more likely to persist within the population if those particular individuals are more capable of controlling or resisting infection. Co-evolution between hosts and pathogens naturally occurs as a result of these interactions. The result is a genetic conflict referred to as a “Red Queen” scenario, from the character in Lewis Carrol's novel who says: “It takes all the running you can do, to keep in the same place”. This conflict shapes genetic diversity both in the host and in the pathogens genomes, determining fast evolutionary rates. Protein regions directly involved in host-pathogen interactions are expected to evolve under the strongest diversifying (positive) selection [12].

Moreover, natural host defenses, which take much longer to evolve than their pathogen counter-parts, have been supplemented by man-made developments, such as vaccines, antibiotics and modern medical interventions, which place added pressures on pathogens to adapt. Host

innate and adaptive immune responses and modern medical interventions are all selective pressures that contribute to pathogen evolution within the host [44].

Integrating evolutionary analyses of host and pathogen interacting partners into a common framework therefore hold the promise to improve the understanding of the strategies used by both hosts and pathogens to adapt and counter-adapt. In turn, this knowledge has possible biomedical and therapeutic implications, given the ability of different pathogens or distinct strains of the same infectious agent to elude not only natural host defenses but also drugs and vaccination strategies.

1.3.2.3 Candidate genes under pathogen-driven selective pressures

It's common opinion that protein regions at the host-pathogen interface are targeted by the strongest selective pressure. Among the candidate genes with high rates of adaptation I mention:

- innate immune genes that are devoted to antiviral/ antimicrobial response: the mammalian immune system is endowed with a repertoire of molecular sensors called pattern-recognition receptors (PRRs). These molecules detect pathogen-associated molecular patterns (PAMPs) and initiate a downstream signaling cascade that culminates in the production of cytokines and antimicrobial factors. In the host-pathogen arms-race, these molecules represent one of the foremost detection-defence system.

Host restriction factors are other important components of the innate immune response to infection: they generate potent, widely expressed intracellular blocks to viral replication and for these reasons represent obvious targets in host-pathogen arms-races. Specifically, genetic conflicts between host restriction factors and viral components often play out in terms of binding-seeking dynamics (e.g. the host factor adapts to bind the viral component) and binding-avoidance dynamics (e.g. the virus counter-

adapts to avoid binding and restriction by the host factors).

Consistently, several studies have reported adaptive evolution at proteins with a specialized role in antiviral defense [45-51];

- antigen presentation, T cell activation and immunoglobulin G receptors: antigen presentation and T cell activation are central processes in mammalian cell-mediated immune response. Therefore, a convenient strategy for pathogens to elude immune surveillance is to hijack the molecular pathways responsible for these processes [52, 53]. In line with the arms-race scenario, there is evidence of positive selection at several mammalian genes involved in antigen presentation and in the regulation of T cell activation [54, 55];
- other than immune effectors: despite the relevance of adaptive immunity for host defense, host-pathogen interactions are not limited to immune system components. Host gene products that engage in genetic conflicts include those that participate in the coagulation cascade and the contact system, which are commonly used by bacterial pathogens to promote tissue invasion or to elude detection by immune cells. Moreover, incidental receptors are often represented by the products of housekeeping genes, which, from the viral perspective, have two advantages: first they are expressed at high levels in multiple cell types, and second they are highly conserved because most amino acid replacements cause a substantial reduction in fitness (this means that the sequence space available to the host for adaptive change is limited) [12].

As far as pathogens are concerned, protein under host-driven selective pressure could be all proteins that mediate the initial and essential steps of host infection via host cell binding and entry and those represent major targets for immune responses influencing antigenic selection.

1.4 Aim of my thesis

The purpose of my project is to use evolutionary analyses to investigate and describe adaptive events at genes subject to different selective pressures. One aim is to gain a deeper understanding into the evolutionary history of the species. Another important driver for the research is based on the 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. This is particularly relevant when the genes being analyzed have been involved in human diseases or phenotypic traits of medical relevance. Moreover, the selective pressure underlying species adaptations are often environment-driven; therefore evolutionary analysis approaches can provide information on how human and other organisms have adapted to their environment and how environment shifts (as those carried along by agriculture and pathogens) may have resulted in disease susceptibility.

Here I show two sets of studies.

In the first study, I analyze genes subject to dietary pressures: in particular, I exploit the availability of an ever increasing number of mammal sequences to analyze the evolutionary history of genes that encode intestinal brush-border proteins involved in carbohydrate metabolism. This decision was based on the concept that the availability of food resources is a driver of pivotal importance in specie evolution and that the introduction of agriculture resulted in a dietary shift in terms of carbohydrate intake. The availability of ancient DNA made it possible to also test when (if after or before the agricultural revolution) adaptive alleles arose. The identification of dietary adaptation signals may shed light not only on the evolutionary history of our species, but also on the mechanisms that underlie common

metabolic diseases in modern human populations.

In the second set of studies, I use evolutionary large-scale analyses to investigate genes involved in host-pathogen interactions, both in virus/pathogen strains and in primate/mammalian species.

On one hand, I analyze the evolutionary history of molecules most likely involved in this interaction, as engaged at different levels in immune response: antiviral sensing proteins, genes with IFN-inducible properties, and antiviral effectors. This is because molecules that participate in immune response are expected to be engaged in a constant arms-race with pathogens and to harbour the molecular signature of such a conflict.

On the other hand, I study the interaction between pathogens and proteins not directly involved in immune response but known to be engaged by pathogens to invade host cells. That is because not only genes with a specific defense function, but also molecules involved in central homeostatic processes may be engaged in genetic conflicts with pathogens.

The idea is that evolutionary analyses can pinpoint regions and residues directly involved in host-pathogen interaction, and validate the idea whereby protein regions at host-pathogen interfaces are the major target of positive selection. This is ultimately expected to result in the identification of variants in the host or the pathogen that (i) modulate infection susceptibility or virulence and can be prioritized in screening for association with autoimmune diseases and infections; (ii) help to explain changes in pathogen tropism; (iii) explain species-specificity of infections; (iv) provide information on suitable vaccine targets and on treatment option such as the effectiveness of long-term treatments based on antibody combinations.

These works highlight the importance of evolutionary analysis. The methods I used have previously been described; yet, I combined multiple

methodologies to assure that results are conservative.

Large-scale evolutionary analyses in mammals have mostly described general patterns rather than focused on the specific interaction with one or more pathogens. In the context of virology, genome-wide analyses of positive selection during speciation are lacking. These works are innovative because they integrate analysis of the host's and the pathogen's interacting partners into a common framework, to define the molecular determinants of host-range and virulence. The experimental design I used may therefore serve as a model for the analysis of other host-pathogen interactions.

2. METHODS

2.1 Evolutionary analyses

Algorithms, programs, and tests applied for all evolutionary analyses are summarized in Table 1.

2.1.1 Mammalian sequences and samples

Coding sequences were retrieved from the NCBI database (http://www.ncbi.nlm.nih.gov/), from the Ensemble website (http://www.ensembl.org/index.html) and from UCSC server (http://genome.ucsc.edu/). Information for the Neandertal and Denisova high-coverage genomes [56, 57], as well as for a hunter-gatherer Mesolithic European from Spain [58] and a Paleolithic Siberian [59] was obtained by previous works.

To increase the number of sequences or to assess the presence of sequence errors or rare variants in the reference genome assemblies, primates sequences were in some cases obtained by direct sequencing. In particular, coding sequencing information for *Gorilla gorilla*, *Aotus trivirgatus*, *Saguinus oedipus* and *Chlorocebus aethiops* and 3 *Pan troglodytes* was obtained by direct sequencing of cDNA derived from EB(JC), OMK, EBV B95 UK, COS1 and EB176(JC) cells, respectively (obtained by the European Collection of Cell Cultures, ECACC). The genomic DNA of *Colobus guereza*, *Ateles fusciceps*, *Pithecia pithecia, Lagothrix lagothricia* amd 9 *Pan troglodytes*, was kindly provided by the Gene Bank of Primates, Primate Genetics (Germany) and used as a template for gene amplification and sequencing.

The list of species analyzed for each genes varies depending on availability and other factors (e.g. gene loss, reliability of orthology). For all human genes, I checked whether primate genes represented 1-to-1 orthologs

using the Ensembl Compara GeneTrees database [60]. Species for which this information was not available, a BLAST search of the gene coding sequences against the genome of these species was performed using the NCBI BLAST utility. Single primate genes for which BLAST or Ensembl Compara GeneTrees hits were not consistent with the presence of a single orthologs were removed.

2.1.1.1 Sequencing analysis

For the primate sequences obtained through direct sequencing, the cDNAs/DNAs were amplified by PCR and treated with ExoSAP-IT (USB Corporation, Cleveland OH, USA). Purified PCR products were directly sequenced on both strands with a Big Dye Terminator sequencing Kit (v3.1 Thermo Fisher Scientific), and run on a Thermo Fisher Scientific ABI 3130 XL Genetic Analyzer. Sequences were assembled using DNA Baser Sequence Assembler version 4.10.

2.1.2 Viral/protoza/ bacteria sequences

Ebolavirus and *Marburgvirus GP* sequences were retrieved from the NCBI database.

Sequences for *Neisseria gonorrhoeae Por1A* and *Por1B* were obtained from a previous work [61]. *Streptococcus pneumoniae PspC* sequences were retrieved from Iannelli et al. [62]; *Borrelia burgdorferi OspE* sequences were obtained from the NCBI database, with the exclusion of antigenic variants arising during infection; *Borrelia burgdorferi CspZ* sequences derive from human clinical isolates as reported in Rogers et al. [63].

RH5 and *EBA175 Plasmodium* sequences were retrieved from the NCBI database or derived from previous works [64-66].

2.1.3 Alignments and gene trees

One of the most common problems in evolutionary analyses is

reconstructing reliable alignments: alignment errors can generate false positive results due to the incorrect evaluation of the number of nonsynonymous and synonymous substitutions. For sequences showing limited divergence, alignments were generated using softwares that maintain the codon reading frame. In particular the RevTrans 2.0 utility (http://www.cbs.dtu.dk/services/RevTrans/, MAFFT v6.240 as an aligner) [67] was used, which uses the protein sequence alignment as a scaffold for constructing the corresponding DNA multiple alignment. This latter was checked and edited using TrimAl (automated1 mode) [68] and manual editing was used to correct few misalignments in proximity of small gaps (http://phylemon.bioinfo.cipf.es/utilities.html).

Viral sequences show a stronger sequence divergence due to the high mutation rate. Multiple sequence alignment were therefore generated by PRANK, based on a specific algorithm that takes into account the evolutionary distances between sequences. Unreliably aligned codons were then filtered using GUIDANCE [69], an open access tool for the identification of unreliably aligned regions (codons with a score <0.90 were masked as suggested by Privman and colleagues [70]).

The unrecognized action of recombination is another source of false positive results when tests of positive selection are applied [71]. This is because most methods used to infer positive selection assume that the phylogenetic tree and branch lengths are constant across all sites in the alignment, a tenet that is invalid in the presence of recombination. All alignments were thus screened for the presence of recombination breakpoints (and, if necessary, split on the basis of these ones) using GARD (genetic Algorithm Recombination Detection) [72], a genetic algorithm implemented in the HYPHY suite [73] which uses phylogenetic incongruence among fragments to detect recombination events.

Gene trees were generated using phyML [74], a General Time Reversible (GTR) model plus gamma-distributed rates and 4 substitution rate categories.

2.1.4 Detection of natural selection acting on all lineages of a tree

To detect sites targeted by negative selection, the Single Likelihood Ancestor Counting (SLAC) or the Fixed Effects Likelihood (FEL) methods implemented in the HYPHY package were used.

Evidence of positive selection was instead assessed by the codon-based *codeml* program implemented in the PAML (Phylogenetic Analysis by Maximum Likelihood) suite [13, 75], which was developed to infer positive selection and to identify positively selected sites.

This tool analyzes gene alignment to evaluate the nonsynonymous/ synonymous rate ratio, considering the dN/dS ratio for any codon in the gene as a random variable from a statistical distribution, thus allowing ω to vary from site to site, assuming a constant rate at synonymous sites.

To test for selection, site models that allow (M2a, M8) or disallow (M1a, M7 and M8a) a class of sites to evolve with ω >1 were fitted to the data using two different codon frequency models: the F3x4 and the F61 models, which weight in different ways the frequency of each codon in the data analyzed. The nested models (M1a vs M2, M7 and M8a vs M8) were compared through likelihood-ratio tests (degrees of freedom= 2 except for the comparison between M8a and M8 which has degree of freedom= 1) to asses statistical significance. Positively selected sites were identified using the Bayes Empirical Bayes (BEB) analysis (with a cut-off of 0.90). BEB calculates the posterior probability that each codon is from the site class of positive selection (under model M8) [15]. For the identification of specific positively selected sites the Mixed Effect Model of Evolution (MEME) from HYPHY (with the cutoff of 0.1 [16]) was also applied. MEME allows the

distribution of ω to vary from site to site and from branch to branch at a site, therefore allowing the detection of both pervasive and episodic positive selection. The REL (Random Effects Likelihood) [76] and FEL (with the default cutoff of 0.1) tools were also applied to identify positively selected sites. REL models variation in nonsynonymous and synonymous rates across sites according to a predefined distribution, with the selective pressure at an individual site inferred using an empirical Bayes approach; FEL directly estimates nonsynonymous and synonymous substitution rates at each site [76].

SLAC, FEL, MEME and REL analyses were performed either through the DataMonkey server (http://www.datamonkey.org) [77] or run locally (through HYPHY).

2.1.5 Detection of episodic selection

Positive selection can act on all lineages in a tree, but also on specific branches. To explore possible variations in selective pressure among different lineages, other models from the PAML package were used, the so called free-ratio models. The M0 model assumes all branches to have the same ω, whereas M1 allows each branch to have its own ω [75]. The models are compared through likelihood-ratio tests (LRT) (degree of freedom=total number of branches -1). In order to identify specific branches with a proportion of sites evolving with dN/dS>1, the Branch Site-Random Effects Likelihood (Bs-Rel) was used. This method implements branch-site models that simultaneously allow ω variation across branches and sites. One advantage of Bs-Rel is that it requires no prior knowledge about which lineages are of interest (i.e., are more likely have experienced episodic diversifying selection).

To cross-validate branches identified using this approach or to analyze the presence of episodic positive selection on specific branches, branch-site

LRTs from *codeml* (the so-called modified model A and model MA1, “test 2”) [17] were used. In this test, branches are divided *a priori* into foreground (those to be analyzed for positive selection) and background lineages. The test is based on the comparison between the MA model, that allows positive selection on the foreground lineages, with model MA1 that does not allow such positive selection. A false discovery rate (FDR) correction was applied to account for multiple hypothesis testing (i.e., correcting for the number of tested lineages), as suggested [78]. To identify sites that evolve under positive selection on specific lineages, the MEME and BEB analysis from MA (with a cutoff of 0.90) [17] and BUSTED (Branch-site Unrestricted Statistical Test for Episodic Diversification, [16]) were used [17]. BUSTED is a test designed to detect the action of episodic positive selection that is acting on a subset of branches in the phylogeny at a proportion of sites within the alignment. To detect selection at individual sites, twice the difference of likelihood for the alternative and the null model is compared to a χ^2 distribution with one degree of freedom. A site was considered as positively selected if it showed a *p* value <0.05.

Bs-Rel, BUSTED and MEME analyses were performed either through the DataMonkey server (http://www.datamonkey.org) [77] or run locally (through HYPHY).

2.1.6 Detection of positive selection in bacteria

Because recombination rates in bacterial genomes can be very high [79], analysis of positive selection in bacteria was performed using omegaMap, a Bayesian method that simultaneously estimates recombination and selection (inferred through ω estimation) [80]. The program performs Bayesian inferences of ω and ρ (recombination parameter), allowing both parameters to vary along the sequence. An average block length of 10 and 30 codons was used to estimate ω and ρ, respectively. To determine the

influence of the choice of priors on the posteriors, the analyses were repeated with two alternative sets of priors. For each alignment, three independent omegaMap runs, each with 500,000 iterations and a 50,000 iteration burn-in, were compared to assess convergence and merged to obtain the posterior probabilities.

2.1.7 Detection of co-evolving sites

In order to analyze the presence of co-evolving sites, two different methods were applied: BGM (Bayesian Graphical Model)-Spidermonkey [81] and the Mutual Information Server To Infer Coevolution (MISTIC) [82]. BGM-Spidermonkey identifies co-evolving sites from coding sequences; a Bayesian Graphical Model is used to evaluate the connection among codons in the alignment (represented by the nodes of the network). Significant statistical associations between nodes are indicated by the edges of the network, suggesting functional or structural interactions between codons. BGM-Spidermonkey is implemented in the HYPHY package.

MISTIC estimates the relationship between two or more alignment positions. The co-evolutionary association is evaluated by Mutual Information (MI), estimating whether the information from the amino acid at the first position can help to predict the amino acid information at the second position.

For BGM-Spidermonkey sites were filtered based on a minimum count of 4 substitutions across the phylogeny and each site was conditionally dependent on one other site. To be conservative, a pair of residues was considered as co-evolving only if it showed a posterior probability >0.95. Likewise, MISTIC site pairs were required to display an MI rank higher that the 95th percentile calculated using all MI scores from the alignment. Pairs of sites exceeding the threshold for both methods were declared to be co-

evolving.

2.1.8 Detection of positive selection in Homininae

The action of positive selection in the Homininae lineages was analyzed using gammaMap, a population genetics-phylogenetics method developed by Wilson and colleagues [19]. This method compares intra-species variation and inter-specific divergence to estimate the distribution of selection coefficients (γ) along coding regions. In particular, gammaMap assigns the selection coefficient γ to 12 different categories of selective effects, ranging from strongly beneficial (γ=100) to effectively unviable (γ =-500), with γ equal to 0 indicating neutrality. For gammaMap analyses, genotype data from the phase 1 of the 1000 Genomes Project were retrieved from the dedicated website [83]; in particular SNP information for three human populations were retrieved: African (YRI), European (CEU), and Chinese (CHB). For chimpanzees and gorillas, genotype information were retrieved from a previous work [84] for 25 and 27 individuals, respectively.

Ancestral sequences were reconstructed by parsimony from the human, chimpanzee, orangutan and macaque sequences or using the ASR utility from Datamonkey. ASR implements three different methods based on maximum-likelihood or Bayesian inference [76].

gammaMap analysis was performed assuming θ (neutral mutation rate per site), k (transitions/transversions ratio), and T (branch length) to vary among genes following log-normal distributions.

For P (the probability that adjacent codons share the same population-scaled selection coefficient), a value of 0.02 was assumed and the neutral frequency of non-STOP codons was set to 1/61. For population-scaled selection coefficients, a uniform Dirichlet distribution with the same prior weight for each selection class was considered. For each gene, two Markov

chain Monte Carlo runs of 100000 iteration each were run with a thinning interval of 10 iterations. Runs were compared for convergence and merged for the analyses. To be conservative, a codon was declared to be targeted by positive selection when the cumulative posterior probability of $\gamma \geq 1$ was greater than 0.75, as suggested [85].

2.1.9 Human population genetic analysis

A set of programs was developed to retrieved genotypes from the 1000 Genomes Pilot Project MySQL database [86] and to analyze them according to selected regions/populations. These programs were developed in C++ using the GeCo++ [87] and the Libsequence [88] libraries. Genotype information for the genes of interest were obtained and, in particular, three human populations with different ancestry were analyzed: Europeans (CEU), Africans (Yoruba ,YRI), and East Asians (Han Chinese in Bejing, CHB). In order to obtain a control set of approximately 1000 genes to use as a reference set, 1200 genes were initially selected by random sampling of those included in the RefSeq list. For these genes orthologous regions were retrieved in the chimpanzee, orangutan or macaque genomes (outgroups) using the LiftOver tool. Genes showing less than 80% human-outgroup aligning bases were discarded. This originated a final set of 987 genes, referred to as control set. These data were used to calculate θ_W [22], π [89] as well as Tajima's D [24] over each entire gene regions. Normalized Fay and Wu's H (DH) was also calculated in 500bp sliding windows moving with step of 500bp [90, 91].

The pairwise F_{ST} [92] and the DIND (Derived Intra-allelic Nucleotide Diversity) [93] test were calculated for all SNPs mapping to the analyzed genes, as well as for SNPs mapping to the control set. F_{ST} values are not

independent from allele frequencies, so variants were binned in 50 classes based on the minor allele frequency (MAF) and the F_{ST} empirical distribution was calculated for each MAF class using the control data set. The same procedure was applied for the DIND test; statistical significance was thus calculated by obtaining an empirical distribution of DIND values for variants located within control genes; in particular, the DIND test was calculated using a constant number of 40 flanking variants (20 upstream and 20 downstream), as described [54]. DIND values for the three human populations were binned in 100 derived allele frequency (DAF) classes, and for each class the distributions were calculated. As suggested [93], for values of iπD=0 the DIND value was set to the maximum obtained over the corresponding class plus 20.

As a confirmatory signature of positive selection in human population, DH was also calculated. Sliding window analyses have an inherent multiple testing problem that is difficult to correct because of the non-indipendence of windows. In order to partially account for this limitation, DH was also calculated for the control gene set, and the distribution of the statistic was obtained for the corresponding windows.

Table 1. Summary of algorithms, programs, and tests used for bioinformatics analysis.

Computational resources/ statistics	Description	*References*
Evolutionary analisis inter-species		
Databases		
NCBI (National Center for Biotechnology Information) database	The National Center for Biotechnology Information database provides access to biomedical and genomic information.	
Ensembl	It is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation.	
Server		
DataMonkey sever	Web server for HyPhy, a computational phylogenetics software package developed to	[77]

	perform maximum likelihood analyses of genetic sequence data and equipped with tools to test various statistical hypotheses.	
MISTIC (Mutual Information Server To Infer Coevolution)	Web tool that aims to estimate the mutual coevolutionary relationship between two residues in a protein family using corrected Mutual Information (MI).	[82]
Utilities		
EnsemblCompara GeneTrees	The database allows performing cross-species analyses to infer gene orthology and paralogy by using phylogenic gene trees generated by maximum likelihood.	[60]
RevTrans 2.0 utility	It virtually translates a set of DNA sequences, aligns the peptide sequences, and uses this as a scaffold to construct the corresponding DNA multiple alignment.	[67]
TrimAl	This tool allows the automated removal of spurious sequences or poorly aligned regions from a multiple sequence alignment.	[68]
GARD (Genetic Algorithm Recombination Detection)	It is a genetic algorithm of the Hyphy package developed to search for recombination breakpoints in multiple sequence alignments and to identify putative recombinant sequences.	[72]
SLAC (Single Likelihood Ancestor Counting)	A tool from the Hyphy package for the estimation of the average dN/dS ratio; it uses likelihood-based branch lengths, nucleotide and codon substitution parameters and ancestral sequence reconstructions.	[76]
PhyML	This software estimates maximum likelihood phylogenies from alignments of nucleotide or amino acid sequences.	[94].
codeml	This software is from the PAML package and applies likelihood ratio tests to compare models of gene evolution that allow or disallow a class of codons to evolve with dN/dS >1	[13]
BEB (Bayes Empirical Bayes analysis)	This method is used to identify positively selected sites; it calculates the posterior probability that each codon is from the site class with dN/dS >1 (under models allowing dN/dS>1).	[14, 15]
MEME (Mixed Effects Model of Evolution)	This methods identifies positively selected sites by allowing the distribution of dN/dS to vary from site to site and from branch to branch at a site, thus detecting both pervasive and episodic positive selection. It is included in the HyPhy package	[16]
BSREL (Branch-site REL)	This tool performs a series of LRT tests to find lineages on which a proportion of sites evolve with dN/dS > 1, without making any a priori assumptions.	[18]
BGM (Bayesian Graphical Model)-Spidermonkey	A tool implemented in the HYPHY package that identifies co-evolving sites from coding sequences.	[81]
Population genetics-phylogenetics analysis		
Utilities		
gammaMap	This program is based on a combined population genetics-phylogenetics model of selection. It estimates the distribution of selection coefficients,	[19]

	and allows localization of the signal of selection using a Bayesian sliding window approach. The signature of selection is detected from the contrast in the dN/dS ratio within and between species.	
ASR (Ancestral Sequence Reconstruction)	This is a utility from Datamonkey that reconstructs ancestral sequences using three likelihood-based methods.	[76]
Population genetics analysis		
Database		
1000 Genomes	This database collects information about human genetic variation	[86]
Browser		
UCSC table browser	Provides access to information about location and annotation of genomic regions.	[95]
C++ libraries		
GeCo++	This library allows to manage genomic element annotation, sequences, and positional genomic features; it provides users with tools to keep track of genomic variations.	[87]
libsequence	This library facilitates writing and implementation of evolutionary genetics applications; it is mainly dedicated to the analysis of SNP data.	[88]
Statistics		
θ_W	This parameter estimates of the expected per site heterozigosity.	[22]
π	It is defined as the average number of nucleotide differences per site between two DNA sequences.	[89]
Tajima's D	This test is based on the allele frequency spectrum (i.e. the distribution of allele frequencies at polymorphic sites); low negative values of D indicate an excess of rare alleles and suggest either purifying or positive selection.	[24]
DH	The test is based on the idea that directional selection at one site may drive linked mutations to high frequency; this also applies to derived alleles (which usually display lower frequency). Negative values indicate an excess of high frequency derived alleles and represent a signature of selective sweeps.	[90, 91]
F_{ST}	This parameter, also known as fixation index, measures variations in the allele frequency between two populations. F_{ST} largely depends on demographic history (which affects all loci equally) but natural selection may drive allele frequencies to differ more or less than expected on the basis of demography alone. Specifically, local adaptation may cause an allele to increase in frequency in one population and therefore result in high F_{ST} (high differentiation with another population)	[92]
DIND test (Derived Intra-allelic Nucleotide Diversity)	The DIND test evaluates haplotype homozygosity. It is based on the difference of nucleotide diversity between haplotypes carrying the derived and the ancestral alleles It has higher power to detect recent selective events compared with the	[93]

	commonly used sequence-based neutrality tests.	
Utilities		
omegaMap	This is a Bayesian method that simultaneously estimates recombination and selection (inferred through ω estimation).	[80]
LiftOver tool	This tool, available from the UCSC genome browser, converts a given genome position from a genome assembly to the corresponding position in another assembly.	[96]

2.2 Protein 3D structures, in silico mutagenesis, and protein-protein docking

For the *in silico* mutagenesis and the protein-protein docking I worked in association with a group at the University of Milan Bicocca.

Protein 3D structures were derived from the Protein Data Bank (PDB) or predicted using three different methods: MODELLER [97] with loop refinement, I-TASSER [98, 99] with a defined template, or I-TASSER without any template. The quality of each model was assessed with VADAR (Volume, Area, Dihedral Angle Reporter), which uses several algorithms to calculate different parameters for individual residues and for the entire protein [100]. The overall quality was estimated with respect to its geometry and energy (packaging defects, free energy of folding, core hydrophobic and charged residues). Secondary structures were validated through the use of PSIPRED [101] server. According to these criteria, the best among the three models was used for our analysis.

Protein-protein interaction analyses were performed using PIC (Protein Interaction Calculator) [102].

Protein-protein docking analysis was performed using ClusPro [103]. To validate the method, a first run of docking was carried out using the interacting partners to verify that the output was comparable to the three-

dimensional structure of the PDB file. After mutagenesis, a new run of docking was performed. The 10 best cluster structures were analyzed to verify whether the 'native protein-protein conformation' could be found among them. Because the sampling on the conformation is very extensive (10^9 combinations), if the 'native conformation' is not found in the clusters, it is safe to assume that this type of binding is no longer stable and the mutation results in a perturbation of the native binding.

In silico mutations were generated with the FoldX tool run-muta [104]. Specific mutation were performed 5 times to ensure convergence. Images were created using PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.2 Schrödinger, LLC).

2.3 Haplotype Association with HIV-1 Infection Susceptibility

2.3.1 Human subjects, genotyping and statistical analysis

191 males exposed to HIV-1 infection by injection drug use (IDU) and enrolled in prospective cohort studies in Spain (Valme Hospital, Sevilla) who had shared needles for >3 months were recruited. Concurrent markers of HCV infection were present in 100% of IDU subjects. Eighty-five of these subjects were HIV-1 negative (IDU-HESN (HIV-1 exposed seronegative individuals)) and 106 were HIV-1 positive (IDU-controls (CTR)).

Thirty-eight Spanish HESN exposed to the virus through unprotected sexual intercourse (Sexual Exposed (SexExp)-HESN) were also recruited. These subjects are female partners of HIV-1 positive patients (without treatment and viremic, mean number of unprotected sexual intercourse per year: 110, mean number of years as sexual partners: 5, range 3–17). Healthy Controls (HCs) (n = 180) that were anonymous blood donors from Jaen Hospital were also enrolled. All these individuals were seronegative for both HIV-1 and HCV. All subjects were Spanish of Caucasian origin. The

study was designed and performed according to the Helsinki declaration and was approved by the Ethics Committee of the participating hospitals and the University of Jaen. All patients and healthy blood donors provided written informed consent to participate in this study.

As for Italian SexExp-HESN, inclusion criteria were a history of multiple unprotected sexual episodes for more than 4 years at the time of 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 [105]. SexExp-HESN and 188 HCs were recruited at the S. M. Annunziata Hospital, Florence; all of them were Italian of Caucasian origin. The study was reviewed and approved by the institutional review board of the S. M. Annunziata Hospital, Florence. Written informed consent was obtained from all subjects.

Variants hypothesized to be associated with HIV-1 infection susceptibility were genotyped through PCR amplification and direct sequencing. PCR products were treated with ExoSAP-IT (USB Corporation, Cleveland OH), directly sequenced on both strands with a Big Dye Terminator sequencing Kit version 3.1 (Thermo Fisher Scientific), and run on a Thermo Fisher Scientific ABI 3130 XL Genetic Analyzer (Thermo Fisher Scientific). Sequences were assembled using DNA Baser Sequence Assembler version 4.10, and inspected manually by two distinct operators. Genetic association analyses were performed by logistic regression and results from the three cohorts were combined using a random-effect metaanalysis; all analyses were performed using PLINK [106].

2.3.2 HIV Infection Assay

This part was performed in association with a group at the Department of Biomedical and Clinical Science “L. Sacco” of the University of Milan.

PBMCs from 50 HESN subjects were separated on lymphocyte separation

medium (Organon Teknica, Malvern, PA); 10 × 10^6 cells/mL were cultured for 2 days at 37 °C and 5% CO_2 in RPMI 1640 containing FBS (20%), phytohemagglutinin (7.5 µg/mL), and interleukin-2 (IL-2) (15 ng/mL). After viability assessment, 2.5 × 10^6 cells were resuspended in medium containing 1 ng of HIV-1_{Ba-L} p24 viral input/10^6 PBMC and incubated for 3 h at 37 °C. Cells were then washed and resuspended in 3 mL of complete medium with IL-2 (15 ng/mL). Cells were plated in 24-well tissue culture plates and incubated at 37 °C and 5% CO_2. After 7 days supernatants were collected for p24 antigen enzyme-linked immunosorbent assay (ELISA) analyses. Absolute levels of p24 were measured using the Alliance HIV-1 p24 ELISA Kit (PerkinElmer). The HIV-infection assay was performed in triplicate. To account for minor differences in virus titer, p24 levels were normalized within experiment. HIV-1_{Ba-L} was provided through the EU programme EVA centre for AIDS Reagents NIBSC, UK.

2.3.3 IFN-α Stimulation and Transcript Quantification

This part was performed in association with a group at the Department of Biomedical and Clinical Science “L. Sacco” of the University of Milan. Whole blood was collected from 45 HCs by venupuncture in Vacutainer tubes containing EDTA (Becton Dickinson, NJ), and PBMCs were separated on lymphocyte separation medium (Organon Teknica, Malvern, PA). Based on data derived from a kinetic study, 5 × 10^5 freshly isolated PBMCs were incubated for 3 h with medium alone or 400 U/ml IFN-α (Sigma Aldrich). RNA was extracted from cultured PBMC by using the acid guanidium thiocyanate–phenol–chloroform method. The RNA was dissolved in RNase-free water, and purified from genomic DNA with RNase-free DNase (RQ1 DNase; Promega, Madison, WI). 1 µg of RNA was reverse transcribed into first-strand cDNA in a 20µl final volume containing

1 µM random hexanucleotide primers, 1 µM oligo dT, and 200 U Moloney murine leukemia virus reverse transcriptase (Clontech, Palo Alto, CA). cDNA quantification for the genes of interest and *GAPDH* was performed by a real-time PCR strategy (DNA Engine Opticon 2; MJ Research, Ramsey, MN). Reactions were performed using a SYBR Green PCR mix (5 Prime, Gaithersburg, MD). Results were expressed as ΔΔCt and presented as ratios between the target gene and the *GAPDH* housekeeping mRNA.

3. RESULTS AND DISCUSSION

3.1 Adaptation to dietary selective pressure

3.1.1 Natural selection at the brush-border: adaptations to carbohydrate diets in humans and other mammals

The ever-increasing availability of genome sequences from different organisms and from multiple individuals of the same species, together with resequencing data of ancient DNA samples, now allows to perform comprehensive evolutionary analyses of biological pathways. Herein I exploited this wealth of information to investigate the evolutionary history of genes that encode intestinal brush-border proteins involved in carbohydrate metabolism. This decision was based on the well-accepted concept that the availability of food resources is a driver of pivotal importance in the evolution in mammals and by the fact that one of the most important turning-points of human history, the introduction of agriculture, resulted in a dietary shift in terms of carbohydrate intake. In this respect, the availability of human DNA samples of pre-agricultural populations allows testing of specific hypotheses as to when adaptive alleles at genes involved in sugar metabolism arose. The text-book examples of positive selection at the LCT (lactase) locus in pastoralism [107], as well as the increase in amylase gene copy number in human populations that consume starch-rich diets [108, 109], were the starting point for the work, which focuses on 9 genes encoding apical brush-border proteins involved in carbohydrate digestion and absorption. Specifically, I analyzed MGAM (maltase-glucomylase), SI (sucrase-isomaltase), LCT and TREH (trehalase), which break down complex sugar into monosaccharides; SLC2A2, SLC5A1 and SLC2A5, that are specialized molecules located at the apical brush-border membrane and transport monosaccharides to enterocytes; TAS1R2 and TAS1R3,

sweet receptors at the intestinal brush-border apical membrane, which probably activate gut hormone secretion through glucose sensing.

I performed an in-depth analysis of the evolutionary history of these 9 genes encoding apical brush-border proteins in 40 mammalian species.

Our results indicated pervasive selection in mammals and human populations at genes coding for brush-border carbohydrate metabolism genes. Episodic positive selection was also detected for several mammalian lineages. For *TREH*, for instance, two positively selected lineages (microbat and platypus) have a diet that includes trehalose. I also detected positive selection at SI in both bat species (megabat is frugivorous, microbat insectivore), with microbat also showing selection signatures at *MGAM* and *TREH*. As an adaptation to flight, bats generally display a reduced small intestinal nominal surface area compared to non-flying mammals, and resort to higher sugar paracellular absorption as a compensation [110]. Positive selection at individual mammalian lineages might therefore reflect specific adaptations to flight (bats) and to specialized diets (insects, crustaceans).

Several positive selected sites lend themselves to further exploration by means of biochemical and molecular biology tools. This will be particularly interesting for sites that involve the catalytic sites or show parallel evolution in MGAM and SI, as well as for those flanking missense mutations.

As for humans, *SLC5A1* (the major transporter of dietary glucose) displayed several positively selected codons in the human but not in the chimpanzee lineage. However, the selective pressure predated agriculture. Indeed, this was the case for most non-coding positively selected variants identified in human populations. Some of these are “modern alleles” when Neandertals and Denisovans are taken as a comparison, but not as modern as to unequivocally support their agriculture-driven spread. Indeed,

based on the sequencing of a Denisova and a Neandertal individual, and on allele frequency in extant human populations, Prufer and colleagues [56] compiled a list of modern-human-specific-alleles, suggested to represent changes that were most important during the recent evolutionary history of our species. Results herein indicate that modern alleles at *SLC5A1* and *SI* were indeed driven to high frequency by natural selection in human populations. Nevertheless, most of these positively selected modern alleles were already present in the Mesolithic and Paleolithic and, therefore, predate the emergence of agriculture. Whether the onset of selection occurred before the Paleolithic or these alleles segregated as neutral standing variation in these early populations remains to be evaluated, possibly through the sequencing of additional ancient samples. Early hominins exploited underground storage organs (USOs, rich in starch) and plant food rich in fermentable carbohydrates (acorns and nuts) [111, 112]. Thus, the introduction of agriculture might have spurred the frequency increase of variants that were already weakly adaptive in hunter-gatherers, resulting in a continuum rather than an abrupt onset of selective events. A similar concept has been proposed for traits unrelated to diet [58]. Clearly, these data have a relevance in explaining the high prevalence of diabetes and obesity seen in modern populations.

Some of the selective events I identified open many interesting research avenues. As an example, I detected positive selection at *SLC2A5* (fructose transporter) in Europeans. Reminiscent of lactose intolerance, some degree of fructose intolerance, associated with gastrointestinal symptoms, is quite common in humans [113, 114] and fructose absorption is reduced by the presence of sorbitol [114]. Thus, selection at *SLC2A5* might have been driven by the domestication and widespread consumption in temperate areas of fruit crops (e.g. apples and pears) that contain excess

fructose plus sorbitol [114]. It will be extremely interesting to test whether the positively selected variant I identified herein (and which is in LD with an eQTL), modulates fructose absorptive capacity.

The analyses provides also valuable informations concerning the susceptibility of human populations to metabolic diseases; e.g. the selection targets at *SLC2A2* are in phase with the risk allele for fasting glucose levels and with the non-risk allele for gamma-glutamyl transferase levels.

Personal contribution to the work: I particularly focused on the evolutionary analysis in mammals and on the lineage-specific selection analysis. I also produced figures and tables for the manuscript.

Natural selection at the brush-border: adaptations to carbohydrate diets in humans and other mammals

Chiara Pontremoli,[1]* Alessandra Mozzi,[1]* Diego Forni,[1]* Rachele Cagliani,[1] Uberto Pozzoli,[1] Giorgia Menozzi,[1] Jacopo Vertemara,[1] Nereo Bresolin,[1,2] Mario Clerici,[3,4] Manuela Sironi[1]

[1] Bioinformatics, Scientific Institute IRCCS E.MEDEA, 23842 Bosisio Parini, Italy.
[2] Dino Ferrari Centre, Department of Physiopathology and Transplantation, University of Milan, Fondazione Ca' Granda IRCCS Ospedale Maggiore Policlinico, 20122 Milan, Italy.
[3] Department of Physiopathology and Transplantation, University of Milan, 20090 Milan, Italy.
[4] Don C. Gnocchi Foundation ONLUS, IRCCS, 20148 Milan, Italy.
* these authors equally contributed to this work

Corresponding author: Manuela Sironi, PhD, Bioinformatics - Scientific Institute IRCCS E.MEDEA, 23842 Bosisio Parini, Italy. Tel: +39-031877915; Fax:+39-031877499; e-mail: manuela.sironi@bp.lnf.it

Abstract

Dietary shifts can drive molecular evolution in mammals and a major transition in human history, the agricultural revolution, favored carbohydrate consumption. We investigated the evolutionary history of 9 genes encoding brush-border proteins involved in carbohydrate digestion/absorption. Results indicated widespread adaptive evolution in mammals, with several branches experiencing episodic selection, particularly strong in bats. Many positively selected sites map to functional protein regions (e.g. within glucosidase catalytic crevices), with parallel evolution at *SI* and *MGAM*. In human populations five genes were targeted by positive selection acting on non-coding variants within regulatory elements. Analysis of ancient DNA samples indicated that most derived alleles were already present in the Paleolithic. Positively selected variants at *SLC2A5* (fructose transporter) were an exception and possibly spread following the domestication of specific fruit crops. We conclude that agriculture determined no major selective event at carbohydrate metabolism genes in humans, with implications for susceptibility to metabolic disorders.

Keywords: *MGAM*; *SI*; *LCT*; *TREH*; *SLC2A2*; natural selection.

Introduction

Diet played an extremely important role in the evolution of mammals and pathways that allow nutrient breakdown and absorption, as well as taste perception, evolved in response to changes in trophic strategies (Karasov et al. 2011). In particular, simple and complex sugars account for a different proportion of energy intake in diverse species and a positive relationship is observed between the dietary intake of carbohydrates and the presence of gut enzymes and transporters necessary for their digestion and absorption (Karasov et al. 2011).

In humans, culture has paralleled and often affected genetic evolution; in particular, the domestication of plant and animals determined dramatic dietary shifts during the evolution of our species. One of the most prominent signals of positive selection in the genome of European populations is observed at the *LCT* gene, encoding a small intestine brush-border enzyme that catalyzes the hydrolysis of lactose into monosaccharides that can be absorbed (Fig. 1A) (Tishkoff et al. 2007). Variants that allow *LCT* expression after weaning are strongly selected for in populations that historically relied on animal husbandry (Tishkoff et al. 2007). Likewise, the development of agriculture

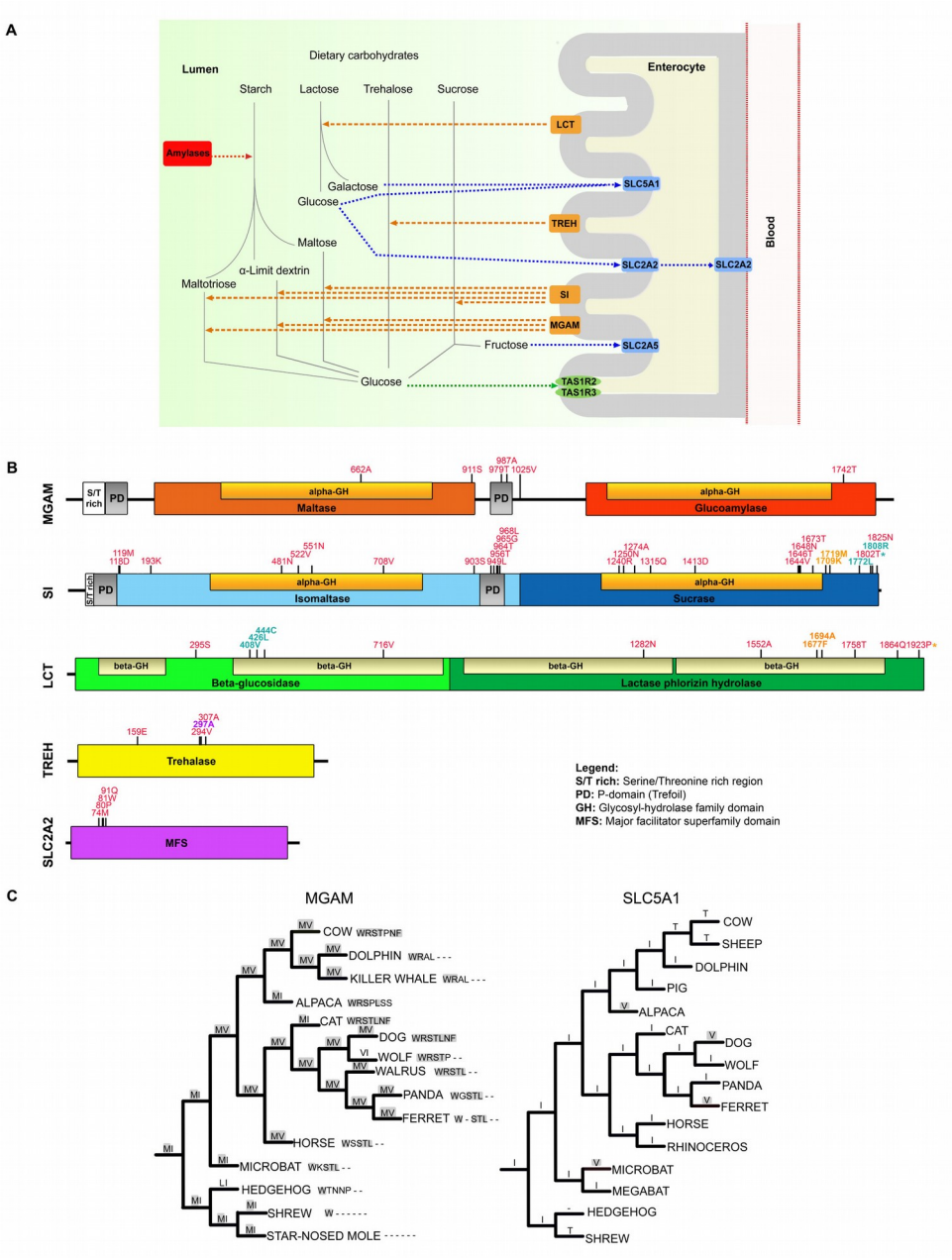


Figure 1. Analyzed genes, protein domain structure, and dog gene analysis. (A) The image is modified from KEGG (hsa04967) and gene products are color-coded with enzymes in orange, transporters in blue, and taste receptors in green. Amylases (not included in this study) are shown in red. (B) Domain representation of positively selected genes. Sites selected in whole phylogeny are in red; positively selected sites in the human, chimpanzee, and gorilla lineage are in cyan, orange, and violet, respectively. Asterisks denote lineage-specific sites that are also selected in whole phylogeny. Positions refer to the human sequence. (C) MGAM and SLC5A1 phylogenetic tree for Laurasiatheria. Aminoacid status at positions 797 and 1001, as well as at the 7 C-terminal positions is shown for MGAM. Gray shading indicates identity with the dog sequence. Position 244 is reported for SLC5A1.

Table 1
List of the Nine Brush-Border Genes Analyzed and Average Nonsynonymous/Synonymous Substitution Rate Ratio (dN/dS)

Gene Symbol	Aliases	Protein Name	Protein Size (amino acids)	Number of Species	Average dN/dS (95% confidence intervals)
MGAM	*MGA, MGAML*	Maltase glucoamylase	1,854	43	0.250 (0.243, 0.257)
SI	—	Sucrase isomaltase	1,833	40	0.286 (0.279, 0.293)
LCT	*LPH*	Lactase-phlorizin hydrolase	1,934	42	0.263 (0.257, 0.269)
TREH	*TREA*	Trehalase	583	43	0.250 (0.240, 0.262)
SLC2A2	*GLUT2*	Solute carrier family 2, facilitated glucose transporter member 2	524	46	0.261 (0.249, 0.274)
SLC5A1	*NAGT, SGLT1*	Sodium/glucose cotransporter 1	664	46	0.172 (0.165, 0.182)
SLC2A5	*GLUT5*	Solute carrier family 2, facilitated glucose transporter member 5	501	42	0.200 (0.191, 0.210)
TAS1R2	*GPR71, T1R2, TR2*	Taste receptor type 1 member 2	839	39	0.272 (0.264, 0.281)
TAS1R3	*T1R3, TR3*	Taste receptor type 1 member 3	852	29	0.238 (0.230, 0.247)

resulted in starch being an increasingly abundant component in human diets. In our species, duplication of the pancreatic *AMY2* gene originated the salivary amylase gene (*AMY1*), which shows extensive copy number variation(Perry et al. 2007). The number of *AMY1* copies is higher in populations that consume high-starch diets, indicating selection for increasing starch digestion capacity (Perry et al. 2007). Analysis of dog genomes also revealed polymorphic increase in *AMY2B* (pancreatic) copy number during domestication, suggesting that these animals adapted to a diet rich in agricultural refuse (Axelsson et al. 2013; Freedman et al. 2014).In most mammals amylases catalyze the first step in the digestion of starch; the following reactions occur in the small intestine where, in addition to LCT, three brush-border enzymes, trehalase (TREH), maltase-glucoamylase (MGAM), and sucrase-isomaltase (SI) break down complex sugars into monosaccharides (Fig. 1A, Table 1). These latter are then transported to enterocytes by specialized molecules (SLC5A1, SLC2A2, and SLC2A5), located at the apical brush-border membrane (Fig. 1A, Table 1).

In addition to enzymes and transporters, sweet taste receptors (TAS1R2 and TAS1R3) have also been observed at the intestinal brush-border apical membrane in different mammals, where they probably activate gut hormone secretion through glucose sensing (Fig. 1A, Table 1).

In line with the central role of starch metabolism in humans and other mammals, the *MGAM* and *SLC5A1* loci were targeted by natural selection in dogs (Axelsson et al. 2013). In humans, signals of selection at genes involved in starch and sucrose metabolism have been detected for populations that rely on roots and tubers as staple foods (Hancock et al. 2010). Nonetheless, the evolution of brush-border carbohydrate metabolic genes has never been analyzed in detail. Herein we use both inter- and intra-species comparisons to analyze the evolution of these 9 genes in mammals and human populations. For the inter-species analyses, we focused on coding regions by applying different methods to assess whether brush-border carbohydrate metabolic genes were targets of either pervasive or episodic positive selection. In this context, positive selection is defined by a faster rate of accumulation of nonsynonymous (aminoacid-replacing) compared to synonymous (non-aminoacid replacing) substitutions, a pattern that may involve only a limited number of sites in a protein. If the selective pressure acted on a limited number of lineages in a phylogeny, it is said to be “episodic”. As for intra-species analyses, we focused on human populations and integrated information concerning archaic hominins: this allowed testing of specific hypotheses as to when adaptive alleles at genes involved in sugar metabolism arose or spread. In this case, we analyzed both coding and non-coding regions and we define positive selection as the frequency increase in a population of a beneficial variant/haplotype (also referred to as selective sweep). The general underlying premise for this study is that natural selection acts on functional genetic variants with a phenotypic effect. Therefore, evolutionary analysis can provide information on the location and nature of adaptive changes that modulate phenotypic diversity in humans and other mammals.

Materials and Methods

Algorithms, programs, and tests applied for all analyses are summarized in Supplementary Table S1, Supplementary Material online.

Evolutionary analysis in mammals

Mammalian sequences genes were retrieved from the NCBI database (as of January 7th, 2015) (Supplementary Table S2, Supplementary Material online). Mammalian orthologs of human brush-border genes were included only if they represented 1-to-1 orthologs as reported in the EnsemblCompara GeneTrees (Vilella et al. 2009). The *MGAM* gene may have undergone domain duplications in some mammals (Naumov 2007). Although all the sequences we obtained from NCBI were comparable in size to the human sequence, we cannot exclude annotation errors and, therefore, aligning of paralogous domains. However, we note that, even in this case, our results would not be significantly affected because the methods we used to detect positive selection are equally applicable to paralogous and orthologous regions (Bielawski and Yang 2003).

DNA alignments were performed using the RevTrans 2.0 utility (Wernersson and Pedersen 2003), which uses the protein sequence alignment as a scaffold for constructing the corresponding DNA multiple alignment. Alignment uncertainties were removed using trimAl (automated1 mode) (Capella-Gutierrez et al. 2009). Alignments were checked by hand before running selection tests.

Recombination may yield false positive results when tests of positive selection are applied (Anisimova et al. 2003). This is because most methods used to infer positive selection assume that the phylogenetic tree and branch lengths are constant across all sites in the alignment, a tenet that is invalid in the presence of recombination. We thus screened all alignments for the presence of recombination breakpoints (the locations where recombination events occur in the alignments) using GARD (Genetic Algorithm Recombination Detection) (Kosakovsky Pond et al. 2006). No evidence of recombination was detected for *LCT*, *SLC2A2*, and *TAS1R2*, whereas breakpoints were detected for the remaining genes.

SLAC (Single Likelihood Ancestor Counting) was applied to calculate the average non-synonymous substitution/synonymous substitution rate (dN/dS) for the 9 genes (Kosakovsky Pond and Frost 2005). To detect positive selection we used the site models implemented in PAML (Yang 1997; Yang 2007); NSsite models that allow (M2a, M8,) or disallow (M1a, M7) sites to evolve with dN/dS >1 were fitted to the data with two models of equilibrium codon frequencies: the F3x4 model (codon frequencies estimated from the nucleotide frequencies in the data at each codon site) and the F61 model (frequencies of each of the 61 non-stop codons estimated from the data) (Supplementary Tables S3, Supplementary Material online). These analyses were performed either for whole gene alignments or independently for sub-regions defined in accordance with the recombination breakpoints. In these latter cases Bonferroni correction for multiple tests was applied to the maximum-likelihood ratio tests (LRT) *p* values (Supplementaty Tables S3, Supplementary Material online). Trees were generated by maximum-likelihood using the program PhyML (Guindon et al. 2009). Whenever maximum-likelihood trees showed differences (always minor) from the accepted mammalian phylogeny, analyses were repeated using the accepted tree, and the same results were obtained in all cases (not shown). Sites under selection with the M8 model were identified using Bayes Empirical Bayes (BEB) analysis with a significance cutoff of 0.90 (Anisimova et al. 2002; Yang et al. 2005). For MEME (Mixed Effects Model of Evolution) (Murrell et al. 2012) the default cutoff of 0.10 was used.

To explore possible variations in selective pressure among different mammals for the five positively selected genes, we tested whether models that allow dN/dS to vary along branches had significant better fit to the data than models that assume one same dN/dS across the entire phylogeny (Yang and Nielsen 1998). This condition was verified for all genes (Supplementary Table S4, Supplementary Material online).

To identify specific branches with a proportion of sites evolving with dN/dS>1, we used BS-REL (Kosakovsky Pond et al. 2011). This method implements branch-site models that simultaneously allow dN/dS variation across branches and sites. One advantage of BS-REL is that it requires no prior knowledge about which lineages are of interest (i.e. are more likely have experienced episodic diversifying selection). Branches identified using this approach were cross-validated using the branch-site likelihood

ratio tests from codeml (the so-called modified model A and model MA1, "test 2") (Zhang et al. 2005). In this test, branches are divided *a priori* into foreground (those to be analyzed for positive selection) and background lineages, and a likelihood ratio test is applied to compare a model that allows positive selection on the foreground lineages with a model that does not allow such positive selection. A false discovery rate correction was applied to account for multiple hypothesis testing (i.e. we corrected for the number of tested lineages), as suggested (Anisimova and Yang 2007). MEME and BEB analysis from MA (with a cutoff of 0.90) were used to identify sites that evolve under positive selection on specific lineages (Supplementary Table S5, Supplementary Fig. S1 and S2, Supplementary Material online).

Ancestral site reconstruction was obtained through the DataMonkey sever by ASR utility, which implements three different methods (Delport et al. 2010) .

GARD, MEME, SLAC, and BS-REL analyses were performed either through the DataMonkey server (Delport et al. 2010) or run locally (through HyPhy) (Supplementary Table S1, Supplementary Material online).

Population genetics-phylogenetics analysis

Data from the Pilot 1 phase of the 1000 Genomes Project were retrieved from the dedicated website (1000 Genomes Project Consortium et al. 2010). SNP genotype information for 25 unrelated chimpanzees and 27 unrelated gorillas were retrieved from (Prado-Martinez et al. 2013). Coding sequence information was obtained for the 9 genes and the ancestral sequence was reconstructed by parsimony from the human, chimpanzee, orangutan and macaque sequences. Analyses were performed with gammaMap (Wilson et al. 2011).

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

Population genetics analyses

A set of programs was developed to retrieve genotypes from the 1000 Genomes Pilot Project MySQL database (1000 Genomes Project Consortium et al. 2010) and to analyse them according to selected regions/populations. These programs were developed in C++ using the GeCo++ (Cereda et al. 2011) and the libsequence (Thornton 2003) libraries. Genotype information was obtained for the 9 brush-border genes. In order to obtain a control set of ~1,000 genes to use as a reference set, we initially selected 1,200 genes by random sampling of those included in the RefSeq list. For these genes we retrieved orthologous regions in the chimpanzee, orangutan or macaque genomes (outgroups) using the LiftOver tool; genes showing less than 80% human-outgroup aligning bases were discarded. This originated a final set of 987 genes, hereafter referred to as control set. Compared to the control set, no brush-border gene was exceptional in terms of recombination rate and none (with the exclusion of *TAS1R3*, which displayed no selection signature) had unusually high GC content, which may bias selection inference (Pollard et al. 2006) (Fig. S3, Supplementary Material online). Nucleotide diversity over whole gene regions was measured as π (Nei and Li 1979) and θ_W (Watterson 1975). DH (Fay and Wu 2000; Zeng et al. 2006) was also calculated in 5kb sliding windows moving with a step of 500 bp. Sliding window analyses have an inherent multiple testing problem that is difficult to correct because of the non-independence of windows. In order to partially account for this limitation, we applied the same procedure to the control gene set, and the distribution of DH was obtained for the corresponding windows. This allowed calculation of the 5th percentile and visualization of regions below this threshold.

F_{ST} (Wright 1950) and the DIND (Derived Intra-allelic Nucleotide Diversity) test (Barreiro et al. 2009) were calculated for all SNPs mapping to the control and brush-border gene sets. Because F_{ST} values are not independent from allele frequencies, we binned variants based on their MAF (Minor Allele Frequency, 50 classes) and alculated the percentiles distributions for each

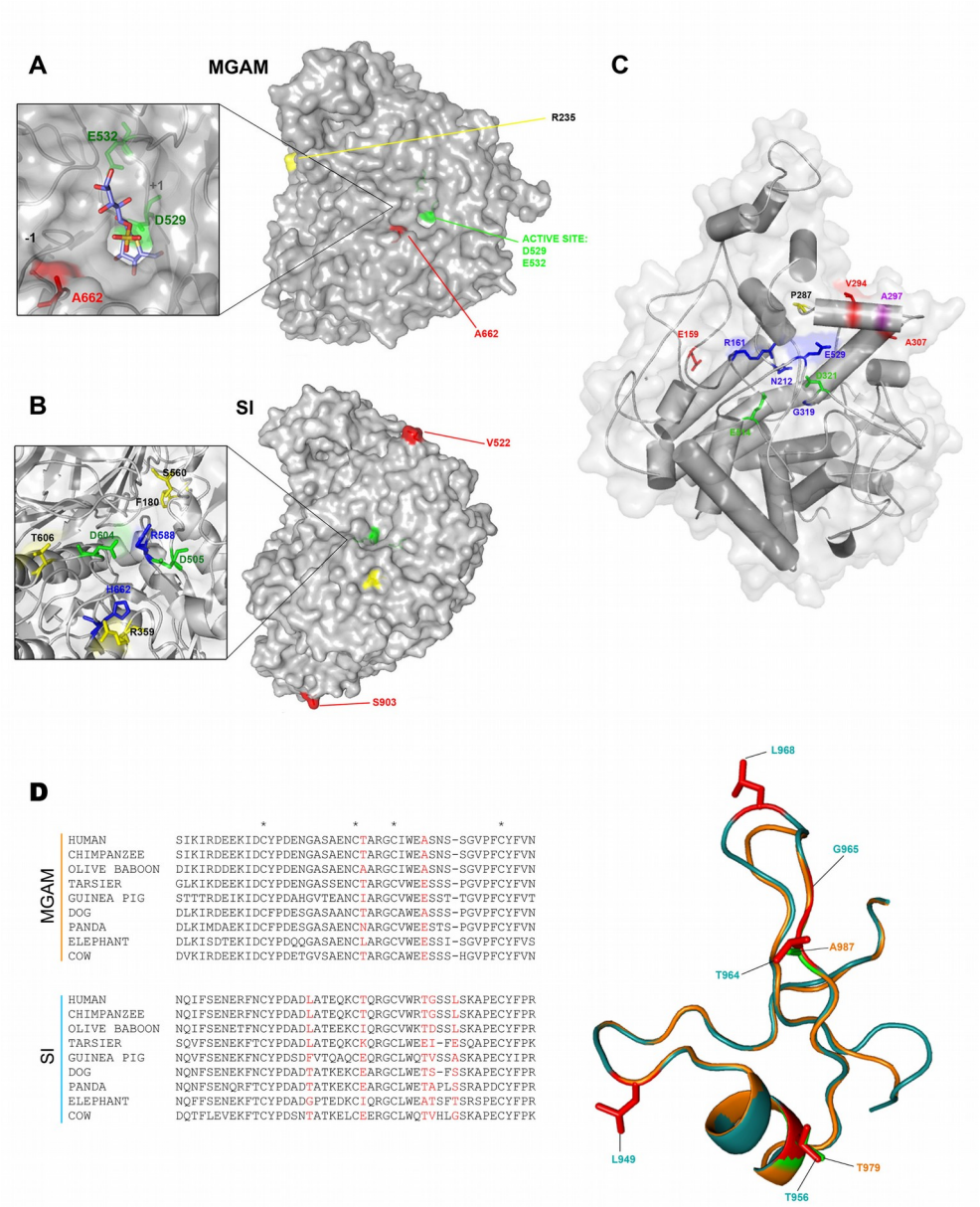


Figure 2. 3D mapping of selected sites. Surface representation of MGAM maltase domain (PDB: 3L4V) in complex with kotalanol (blue stick) (A) and SI isomaltase domain (PDB: 3LPP) (B). Catalytic crevices are shown in the enlargements; color codes as follows: red, positively selected sites in the whole phylogeny; yellow, lineage-specific sites; orange and cyan, positively selected sites in the chimpanzee and human lineages, respectively; green, catalytic residues; blue, aminoacids involved in ligand binding (Sim et al. 2010). (C) Mapping of positively selected sites onto the TREH structure; color codes are as above; violet: positively selected residues in gorilla. (D) Multiple alignment of MGAM and SI trefoil domains for a few of representative mammalian species; positively selected sites (whole phylogeny) are in red. Asterisks indicate conserved cysteine residues. The structural superimposition of trefoil domains of MGAM (orange) (PDB code: 3TON) and SI (light blue) (Protein Model Portal code: P14410, Model 2) is also shown. Positively selected sites on whole phylogeny are represented as sticks, green for SI and red for MGAM.

MAF class. As for the DIND test, we calculated statistical significance by obtaining an empirical distribution of DIND-DAF (Derived Allele Frequency) value pairs for variants located within control genes. Specifically, DIND values were calculated for all SNPs using a constant number of 40 flanking variants (20 up- and down-stream). The distributions of DIND-DAF pairs for Yoruba (YRI), Europeans (CEU), and Chinese plus Japanese (CHBJPT) was binned in DAF intervals (100 classes) and for each class the percentiles distributions were calculated. As suggested previously (Barreiro et al. 2009), for values of $i\pi_D = 0$ we set the DIND value to the maximum obtained over the whole dataset plus 20. Due to the nature of low-coverage data, for low DAF values most $i\pi_D$ resulted equal to 0 (i.e. the 95th percentile could not be calculated); thus, we did not calculated DIND in these ranges and we consequently cannot detect selection acting on low frequency derived alleles.

For the DIND test, an approach based on coalescent simulations was also applied to assess statistical significance. In particular, coalescent simulations were performed using the cosi package (Schaffner et al. 2005) with 2000 iterations. Simulations were conditioned on mutation and recombination rates, and on a region length of 20,000 bp. We simulated demographic patterns using parameters for YRI, CEU and AS as described in (Grossman et al. 2010) with a data thinning procedure that improves fitting to the 1000 Genomes empirical data (Engelken et al. 2014). Estimates of the population recombination rate parameter ρ were obtained from UCSC table browser.

Results

Most brush-border carbohydrate digestion/absorption genes evolve adaptively in mammals.

We analyzed the evolutionary history of genes involved in carbohydrate metabolism. These were selected on the basis of KEGG pathway "Carbohydrate digestion and absorption" (hsa04973) with the inclusion of brush-border proteins only and the addition of *TREH* (GO:0044245, polysaccharide digestion) (Fig. 1A, Table 1). We obtained coding sequence information from public databases. Except for *TAS1R3*, at least 39 species were available for each gene (Table 1, Supplementary Table S2, Supplementary material online). We first calculated the average non-synonymous substitution/synonymous substitution rate (dN/dS) for the 9 genes: in all cases dN/dS was lower than 1 (Table 1), indicating a major role for purifying selection in shaping genetic diversity. Although constraints on protein function and structure often result in purifying selection being the primary force that shapes diversity at coding sequences, diversifying selection might involve specific sites or domains. To test this possibility, we applied maximum-likelihood ratio tests (LRT) implemented in the codeml program (Yang 2007) after accounting for the presence of recombination. Specifically, we compared models of gene evolution that allow (NSsite models M2a and M8, positive selection models) or disallow (NSsite models M1a and M7, null models) a class of codons to evolve with dN/dS >1. To assure reliability, different codon substitution models were used (Supplementary Table S3, Supplementary Material online). Results indicated that five brush-border genes were targeted by positive selection in mammals (Fig. 1B, Supplementary Table S3, Supplementary Material online). In order to identify specific sites subject to positive selection, we applied the Bayes Empirical Bayes (BEB) analysis (Yang et al. 2005), which calculates the posterior probability that each codon is from the site class of positive selection (under model M8). An additional method, the Mixed Effects Model of Evolution (MEME) (Murrell et al. 2012) was also applied. MEME allows the distribution of dN/dS to vary from site to site and from branch to branch at a site, therefore allowing the detection of both pervasive and episodic positive selection; the method has been shown to have more power than methods that assume constant dN/dS across lineages (Murrell et al. 2012). To be conservative, only sites detected using both BEB and MEME were considered targets of positive selection (Fig. 1B); their functional implications are analyzed below.

Different selective pressure among lineages

We next explored possible variations in

selective pressure among different mammals for the five positively selected genes (Fig. 1B, Supplementary Table S4 and S5, Supplementary Fig. S1 and S2, Supplementary Material online).

SI showed the strongest evidence of episodic selection: several positively selected residues were identified for rodents and bats, with microbat also showing positive selection at *MGAM* (Supplementary Table S5, Supplementary Fig. S2, Supplementary Material online). Interestingly, microbat and platypus, the only two lineages that experienced episodic selection at *TREH* (Supplementary Table S5, Supplementary Fig. S1, Supplementary Material online) have a diet that includes trehalose, as these animals feed on insects and crustaceans, respectively.

It was recently suggested that *MGAM* and *SLC5A1* were positively selected in dog. The putative adaptive coding changes are present in the dog reference genome (a boxer) and are accounted for by position M797 and V1001 (dog residues) in MGAM, where a 2 aminoacid C-terminal extension was also noted (Axelsson et al. 2013). Although we did not find evidence of positive selection for any of the analyzed genes in dog (Supplementary Table S5, Supplementary Fig. S1 and S2, Supplementary Material online), we analyzed these residues by taking into account the known phylogeny of mammals and by ancestral state reconstruction at internal nodes (this was not feasible for the C-terminal extension). As shown in Figure 1C, dogs share the M797 and V1001 residues with several related species and these aminoacids represent the ancestral state at most nodes. Inference on the C-terminal extension was more difficult, due to extensive variability in this region; dog shares the 2 aminoacids extension with cat, cow, and alpaca, although with minor differences in these two latter species (Fig. 1C). A similar analysis for the *SLC5A1* putatively selected site (V244) (Axelsson et al. 2013) indicated frequent substitutions at this position, with valine being shared by dog, ferret, and other species (Fig. 1C). Calculation of dN/dS for this position in the whole phylogeny indicated a value of 1.19, close to selective neutrality.

Several positively selected sites impinge on functional protein regions

We detected one positively selected site in the maltase domain of MGAM (A662, Fig. 1B), which is in close spatial proximity to the active site (Sim et al. 2010) (Fig. 2A). Similarly, in the SI isomaltase subunit some lineage-specific positively selected sites were found to be located in nearby the substrate-binding and active sites (Fig. 2B, Supplementary Table S5, Supplementary Material online) (Sim et al. 2010). As for TREH, two of the selected sites we identified, E159 (whole phylogeny) and P287 (microbat) are also in proximity to residues involved in substrate binding (Fig. 2C, Supplementary Table S5, Supplementary Material online).

A part from these sites, most selected residues in MGAM and SI are surface-exposed, with some of them defining continuous surface patches (Supplementary Fig. S4, Supplementary Material online). Moreover, a considerable proportion of positively selected sites maps to the trefoil or P domains (PD, Fig. 1B). The superimposition of the two PDs revealed that the two positively selected sites of MGAM (T979, A987) correspond to T956 and T964, which are positively selected in SI (Fig. 2D).

Although four glycosyl-hydrolase domains of SI and MGAM share limited sequence identity, their 3D structure is remarkably similar. Structural superimposition indicated that, in addition to the trefoil domain, other corresponding regions were targeted by selection (Fig. 3B and C).

In SI, missense mutations responsible for congenital sucrase-isomaltase deficiency (CSID) or identified in chronic lymphocytic leukemia patients (CLL) have been shown to alter the cellular trafficking of the protein, its folding, membrane turnover and localization (Spodsberg et al. 2001; Rodriguez et al. 2013). We noted that mutations R91T (CLL, endoplasmic reticulum accumulation) and Q117R (CSID, mis-sorting to the basolateral membrane) (Spodsberg et al. 2001; Rodriguez et al. 2013) immediately flank positively selected sites (Fig. 3A). 3D mapping and structural comparisons indicated that CSID mutations Q1098P, C1229Y, and W1493C (Propsting et al. 2003; Alfalah et al. 2009; Rodriguez et al. 2013) are located in close spatial proximity to positively selected sites in either SI or MGAM (Fig. 3B and C).

Parallel and divergent evolution of brush-border proteins in humans, chimpanzees, and gorillas

We next applied a population genetics-phylogenetics approach to study the evolution of brush-border genes in the human, chimpanzee, and gorilla lineages. Specifically, we used gammaMap (Wilson et al. 2011) that jointly uses intra-specific variation and inter-specific diversity to estimate the distribution of selection coefficients (γ) along coding regions. gammaMap envisages 12 classes of γ, ranging from strongly beneficial (γ=100) to inviable (γ =-500), with γ equal to 0 indicating neutrality.

We observed a general preponderance of codons evolving under negative selection ($\gamma < 0$) in all genes and in all species. The most striking difference was observed for *SLC5A1*, which showed a preponderance of negative γ values in chimpanzee and to a lesser extent in gorilla, but not in our species, where an appreciable fraction of codons showed γ values higher than 5 (Fig. 3D). We thus used gammaMap to identify specific codons evolving under positive selection (cumulative probability >0.80 of $\gamma \geq 1$) (Supplementary Table S6, Supplementary Material online). Seven positively selected codons were identified for SLC5A1 in humans, none in chimpanzees or gorillas. Whereas two of these (A411 and H615) might have hitchhiked with a regulatory variant (see below), analysis of the remaining sites indicated that E341 and G312 flank one of the transmembrane helices composing the so-called “sugar bundle”, which forms extensive contacts with carbohydrate molecules (Sala-Rabanal et al. 2012) (Fig. 3E). One additional site (L645) is in the immediate vicinity of a C-terminal luminal region that acts as a stereo-specific sugar binding region (Fig. 3E) (Wimmer et al. 2009).

The location relative to 3D structures of other positively selected sites (Supplementary Table S6, Supplementary Material online) are shown in Figures 1B, 3B, 3C, and Supplementary S4.

Pre-agricultural origin of most positively selected alleles

We finally investigated whether natural selection acted on genes involved in carbohydrate digestion/absorption during the recent evolutionary history of human populations. We excluded *LCT* from this analysis, as its selection pattern has been described in detail (Tishkoff et al. 2007). Natural selection leaves signatures that can be detected using appropriate tests.

For instance, the increase in frequency of a selected haplotype (selective sweep) may result in a temporary reduction in the level of genetic variability (measured by θ_W (Watterson 1975) and π (Nei and Li 1979)) and in a shift of the site frequency spectrum, leading to a deficiency of intermediate frequency alleles (indicated by negative values of Tajima's D (Tajima 1989)). Also, a selective sweep may determine an excess of high frequency derived alleles (which can be assessed with the normalized Fay and Wu's H (DH) test (Zeng et al. 2006)) and low nucleotide diversity associated with the derived allele (Barreiro et al. 2009). This latter feature can be searched for using the DIND (Derived Intra-allelic Nucleotide Diversity) test (Barreiro et al. 2009). Thus, using the 1000 Genomes Pilot Project data (1000 Genomes Project Consortium et al. 2010) for Yoruba (YRI), Europeans (CEU), and Chinese plus Japanese (CHBJPT), we estimated nucleotide diversity and Tajima's D (Tajima 1989) over whole gene regions. We also calculated pairwise F_{ST}, an estimate of population genetic differentiation, and performed the DIND test for all SNPs mapping to these genes and in their 50 kb flanks (25 kb up- and down-stream). For all tests statistical significance (in terms of percentile rank) was obtained by deriving empirical distributions*;* coalescent simulations were also performed for the DIND test. We considered genes as positive selection targets if significant results were obtained for the same population in at least two statistics based on different features; we also considered SNPs with a significant DIND test in all populations or with extremely high DIND ranks (>0.999). Moreover, we obtained normalized values for Fay and Wu's H (DH) (Zeng et al. 2006), in sliding windows along the analyzed genomic regions; DH was used as a confirmatory signature but not in the initial detection of selection targets (Supplementary Table S1, Supplementary Material online).

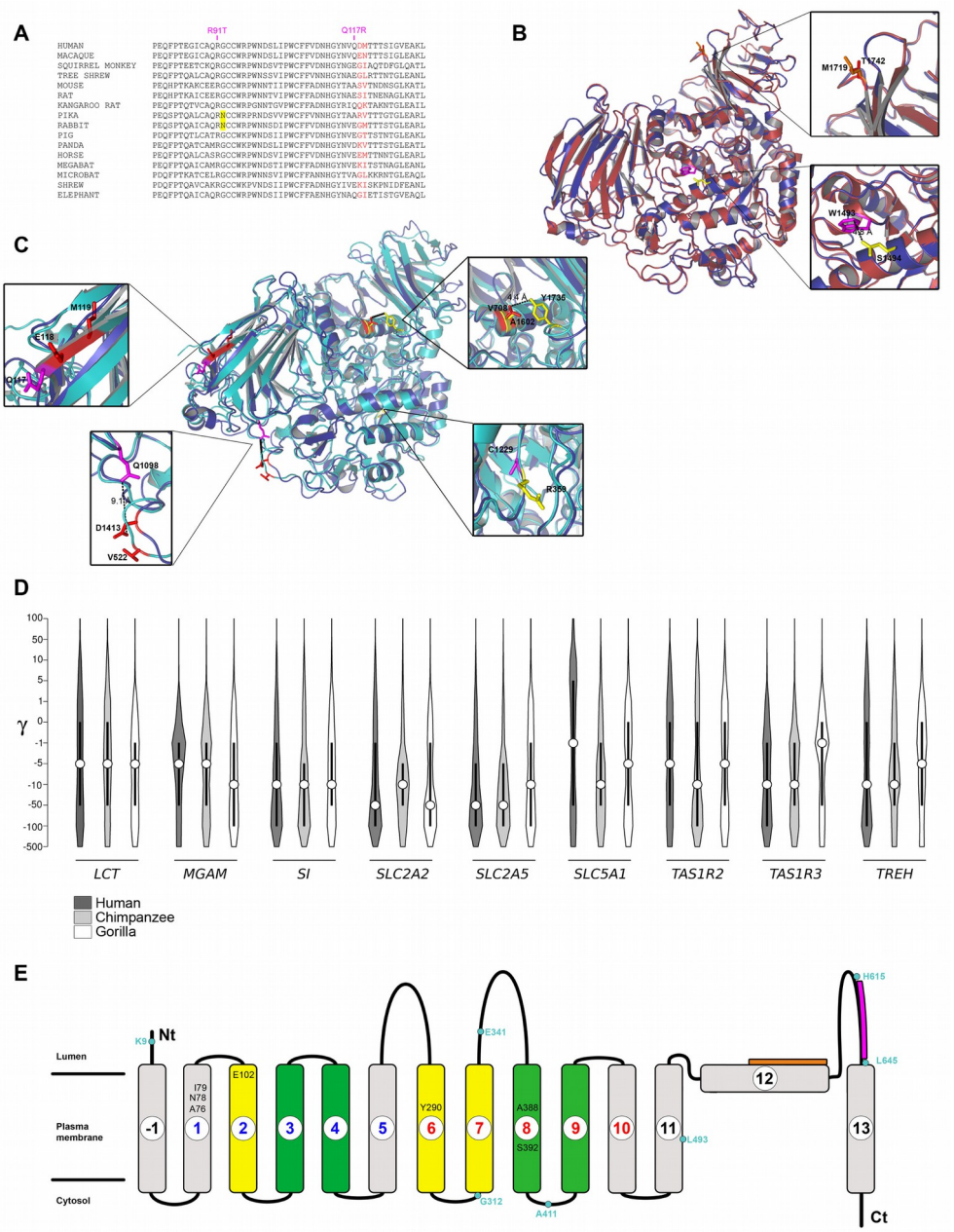


Figure 3. Parallel evolution at MGAM and SI, and lineage-specific selection. (A) Multiple alignment of SI aminoacids 78-130 for a few of representative mammalian species. The location of mutations R91T and Q117R is shown. (B and C) Superimposition of the structure of the sucrase domain (SI, Protein Model Portal code: P14410 Model 2, blue) with glucoamylase (MGAM, PDB code: 3TON, red) (B) and with isomaltase (SI, PDB code: 3LPP , pale blue) (C). Enlargements highlight positively selected sites or residues subjected to pathological mutation located in the corresponding regions of the two different domains. Color codes are as in figure 2A-C. Human missense mutations affecting the protein sorting are reported in magenta. (D) Violin plot of selection coefficients (median, white dot; interquartile range, black bar). Selection coefficients (γ) are classified as strongly beneficial (100, 50), moderately beneficial (10, 5), weakly beneficial (1), neutral (0), weakly deleterious (−1), moderately deleterious (−5, −10), strongly deleterious (−50, −100), and inviable (−500). (E) Topological representation of SLC5A1; transmembrane helices forming the sugar- and hush-bundle are represented in yellow and green, respectively. The location of the stereo-specific and non-stereo-specific binding motifs is shown in magenta and orange, respectively. Positively selected sites in the human lineage are in cyan. Residues in black are involved in sugar or Na^+ binding.

In *SI*, the DIND test detected 2 outlier linked variants in YRI, which also had unusually high F_{ST} (Table 2); rs6788812 represented a DIND outlier in CHBJPT, as well, and was located in a local valley of DH for this population (in line with DH having maximum power for high-frequency sweeps (Zeng et al. 2006)) (Fig. 4A, Table 2). These results suggest that a common selective sweep determined the frequency increase of these variants in Asia and Africa. No selection signal was detected in CEU and analysis of ancient DNA samples indicated that the Denisova and Altai Neandertal (Meyer et al. 2012; Prufer et al. 2014) carry the ancestral allele, whereas a Mesolithic European individual from the La Brana-Arintero site (Olalde et al. 2014) harbors the derived allele at rs6788812 (Table 2, Fig. 5).

In YRI another variant (rs11919067) had an extremely high DIND rank and a linked SNP (rs112446029) represented a DIND outlier, although with lower rank (Table 2); both variants have high DAF (derived allele frequency) in YRI, whereas the derived allele is fixed outside Africa (Table 2). Sliding-window analysis of DH in YRI detected a local valley where rs11919067 is located (Fig. 4A). Overall, these results suggest that a selective sweep drove the frequency increase of these variants in all populations and that the process is complete in non-Africans. Interestingly, rs11919067 and rs112446029 have been cataloged in a list of "modern-human-specific sites"- i.e. positions where the Denisova or Altai Neandertal sequences display the ancestral allele, whereas most modern humans carry the derived allele (Prufer et al. 2014) (Table 2, Fig. 5). The catalog also includes rs9917722 (T1802S), which we identified in the gammaMap analysis (Table 2, Supplementary Table S6, Supplementary Material online). Analysis of all modern-human-specific sites in *SI* (Fig. 4A) indicated that they mainly cluster in two regions, one where rs9917722 and rs6788812 are located, and the other encompassing rs11919067 and rs112446029. In YRI rs9917722 shows no linkage disequilibrium (LD) with rs11919067 and rs6788812 (r^2=0.003 and 0.085, respectively). Overall, these data suggest that distinct selective events, have occurred at SI after the modern-human lineage split from the common ancestor with Denisovans and Neandertals. Interestingly, analysis of an Upper Paleolithic sample from Siberia (Raghavan et al. 2014) indicated that this individual carried the derived allele at rs11919067, rs112446029, and rs9917722 (Fig. 5).

Signals of positive selection in all populations were also detected at another brush-border enzyme, TREH. Indeed, the same *TREH* variant (rs527619) was identified as a DIND outlier in all three populations, although with different DAF (Table 2, Fig. 4B).

As for transporters, *SLC5A1* showed reduced nucleotide diversity in CHBJPT and low Tajima's D in CEU and CHBJPT (Supplementary Table S7, Supplementary Material online). The DIND test detected five linked outlier variants in CEU with a DAF of 0.94; the derived allele is fixed in YRI and AS (Table 2). The SNPs are in a local valley of DH in Europeans, and a very local and limited reduction in DH was also observed in YRI (DH loses power at sweep completion) (Fig. 4C).

Four of the five *SLC5A1* SNPs we detected are listed in the modern-human-specific site catalog, which also includes rs17683430 (A411T, detected by gammaMap) (Table 2, Supplementary Table S6, Supplementary Material online). Analysis of modern-human-specific sites along the *SLC5A1* gene indicated that they are scattered across a relatively large region with a clustering around the five variants detected (Fig. 4C); in Europeans these are in tight LD with rs17683430 and with rs33954001 (also detected by gammaMap, Supplementary Table S6, Supplementary Material online) (r^2>0.86), suggesting these SNPs hitchhiked to high frequency due to LD with one of the DIND outlier variants. Analysis of the Mesolithic and Paleolithic samples (Olalde et al. 2014; Raghavan et al. 2014) revealed that the derived allele was already present at all selected variants (Fig. 5) .

SLC2A2 also showed low Tajima's D values in CHBJPT (Supplementary Table S7, Supplementary Material online). Four variants were DIND outliers in CHBJPT and displayed unusually high F_{ST} in the YRI/CHBJPT comparison. The variants have high DAF in CHBJPT (Table 2) and are located in a local DH valley, strongly supporting selective sweep has occurred in Asian populations (Fig. 4D). Interestingly, in CEU the four variants are in tight LD (r^2>0.9) with two GWAS SNPs

Table 2
Candidate Targets of Positive Selection in Human Populations

Gene	SNP ID	Derived Allele[a]	DAF[b]			DIND Rank (population)	DIND *P* Value[c] (population)	F_{ST} Rank (comparison)	Notes
			YRI	CEU	CHBJPT				
SI	rs41273563	C	0.32	0.89	0.87	0.97 (YRI)	0.031 (YRI)	0.96 (YRI/CHBJPT)	
	rs11919067	C	0.98	1	1	>0.999 (YRI)	<0.001 (YRI)	—	Modern-human-specific site
	rs112446029	A	0.98	1	1	0.99 (YRI)	0.008 (YRI)	—	Modern-human-specific site
	rs6788812	G	0.32	0.89	0.87	0.98 (YRI) 0.98 (CHBJPT)	0.024 (YRI) 0.044 (CHBJPT)	0.95 (YRI/CEU)	
	rs9917722	G	0.85	1	1	—		—	Modern-human-specific site; identified by gammaMap
TREH	rs527619	A	0.52	0.42	0.27	0.98 (YRI), 0.97 (CEU), 0.99 (CHBJPT)	0.007 (YRI), 0.073 (CEU), 0.081 (CHBJPT)	—	
SLC5A1	rs117628874	T	1	0.94	1	0.95 (CEU)	0.012 (CEU)	—	Modern-human-specific site
	rs74399071	G	1	0.94	1	0.96 (CEU)	0.012 (CEU)	—	Modern-human-specific site
	rs79022443	T	1	0.94	1	0.97 (CEU)	0.010 (CEU)	—	
	rs78578916	G	1	0.94	1	0.98 (CEU)	0.002 (CEU)	—	Modern-human-specific site
	rs2899174	T	1	0.94	1	0.95 (CEU)	0.012 (CEU)	—	Modern-human-specific site
	rs17683430	G	1	0.94	1	—		—	Modern-human-specific site; identified by gammaMap
SLC2A2	rs11720640	G	0.64	0.86	0.96	0.99 (CHBJPT)	0.047 (CHBJPT)	0.95 (CHBJPT/YRI)	In LD with rs11920090 and rs10513686 (GWAS)
	rs1905504	T	0.64	0.86	0.96	>0.999 (CHBJPT)	0.047 (CHBJPT)	0.95 (CHBJPT/YRI)	In LD with rs11920090 and rs10513686 (GWAS)
	rs7635100	G	0.64	0.86	0.96	>0.999 (CHBJPT)	0.047 (CHBJPT)	0.95 (CHBJPT/YRI)	In LD with rs11920090 and rs10513686 (GWAS)
	rs6780208	A	0.64	0.86	0.96	>0.999 (CHBJPT)	0.047 (CHBJPT)	0.95 (CHBJPT/YRI)	In LD with rs11920090 and rs10513686 (GWAS)
	rs78425790	T	0.90	0.98	1	0.97 (YRI)	0.028 (CHBJPT)	0.99 (YRI/CEU), 0.99 (YRI/CHBJPT)	
	rs7649712	G	0.90	0.98	1	0.98 (YRI)	0.026 (YRI)	0.99 (YRI/CEU), 0.99 (YRI/CHBJPT)	
	rs75513459	G	0.90	0.98	1	0.97 (YRI)	0.031 (YRI)	0.99 (YRI/CEU), 0.99 (YRI/CHBJPT)	
	rs79438006	T	0.90	0.98	1	0.96 (YRI)	0.034 (YRI)	0.99 (YRI/CEU), 0.99 (YRI/CHBJPT)	
	rs75975268	C	0.90	0.98	1	0.97 (YRI)	0.030 (YRI)	0.99 (YRI/CEU), 0.99 (YRI/CHBJPT)	
	rs74828611	A	0.90	0.98	1	0.98 (YRI)	0.029 (YRI)	0.99 (YRI/CEU), 0.99 (YRI/CHBJPT)	
SLC2A5	rs875996	A	0	0.18	0.16	0.97 (CEU)	0.014 (CEU)	0.99 (YRI/CEU)	In LD with rs113568511 (eQTL)
	rs34605482	T	0	0.18	0.17	0.95 (CEU)	0.034 (CEU)	0.99 (YRI/CEU)	In LD with rs113568511 (eQTL)

[a]To avoid misattribution (Hernandez et al. 2007), the derived allele was inferred by parsimony through incorporating sequence information for at least four primate species.
[b]DAF.
[c]*P* value calculated by coalescent simulations.

(rs11920090 and rs10513686) associated with fasting glucose-related traits and gamma-glutamyl transferase (GGT) levels (Dupuis et al. 2010; Chambers et al. 2011; Manning et al. 2012). Both the Mesolithic and the Paleolithic samples carried the derived allele at most SNPs (Fig. 5). Thus, in analogy to the *SLC5A1* and *SI* variants, the selected haplotype was present in the Paleolithic (Fig. 5).

In *SLC2A2*, the DIND test also detected six outliers in YRI, which also display high F_{ST} values (Table 2). These variants have a DAF of 0.90 in YRI and fall in a DH valley (Fig. 4D); the derived allele is fixed or almost fixed in non-Africans, suggesting a complete sweep that predated the split of modern humans from Neandertal and Denisovans, as these hominins also carry the derived alleles (Fig. 5).

Finally, in *SLC2A5* two DIND outlier variants in CEU also displayed a high F_{ST} ranks (Table 2), suggesting that a selective sweep has occurred in Europeans. The two variants are in LD (r^2= 0.76) with rs113568511, identified as an eQTL (expression quantitative trait locus) for *SLC2A5* in lymphoblastoid cell lines (Lappalainen et al. 2013).

Several selected variants we detected map within ENCODE functional elements (Fig. 4). Overall, we analyzed 8 genes (*LCT* was omitted) and we found one with a significant DIND test for the same variant in three populations (*TREH*) and three with at least two variants showing outlier values both for the DIND and F_{ST} tests (*SLC2A2*, *SI*, and *SLC2A5*) (Table 2). To obtain an estimate of whether these findings are unusual and of the incidence of false positives, we adopted a resampling approach. Specifically, we drew 100 samples of 8 randomly selected genes and we calculated the DIND tests and F_{ST} for all variants mapping to these genes. For each sample we counted the number of positively selected genes, defined as those carrying at least one variant with significant DIND test in three populations or at least two variants showing outlier values both for the DIND and F_{ST} tests in the same population. Results indicated that the probability of drawing a set of genes showing the same or a higher number of selected genes as those in the brush-border set is 0.02.

Discussion

Adaptive evolution in mammals

We explored the evolutionary history of genes encoding brush-border proteins involved in carbohydrate digestion and absorption. This decision was based on the well-accepted concept that the availability of food resources is a driver of pivotal importance in evolution in mammals and that individual mammalian lineages might have adapted to specialized diets (e.g. insects, crustaceans) or lifestyles (e.g. flight).

We found evidence of positive selection at the four brush-border enzymes, indicating stronger selective pressure compared to transporters and taste receptors. Episodic positive selection was also detected for several mammalian lineages. Whereas for *TREH* two positively selected lineages (microbat and platypus) have a diet that includes trehalose, mammals showing evidence of positive selection at *MGAM* and *SI* display different food habits. Thus, as previously reported for *TAS1R2* (Zhao et al. 2010; Jiang et al. 2012), inference of the underlying selective pressures remains uncertain. Nonetheless, we detected positive selection at *SI* in both bat species (megabat is frugivorous, microbat insectivore), with microbat also showing selection signatures at *MGAM* and *TREH*. As an adaptation to flight, bats generally display a reduced small intestinal nominal surface area compared to non-flying mammals, and resort to higher sugar paracellular absorption as a compensation (Caviedes-Vidal et al. 2007). Because polysaccharides require digestion before they can be metabolized, fast and efficient digestion of complex sugars would be strongly advantageous in these species, which daily ingest large amounts of food (up to 50% of their body weight) to meet energy requirements. Whether positive selection at *SI* and *MGAM* is part of a more general adaptation to flight in these animals remains an interesting possibility worth further investigation.

Positively selected sites in enzyme encoding genes

The rate of starch-generated glucose depends on the activity of MGAM and SI, which have complementary substrate specificity in humans

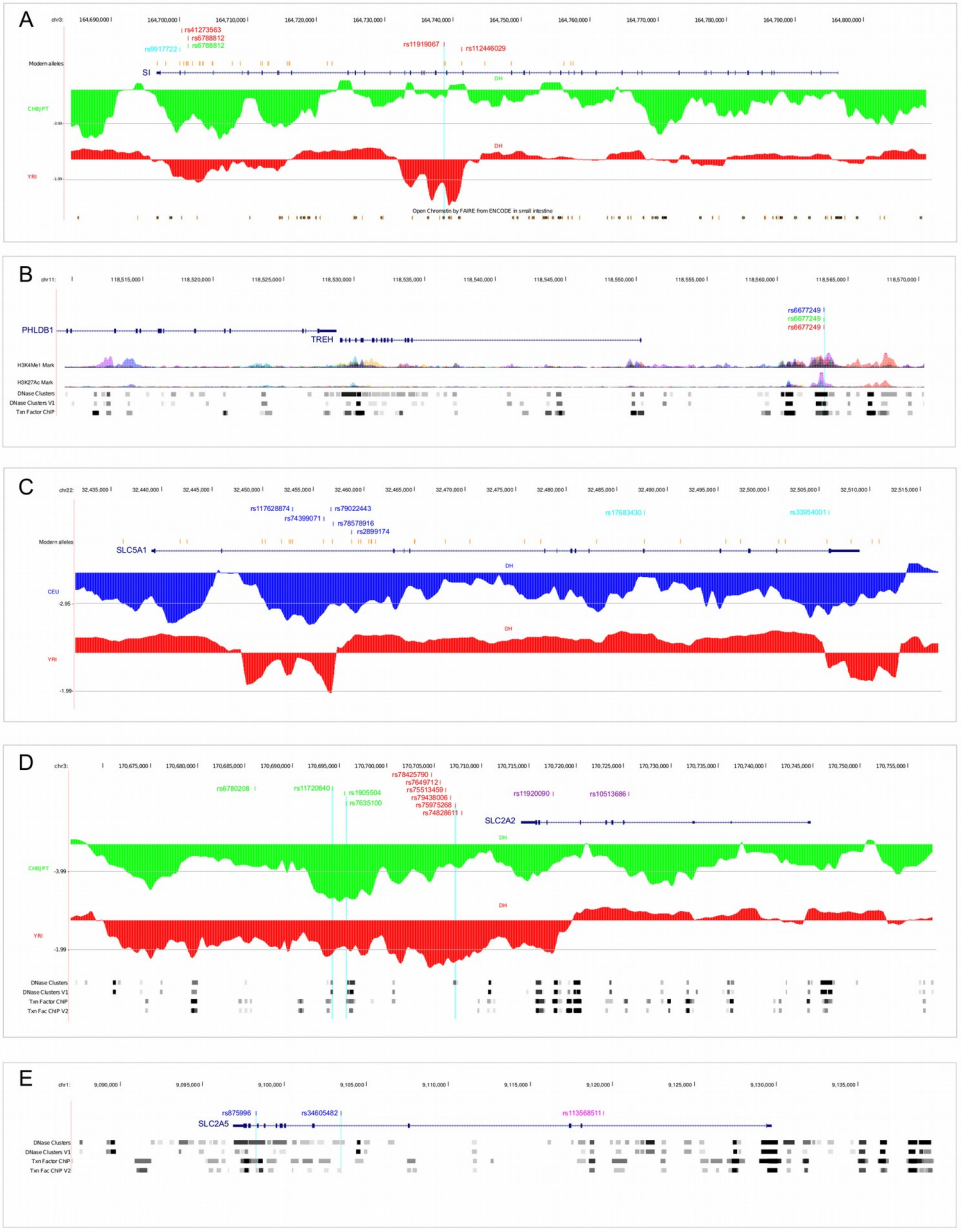


Figure 4. Location of the most likely selection targets. Candidate targets in human populations and their genomic locations (GRCh37/hg19) are shown for *SI* (A), *TREH* (B), *SLC5A1* (C), *SLC2A2* (D), *SLC2A5* (E) within the UCSC Genome Browser view. Relevant ENCODE annotation tracks are shown as gray horizontal shading or colored peaks in case of histone marks. Candidate selection targets falling in putative regulatory regions are indicated with cyan vertical lines. For *SI*, *SLC5A1*, and *SLC2A2* a sliding-window analysis of DH is also shown in green (YRI), red (CHBJPT) or blue (CEU). The gray horizontal line represents the 5th percentile of DH. Variants in blue, red and green represent selection targets in CEU, CHBJPT, and YRI, respectively. The location of variants cataloged as modern-human-specific sites are shown in orange. Additional color codes are as follows: cyan, positively selected sites in the human lineage detected by gammaMap; violet, GWAS SNPs; magenta, eQTL.

(Sim et al. 2010). In a few instances we found the corresponding residues of MGAM and SI to be targeted by selection, indicating an important role for these sites. An interesting possibility is that some selected sites in SI and MGAM evolved to hone the folding, cellular trafficking, and membrane turnover of these enzymes, depending on specific molecular (e.g. interaction with chaperones) or physiological (e.g. body temperature) features of distinct mammals. In fact, some of the identified selected sites are located in close spatial proximity to SI missense mutations that affect the enzyme's post-translational fate, sometimes showing temperature-sensitive effects (Propsting et al. 2003; Alfalah et al. 2009; Rodriguez et al. 2013). In analogy, an LCT missense mutation associated with congenital lactase deficiency (G1363S), has been shown to alter protein trafficking and folding, partially depending on temperature (Behrendt et al. 2009).

Adaptive events in primates and human populations

Our study was also motivated by the observation that one of the most important turning-points of human history, the introduction of agriculture, resulted in a dietary shift in terms of carbohydrate intake. In this respect, the availability of genetic information for other primates and for pre-agricultural human populations allows the opportunity to address the tempo and mode of evolution for genes involved in carbohydrate digestion and absorption.

A notable observation is the different evolutionary fate of *SLC5A1* in humans vs. chimpanzees and gorillas. Still, we note that, whereas some sites positively selected in the human *SLC5A1* gene are likely involved in sugar binding, the signal we detected is partially accounted for by hitchhiking of coding variants with the intronic positive selection target(s), as population genetic analysis indicated.

Integration of different tests can improve the power to detect selective sweeps and, importantly, allows identification of the causal variant(s) (Grossman et al. 2013). Our approach includes the DIND test, which is powerful in most DAF ranges (Barreiro et al. 2009; Fagny et al. 2014) and less sensitive than iHS to low genotype quality or low coverage (i.e. it is well suited for the 1000G data) (Fagny et al. 2014). DIND results were combined with pairwise F_{ST} analyses and nucleotide diversity or Tajima's D, whereas DH (Zeng et al. 2006) was calculated in sliding-windows to account for local events and, for this reason, used as an *a posteriori* validation. These analyses indicated that five out of the nine genes we analyzed have been targeted by selection during the history of human populations, with *SI* and *SLC2A2* having experienced distinct events targeting different variants. The majority of sweeps we detected occurred in all analyzed populations, although in some instances they have reached completion (e.g. *SI* and *SLC2A2* in non-Africans and *SLC5A1* in non-Europeans) or proceeded with different timing/strength (e.g. *TREH*).

The availability of an increasing number of ancient DNA sequences allows the unprecedented opportunity to define the time in human history when selection operated, in turn providing information on the possible selective pressures. Based on the sequencing of a Denisova and a Neandertal individual, and on allele frequency in extant human populations, Prufer et al (Prufer et al. 2014) compiled a list of modern-human-specific-alleles, suggested to represent changes that were most important during the recent evolutionary history of our species. Results herein indicate that modern alleles at *SLC5A1* and *SI* were indeed driven to high frequency by natural selection in human populations. Nevertheless, most of these positively selected modern alleles were already present in the Mesolithic and Paleolithic and, therefore, predate the emergence of agriculture. Whether the onset of selection occurred before the Paleolithic or these alleles segregated as neutral standing variation in these early populations remains to be evaluated, possibly through the sequencing of additional ancient samples. Although with uncertainty due to possible gene conversions, the initial expansion of the *AMY1* copy number was dated around 200,000 years ago, a time frame that might coincide with the introduction of starch-rich underground storage organs (USOs) as food sources in hominin diet (Perry et al. 2007). USOs are thought to have played an important

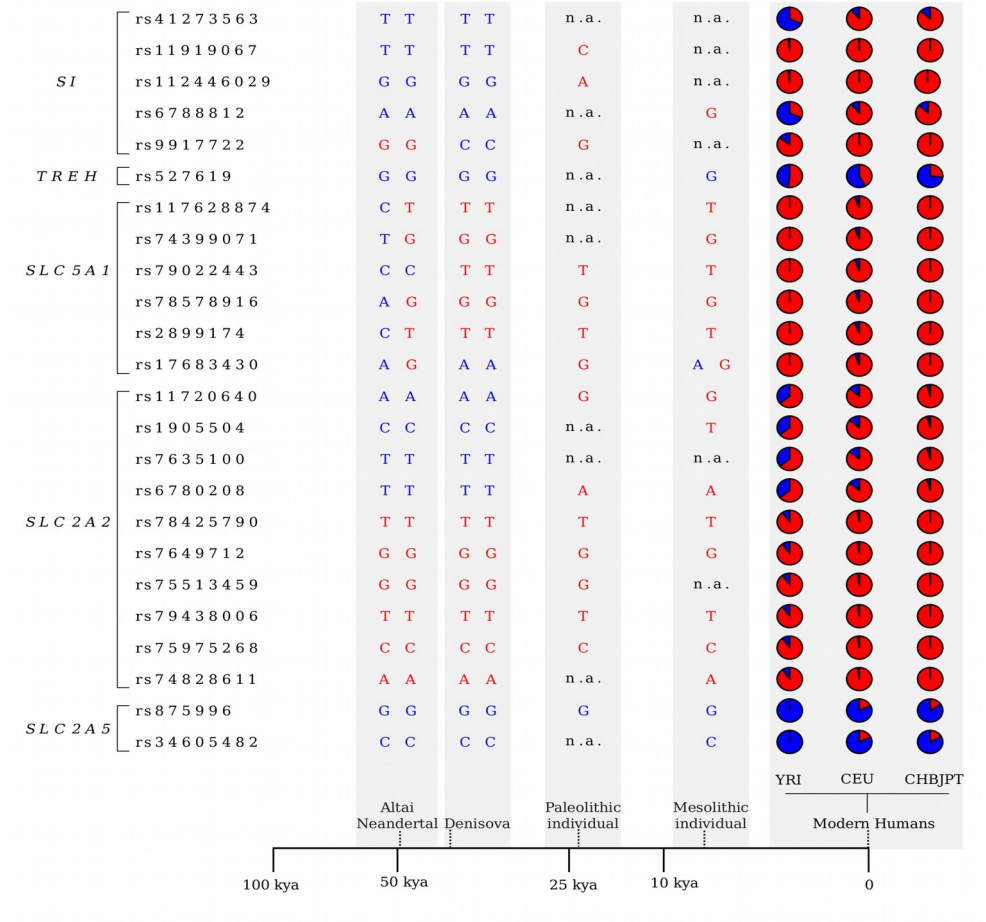


Figure 5. Positively selected variants in human populations. Genotype data are shown for a Neandertal, a Denisova, an Upper Paleolithic Siberian, and a Mesolithic hunter-gatherer; allele frequencies are shown for modern human populations (pie-charts). Only one allele is reported when coverage was not sufficient for genotype inference. Blue and red colors indicate ancestral and derived alleles, respectively. A temporal line with the approximate ages of the individuals is also reported (kya: thousands of years ago).

role in human evolution (Laden and Wrangham 2005). Thus, agriculture might have spurred the frequency increase of variants that were already weakly adaptive in hunter-gatherers, resulting in a continuum rather than an abrupt onset of selective events. A similar concept has been proposed for traits unrelated to diet (Olalde et al. 2014).

As for the more recent selective event at *SLC2A5*, it is worth noting that some degree of fructose intolerance is widespread in humans, and fructose absorption is increased by the co-ingestion of glucose and is reduced by the presence of sorbitol (Skoog and Bharucha 2004). Thus, selection at *SLC2A5* might have been driven by the domestication in temperate areas of fruit crops (e.g. apples and pears) that

contain excess fructose plus sorbitol (Skoog and Bharucha 2004). Clearly, it would be extremely interesting to test whether the positively select variant identified herein (and which is in LD with an eQTL), modulates fructose absorptive capacity.

Selection targets in regulatory regions

In analogy to the well-known selection targets at the *LCT* locus (Tishkoff et al. 2007), the selection signatures we identified in human populations target non-coding polymorphisms, supporting the view that most adaptive changes affect regulatory elements (Grossman et al. 2013). We suggest that regulatory variants may also represent the selection target at the dog *MGAM* and *SLC5A1* genes. Although the analyses we performed were not specifically devised to search for recent selective events in dogs, and surely lack power in this respect, the candidate coding variants Axelsson et al. (Axelsson et al. 2013) proposed can be analyzed within the framework of the known mammalian phylogeny. Overall, these analyses suggest that coding variants are not likely selection targets in the canine *MGAM* and *SLC5A1* genes, in line with the observation that the expression of *MGAM* is higher in dogs compared to wolves (Axelsson et al. 2013).
Deeper understanding of the evolutionary processes associated with human dietary shifts is expected to provide valuable information concerning the susceptibility of human populations to metabolic diseases. Data herein indicate that the selection targets at *SLC2A2* are in phase with the risk allele for fasting glucose levels and with the non-risk allele for GGT levels. This opens the question as to whether the disease alleles hitchhiked with the selected variant, or might be accounted for by the selected haplotype. In either case, further analyses will be required to determine which phenotype selection acted upon.

Acknowledgments

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

References

1000 Genomes Project Consortium, et al. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061-1073.

Alfalah M., M. Keiser, T. Leeb, K. P. Zimmer, and H. Y. Naim. 2009. Compound heterozygous mutations affect protein folding and function in patients with congenital sucrase-isomaltase deficiency. Gastroenterology 136:883-892.

Anisimova M., and Z. Yang. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol. Biol. Evol. 24:1219-1228.

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

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

Axelsson E., et al. 2013. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature 495:360-364.

Barreiro L. B., et al. 2009. Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. PLoS Genet. 5:e1000562.

Behrendt M., M. Keiser, M. Hoch, and H. Y. Naim. 2009. Impaired trafficking and subcellular localization of a mutant lactase associated with congenital lactase deficiency. Gastroenterology 136:2295-2303.

Bielawski J. P., and Z. Yang. 2003. Maximum likelihood methods for detecting adaptive evolution after gene duplication. J. Struct. Funct. Genomics 3:201-212.

Capella-Gutierrez S., J. M. Silla-Martinez, and T. Gabaldon. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972-1973.

Caviedes-Vidal E., et al. 2007. The digestive adaptation of flying vertebrates: high intestinal

paracellular absorption compensates for smaller guts. Proc. Natl. Acad. Sci. U. S. A. 104:19132-19137.

Cereda M., M. Sironi, M. Cavalleri, and U. Pozzoli. 2011. GeCo++: a C++ library for genomic features computation and annotation in the presence of variants. Bioinformatics 27:1313-1315.

Chambers J. C., et al. 2011. Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. Nat. Genet. 43:1131-1138.

Delport W., A. F. Poon, S. D. Frost, and S. L. Kosakovsky Pond. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 26:2455-2457.

Dupuis J., et al. 2010. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat. Genet. 42:105-116.

Engelken J., et al. 2014. Extreme population differences in the human zinc transporter ZIP4 (SLC39A4) are explained by positive selection in Sub-Saharan Africa. PLoS Genet. 10:e1004128.

Fagny M., et al. 2014. Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. Mol. Biol. Evol. 31:1850-1868.

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

Freedman A. H., et al. 2014. Genome sequencing highlights the dynamic early history of dogs. PLoS Genet. 10:e1004016.

Grossman S. R., et al. 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327:883-886.

Grossman S. R., et al. 2013. Identifying recent adaptations in large-scale genomic data. Cell 152:703-713.

Guindon S., F. Delsuc, J. F. Dufayard, and O. Gascuel. 2009. Estimating maximum likelihood phylogenies with PhyML. Methods Mol. Biol. 537:113-137.

Hancock A. M., et al. 2010. Colloquium paper: human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. Proc. Natl. Acad. Sci. U. S. A. 107 Suppl 2:8924-8930.

Jiang P., et al. 2012. Major taste loss in carnivorous mammals. Proc. Natl. Acad. Sci. U. S. A. 109:4956-4961.

Karasov W. H., C. Martinez del Rio, and E. Caviedes-Vidal. 2011. Ecological physiology of diet and digestive systems. Annu. Rev. Physiol. 73:69-93.

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

Kosakovsky Pond S. L., and S. D. Frost. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. 22:1208-1222.

Kosakovsky Pond S. L., et al. 2011. A random effects branch-site model for detecting episodic diversifying selection. Mol. Biol. Evol. 28:3033-3043.

Laden G., and R. Wrangham. 2005. The rise of the hominids as an adaptive shift in fallback foods: plant underground storage organs (USOs) and australopith origins. J. Hum. Evol. 49:482-498.

Lappalainen T., et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501:506-511.

Manning A. K., et al. 2012. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nat.

Genet. 44:659-669.

Meyer M., et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science 338:222-226.

Murrell B., et al. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 8:e1002764.

Naumov D. G. 2007. Structure and evolution of mammalian maltase-glucoamylase and sucrase-isomaltase genes. Mol. Biol. (Mosk) 41:1056-1068.

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

Olalde I., et al. 2014. Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. Nature 507:225-228.

Perry G. H., et al. 2007. Diet and the evolution of human amylase gene copy number variation. Nat. Genet. 39:1256-1260.

Pollard K. S., et al. 2006. Forces shaping the fastest evolving regions in the human genome. PLoS Genet. 2:e168.

Prado-Martinez J., et al. 2013. Great ape genetic diversity and population history. Nature 499:471-475.

Propsting M. J., R. Jacob, and H. Y. Naim. 2003. A glutamine to proline exchange at amino acid residue 1098 in sucrase causes a temperature-sensitive arrest of sucrase-isomaltase in the endoplasmic reticulum and cis-Golgi. J. Biol. Chem. 278:16310-16314.

Prufer K., et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505:43-49.

Raghavan M., et al. 2014. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 505:87-91.

Rodriguez D., et al. 2013. Functional analysis of sucrase-isomaltase mutations from chronic lymphocytic leukemia patients. Hum. Mol. Genet. 22:2273-2282.

Sala-Rabanal M., et al. 2012. Bridging the gap between structure and kinetics of human SGLT1. Am. J. Physiol. Cell. Physiol. 302:C1293-305.

Schaffner S. F., et al. 2005. Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 15:1576-1583.

Sim L., et al. 2010. Structural basis for substrate selectivity in human maltase-glucoamylase and sucrase-isomaltase N-terminal domains. J. Biol. Chem. 285:17763-17770.

Skoog S. M., and A. E. Bharucha. 2004. Dietary fructose and gastrointestinal symptoms: a review. Am. J. Gastroenterol. 99:2046-2050.

Spodsberg N., R. Jacob, M. Alfalah, K. P. Zimmer, and H. Y. Naim. 2001. Molecular basis of aberrant apical protein transport in an intestinal enzyme disorder. J. Biol. Chem. 276:23506-23510.

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

Thornton K. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics 19:2325-2327.

Tishkoff S. A., et al. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. Nat. Genet. 39:31-40.

Vilella A. J., et al. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 19:327-335.

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

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

Wilson D. J., R. D. Hernandez, P. Andolfatto, and M. Przeworski. 2011. A population genetics-phylogenetics approach to inferring natural selection in coding sequences. PLoS Genet. 7:e1002395.

Wimmer B., M. Raja, P. Hinterdorfer, H. J. Gruber, and R. K. Kinne. 2009. C-terminal loop 13 of Na+/glucose cotransporter 1 contains both stereospecific and non-stereospecific sugar interaction sites. J. Biol. Chem. 284:983-991.

Wright S. 1950. Genetical structure of populations. Nature 166:247-249.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586-1591.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555-556.

Yang Z., and R. Nielsen. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. 46:409-418.

Yang Z., W. S. Wong, and R. Nielsen. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22:1107-1118.

Zeng K., Y. X. Fu, S. Shi, and C. I. Wu. 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 174:1431-1439.

Zhang J., R. Nielsen, and Z. Yang. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22:2472-2479.

Zhao H., et al. 2010. Evolution of the sweet taste receptor gene Tas1r2 in bats. Mol. Biol. Evol. 27:2642-2650.

3.2 Adaptations to pathogens

3.2.1 Adaptation of genes involved in the immune response

3.2.1.1 OASes and STING: adaptive evolution in concert

In this work I analyzed the evolutionary history of the MB21D1-TMEM173 and OAS-RNASEL axes. OAS (2'-5'-oligoadenylate synthases) proteins and cyclic GMP-AMP synthase (cGAS, gene symbol: *MB21D1*) patrol the cytoplasm for the presence of foreign nucleic acids (based on their similarities, cGAS and OAS proteins may be considered as a novel family of pattern-recognition receptors, PRRs). Upon binding to dsRNA or dsDNA, OAS proteins and cGAS produce nucleotide second messengers to activate RNase L and STING (stimulator of interferon genes, gene symbol: *TMEM173*), respectively, and initiate antiviral programs. Therefore, these molecules represent central elements of the innate antiviral immune response. Indeed, all of them are constantly involved in genetic conflicts with pathogens and, as a consequence, are commonly thought to be targeted by positive selection.

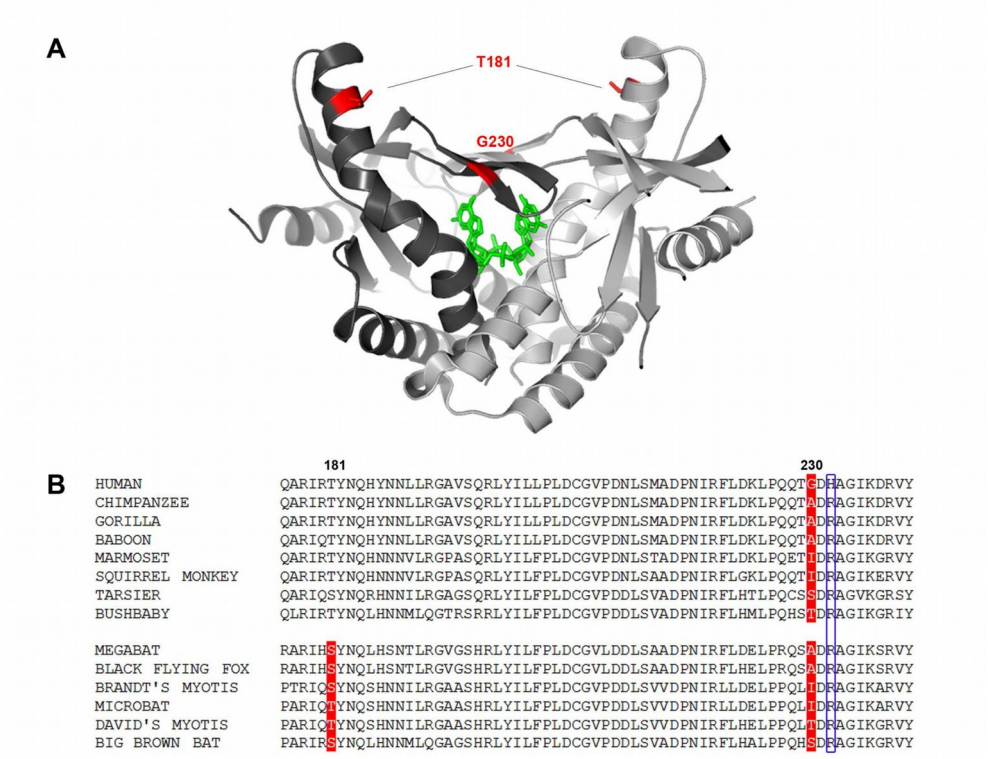
Species-specific differences in the function of PRRs or of their downstream effectors may be common. In the case of TLRs, for instance, the same receptor in distinct species may recognize different ligands or the same ligand with different affinity [115]. Therefore, studying the pattern of inter-species evolution may provide valuable information on the differential susceptibility to infection within and among species. Primates, for example, show marked differences in the susceptibility and severity of several viral infections including those caused by HIV/SIV, HCV, HBV, and Varicella-zoster virus [116, 117]. Moreover, several emerging and re-emerging viral diseases affecting humans originate through the zoonotic transmission from a reservoir animal host [118]. Bats (Chiroptera), as instance, have long been known to harbor and disseminate a wide range of viruses that are

highly pathogenic for humans. On the one hand, this observation suggests that bats have been co-evolving with viruses for a long time and have adapted to high viral exposure. On the other hand, the wide variety of viral families hosted by bats indicates that adaptation most likely involved genes with a role in immune response, rather than molecules acting as incidental viral receptors (as different viruses use distinct strategies to invade the host). Thus, innate immunity genes that are devoted to antiviral response represent excellent candidates as adaptive selection targets in Chiroptera.

I therefore performed an in-depth analysis of the evolutionary history of 7 genes (*OAS1-3*, *OASL*, *MB21D1*, *TMEM173*, and *RNASEL*) in primates and bats. I applied conservative maximum-likelihood models [13, 71, 75] and identified sites and lineages subject to positive selection by the intersection of different methods, to assure reliability [14-16]. Positively selected sites were mapped onto 3D structures in order to understand their functional significance. Furthermore, I performed an evolutionary analysis in the human, chimpanzee, and gorilla lineages using a population genetics-phylogenetics approach [19].

Results indicated widespread evidence of adaptive evolution in both primates and bats, with several genes targeted by positive selection in both mammalian orders. In Chiroptera selective pressure was comparatively stronger for STING and RNase L than for pattern recognition receptors (OASes and cGAS). Several positively selected sites were found to be located in functionally relevant protein regions. As an example, position 230 in STING, a major determinant of response to natural ligands and to mimetic drugs (e.g. DMXAA), was found to be positively selected in both the primate and bat phylogenies. Several positively selected sites were found to be located in functionally relevant protein regions. For instance, T181 immediately flanks the second ER (endoplasmic reticulum) retention

signal in the protein sequence (Fig 3A-B). Also, position 230 in *TMEM173* was found to be positively selected in both the primate and bat phylogenies. This site lies in the flexible loop that acts as a lid above the cyclic dinucleotide binding pocket of the receptor. Substitutions at this site greatly affect the response to natural ligands and to mimetic drugs, such as DMXAA [119-121] (Fig 3). In humans, positions 230 and 232 are polymorphic (G230A and H232R). Different alleles at these sites affect STING binding specificity for different substrates, including the canonical 3'-5' cyclic dinucleotides, known to be synthesized by bacteria, and the noncanonical [G(2'-5')pA(3'-5')p] cyclic dinucleotide, that contains a single 2'-5' phosphodiester bond and is produced by mammalian cGAS (Fig 3B-C) [122-125].



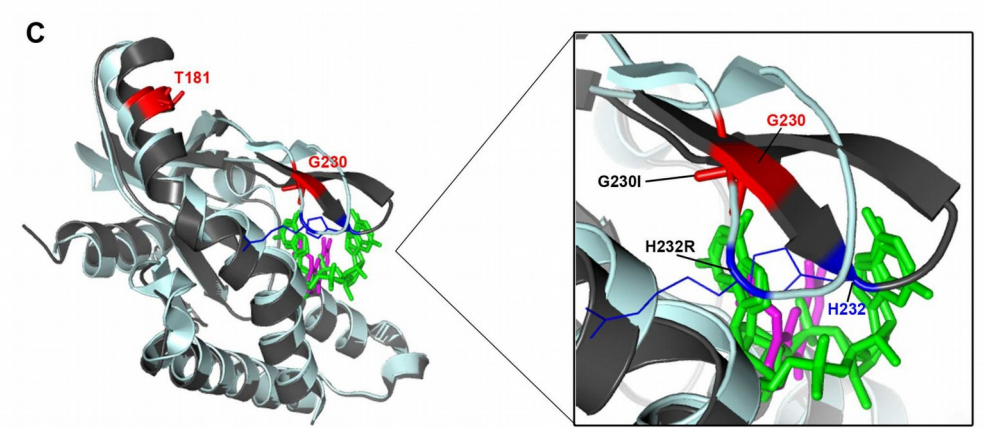


Figure 3: A. Positively selected sites mapped onto the human STING dimeric structure in complex with [G(2',5')pA(3',5')p] (green) (PDB code: 4LOH). The two monomers are colored in dark and light gray. Positively selected sites in both orders are in red. B. Multiple alignment of cGAS amino acids 176-240 for a few of representative primate and bat species. Positively selected in primates and/or bats are in red; position 232 is boxed in blue. C. Superimposition of the structure of the wt STING monomer (dark grey) in complex with [G(2',5')pA(3',5')p]) (green) (PDB code: 4LOH) and the STING double mutant (G230I, H232R) (pale cyan) in complex with DMXAA (magenta) (PDB code: 4QXP). The different conformation of the loop covering the dNTPs binding site is enlarged. [G(2',5')pA(3',5')p] and DMXAA are represented as sticks.

In OAS1, OAS2, and cGAS analysis of positively selected sites and superimpostion of 3D structures revealed parallel evolution, with the corresponding residues selected in different genes (Fig 4).

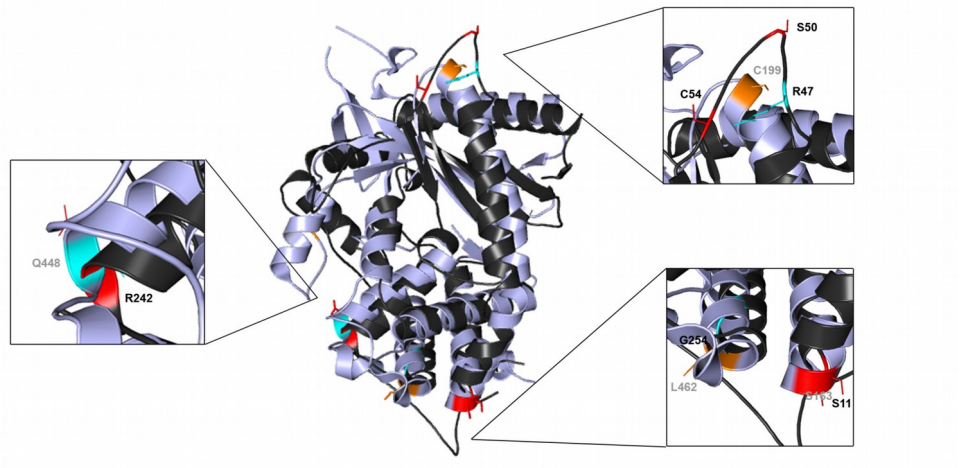


Figure 4: Superimposition of the structure of the cGAS (PDB code: 4O67, light grey) and OAS1 (PDB code: 4IG8, black). Enlargements highlight positively selected sites located in the corresponding regions of the two different enzymes. Color codes are as follows: red, positively selected sites in the primate phylogeny; orange, positively selected sites in the bat phylogeny; yellow, lineage-specific sites; cyan, positively selected sites in the chimpanzee lineage; blue, positively selected sites in the gorilla lineage; green, positively selected sites in more than one lineage among human, chimpanzee and gorilla.

Because this cannot result from gene conversion, it is suggested that selective pressure acting on *OAS* and *MB21D1* genes is related to nucleic acid recognition and to the specific mechanism of enzyme activation, which requires a conformational change.

As for RNASEL, I extended a previous analysis [126] with the inclusion of additional primate sequences and by incorporating information on the protein 3D structure [127]. Interestingly, in primates and bats selection targeted an α-helix/loop element in RNase L that modulates the enzyme preference for ssRNA versus stem loops.

Finally, using a population genetics-phylogenetics approach I analyzed the distribution of selective coefficients along the whole gene regions and I detected positively selected sites in the human, chimpanzee and gorilla lineages. Several interesting findings emerged. For instance, the long-standing balancing selection event that was previously described at the OAS1 locus in chimpanzees [128] was detected. In fact, I identified shared selected sites in the three species, with some of them polymorphic in chimpanzees, but fixed in humans and gorillas. This observation makes perfect sense if, as shown by Ferguson et al. [128], two or more haplotypes originated before the split of great apes and were driven to fixation or maintained in the population as a result of selective forces. I also detected fixed positively selected sites shared between gorillas and humans in *MB21D1*, possibly suggesting a similar scenario as in *OAS1*.

In summary, this work represents a comprehensive analysis of the selective events acting on the MB21D1-TMEM173 and OAS-RNASEL systems in

primates and bats. Because evolutionary studies can provide information on the location and nature of adaptive changes, thus highlighting the presence of functional variation, these data provide information about sites and domains that determine species-specific infection susceptibility. Also, it has previously been suggested that, by generating diversity, selection may induce species-specific differences in the response to pharmacological compounds, suggesting caution when extrapolating results obtained in model organisms [55, 115]. I provide direct proof of this hypothesis by showing that a major determinant of the species-specific differences in the induction of the interferon gene expression in response to DMXAA was a target of positive selection. This mimetic drug showed promising antitumor effects in mice, but failed in human clinical trials. These results suggest that binding affinity for natural ligands drove the evolution of the STING binding crevice and eventually resulted in species-specific response to a synthetic compound. This observation is relevant for the design of DMXAA derivatives for the development of human antitumor and antiviral applications, as well as for the use of STING-stimulating cyclic nucleotides as vaccine adjuvants [129].

Personal contribution to the work: I performed evolutionary analyses and I analyzed data. I also produced figures and tables for the manuscript.

OASes and STING: adaptive evolution in concert

Alessandra Mozzi[1*], Chiara Pontremoli[1*], Diego Forni[1], Mario Clerici[2,3], Uberto Pozzoli[1], Nereo Bresolin[1,4], Rachele Cagliani[1], Manuela Sironi[1]

[1] Bioinformatics, Scientific Institute IRCCS E.MEDEA, 23842 Bosisio Parini, Italy.
[2] Department of Physiopathology and Transplantation, University of Milan, 20090 Milan, Italy.
[3] Don C. Gnocchi Foundation ONLUS, IRCCS, 20148 Milan, Italy.
[4] Dino Ferrari Centre, Department of Physiopathology and Transplantation, University of Milan, Fondazione Ca' Granda IRCCS Ospedale Maggiore Policlinico, 20122 Milan, Italy.

* these authors equally contributed to this work

Corresponding author: Manuela Sironi, PhD, Bioinformatics - Scientific Institute IRCCS E.MEDEA, 23842 Bosisio Parini, Italy. Tel: +39-031877915; Fax:+39-031877499; e-mail: manuela.sironi@bp.lnf.it

Abstract

OAS (2'-5'-oligoadenylate synthases) proteins and cyclic GMP-AMP synthase (cGAS, gene symbol: *MB21D1*) patrol the cytoplasm for the presence of foreign nucleic acids. Upon binding to dsRNA or dsDNA, OAS proteins and cGAS produce nucleotide second messengers to activate RNase L and STING (stimulator of interferon genes, gene symbol: *TMEM173*), respectively; this leads to the initiation of antiviral responses. We analyzed the evolutionary history of the *MB21D1-TMEM173* and *OAS-RNASEL* axes in primates and bats and found evidence of widespread positive selection in both orders. In *TMEM173*, residue 230, a major determinant of response to natural ligands and to mimetic drugs (e.g. DMXAA), was positively selected in Primates and Chiroptera. In both orders selection also targeted an α-helix/loop element in RNase L that modulates the enzyme preference for ssRNA vs. stem loops. Analysis of positively selected sites in *OAS1*, *OAS2*, and *MB21D1* revealed parallel evolution, with the corresponding residues being selected in different genes. As this cannot result from gene conversion, these data suggest that selective pressure acting on *OAS* and *MB21D1* genes is related to nucleic acid recognition and to the specific mechanism of enzyme activation, which requires a conformational change. Finally, a population genetics-phylogenetics analysis in humans, chimpanzees, and gorillas detected several positively selected sites in most genes. Data herein shed light into species-specific differences in infection susceptibility and in response to synthetic compounds, with relevance for the design of synthetic compounds as vaccine adjuvants.

Key words : OAS, cGAS, STING, RNase L, positive selection

Introduction

The innate immune system recognizes invading infectious agents through an array of so-called pattern-recognition receptors (PRRs). These molecules detect pathogen-associated molecular patterns (PAMPs) and initiate a downstream signaling cascade that ultimately triggers antiviral/antimicrobial programs. PRRs belong to diverse molecular families including Toll-like receptors (TLRs), Nod-like receptors (NLRs), RIG-I-like receptors (RLRs), and AIM2-like receptors (ALRs).

Recently, a cytosolic cyclic GMP-AMP synthase (cGAS, official gene symbol: *MB21D1*) was found to act as an antiviral DNA sensor (Sun et al. 2013). Upon binding to DNA, cGAS catalyzes the synthesis of cyclic GMP-AMP (cGAMP), which functions as a second messenger and binds the stimulator of interferon genes (STING, official gene symbol: *TMEM173*). STING, which is located in the endoplasmic reticulum (ER), can also sense cyclic dinucleotides of prokaryotic origin and is targeted by different viruses, including hepatitis C virus (HCV) and Dengue virus.

cGAS can detect a wide range of viruses and

shares structural and functional features with OAS1 (2'-5'-oligoadenylate synthase 1). Although they are not phylogenetically related, OAS1 and cGAS display similar structural fold and activation mechanisms, and both enzymes produce atypical nucleotide second messengers. In fact, OAS1 and its paralogs, OAS2 and OAS3, have long been known to bind viral dsRNA and to catalyze the synthesis of 2'-5' oligoadenylates, which specifically activate the latent form of RNase L (Hovanessian et al. 1977; Kerr and Brown 1978). Inhibition of viral propagation is eventually achieved by RNase L through RNA degradation and induction of apoptosis.

Based on their similarities, cGAS and OAS proteins may be considered as a novel family of PRRs (Civril et al. 2013; Kranzusch et al. 2013), although they impinge on different effector molecules.

Because of their direct role in PAMP recognition, PRRs and their downstream effectors are constantly involved in genetic conflicts with pathogens and, as a consequence, are commonly targeted by positive selection (Wlasiuk and Nachman 2010; Areal et al. 2011; Cagliani et al. 2014a; Cagliani et al. 2014b; Tenthorey et al. 2014). RNase L, for example, evolved adaptively in Primates, with most positively selected sites located in protein domains that directly contact the viral genetic material (Jin et al. 2012). This is in line with the host-pathogen arms race scenario, whereby protein regions directly involved in the recognition and binding of pathogen-derived components should evolve under the strongest selective pressure. This implies that species-specific differences in the function of PRRs or of their downstream effectors may be common. In the case of TLRs, for instance, the same receptor in distinct species may recognize different ligands or the same ligand with different affinity (Werling et al. 2009). Therefore, studying the pattern of inter-species evolution may provide valuable information on the differential susceptibility to infection within and among species. Primates, for example, show marked differences in the susceptibility and severity of several viral infections including those caused by HIV/SIV, HCV, HBV, and Varicella-zoster virus (Varki 2000; Willer et al. 2012). Moreover, several emerging and re-emerging viral diseases affecting humans originate through the zoonotic transmission from a reservoir animal host (Jones et al. 2008). Recent examples of pathogen spillover events include the Ebola and Middle East respiratory syndrome (MERS-CoV) viruses: both originated in bats and subsequently spread to humans either directly or through an intermediate host (Wang et al. 2011; Cotten et al. 2013). Indeed, bats (Chiroptera) have long been known to harbor and disseminate a wide range of viruses that are highly pathogenic for humans. In addition to Ebola virus and MERS-CoV, notable examples include henipavirueses (e.g. Nipah and Hendra viruses), which cause a high fatality rate in humans and other mammals, hepaciviruses, influenza A viruses, as well as a range of paramyxoviruses (Calisher et al. 2006; Drexler et al. 2012; Tong et al. 2012; Quan et al. 2013). With the exception of lyssavirus (e.g. rabies virus), bats are symptomless carriers of these human viral pathogens (Field et al. 1999). On the one hand, the observation whereby several Chiroptera families harbor a range of viral species suggests that bats have been co-evolving with viruses for a long time and have adapted to high viral exposure. On the other hand, the wide variety of viral families hosted by bats indicates that adaptation most likely involved genes with a role in immune response, rather than molecules acting as incidental viral receptors (as different viruses use distinct strategies to invade the host). Thus, innate immunity genes that are devoted to antiviral response represent excellent candidates as adaptive selection targets in Chiroptera. A recent comparison of two bat genomes (*Pteropus alecto* and *Myotis davidii*) reported adaptive evolution at such genes, including *TLR7* and *TBK1*, this latter encoding an interactor of STING (Zhang et al. 2013). Nonetheless, fast evolutionary rates at immune response loci are a common feature of mammalian genomes and surely do not represent a bat-specific trait (Barreiro and Quintana-Murci 2010; Zhang et al. 2013).

Also, an unexpected finding emerged from the analysis of the two bat genomes, as both species were found to have lost the entire cluster of *ALR* genes (Zhang et al. 2013). Overall, as noted elsewhere (Wynne and Wang 2013), the adaptive strategies underlying bat ability to asymptomatically maintain viruses remain elusive. Possibly, detailed analyses of

specific antiviral systems may help address this issue. Starting from this premise, we analyzed the evolutionary history of the *OAS-RNASEL* and *MB21D1-TMEM173* axes in primates and bats.

Materials and Methods

Gorilla sample and sequencing

The genomic DNA of one *Gorilla gorilla* was obtained from the European Collection of Cell Cultures (ECACC). *MB21D1* exons 3 and 5 were PCR-amplified from genomic DNA and directly sequenced using primers
5'-GCCTGAACATATAACATTAAC-3' (exon 3)
and
5'-AGGGTGACTCTAGTTCTTAGA -3'(exon 5)
as forward and
5'-TTATTTCCCCTGTATTTCCAG -3'(exon 3)
and 5'-GCTATGAGATGCCTAAAATCC-3' (exon 5) as reverse. 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 (Life Technologies). Sequences were assembled using AutoAssembler version 1.4.0 (Applied Biosystems), and manually inspected. The obtained sequences have been submitted to the NCBI database.

Evolutionary analyses in primates and bats

Primate and bat sequences were retrieved from the NCBI database (*http://www.ncbi.nlm.nih.gov*, last accessed October 31, 2014). The tree shrew and the horse sequences were also included in primate and bat alignments, respectively. A list of species is reported in supplementary table S1. DNA alignments were performed using the RevTrans 2.0 utility (http://www.cbs.dtu.dk/services/RevTrans/, last accesed October 31, 2014) (Wernersson and Pedersen 2003), which uses the protein sequence alignment as a scaffold to construct the corresponding DNA multiple alignment. This latter was checked and edited by TrimAl to remove alignment uncertainties (http://phylemon.bioinfo.cipf.es/utilities.html, last accessed October 31, 2014) (Capella-Gutierrez et al. 2009). Gene trees were generated by maximum-likelihood using the program phyML (Guindon et al. 2009).

Positive selection was detected using PAML (Phylogenetic Analysis by Maximum Likelihood) analyses (Yang 2007). The site models implemented in PAML were developed to detect positive selection affecting only a few aminoacid residues in a protein: positive selection is characterized by a non-synonymous substitution/synonymous substitution rate (dN/dS, also referred to as ω) ratio >1. To detect selection, site models that allow (M2a, M8) or disallow (M1a, M7) a class of sites to evolve with ω >1 were fitted to the data using the F3x4 model (codon frequencies estimated from the nucleotide frequencies in the data at each codon site) and the F61 model (frequencies of each of the 61 non-stop codons estimated from the data).

Positively selected sites were identified using two different methods: the Bayes Empirical Bayes (BEB) analysis (with a cut-off of 0.90), which calculates the posterior probability that each codon is from the site class of positive selection (under model M8) (Anisimova et al. 2002), and the Mixed Effects Model of Evolution (MEME) (with the default cutoff of 0.1) (Murrell et al. 2012), which allows the distribution of ω to vary from site to site and from branch to branch at a site. Only sites detected using both methods were considered positively selected.

To explore also possible variations in selective pressure among different lineages, we applied the free-ratio models implemented in the PAML package: the M0 model assumes all branches to have the same ω, whereas M1 allows each branch to have its own ω (Yang 1997). The models are compared through likelihood-ratio tests (degree of freedom= total number of branches -1). In order to identify specific branches with a proportion of sites evolving with ω>1, we used BS-REL (Kosakovsky Pond et al. 2011). This method implements branch-site models that simultaneously allow ω variation across branches and sites. BS-REL requires no prior knowledge about which lineages are more likely have experienced episodic diversifying selection. Branches identified using this approach were cross-validated with the branch-site likelihood ratio tests from PAML (the so-called modified model A and model MA1, “test 2”) (Zhang et al.

2005). A false discovery rate (FDR) correction was applied to account for multiple hypothesis testing (i.e. we corrected for the number of tested lineages), as suggested (Anisimova and Yang 2007). MEME and BEB analysis from MA (with a cut-off of 0.90) were used to identify sites that evolve under positive selection on specific lineages (Zhang et al. 2005).

GARD (Kosakovsky Pond et al. 2006), SLAC (Kosakovsky Pond and Frost 2005), MEME (Murrell et al. 2012) and BS-REL analyses were performed through the DataMonkey server (*http://www.datamonkey.org*, last accessed October 31, 2014) (Delport et al. 2010) or run locally (through HyPhy).

Population genetics-phylogenetics analysis

Data from the Pilot 1 phase of the 1000 Genomes Project (1000G) were retrieved from the dedicated website (*http://www.1000genomes.org/*, last accessed October 31, 2014) (1000 Genomes Project Consortium et al. 2010). SNP genotype information for 25 unrelated chimpanzees and 27 unrelated gorillas were retrieved from (Prado-Martinez et al. 2013). Coding sequence information was obtained for each gene and the ancestral sequence was reconstructed by parsimony from the human, chimpanzee, orangutan and macaque sequences. Analyses were performed with gammaMap (Wilson et al. 2011).

For gammaMap analysis, we assumed θ (neutral mutation rate per site), k (transitions/transversions ratio), and T (branch length) to vary among genes following log-normal distributions. For each gene we set the neutral frequencies of non-STOP codons (1/61) and the probability that adjacent codons share the same selection coefficient (p=0.02). For selection coefficients we considered a uniform Dirichlet distribution with the same prior weight for each selection class. For each gene we run 100,000 iterations with thinning interval of 10 iterations.

To be conservative, we declared a codon to be targeted by positive selection when the cumulative posterior probability of $\gamma \geq 1$ was > 0.75, as suggested (Quach et al. 2013).

Three-dimensional structure analysis

Protein 3D structures for human OAS1 (PDB code: 4IG8) (Donovan et al. 2013), cGAS (PDB codes: 4O67 and 4KM5) (Zhang et al. 2014; Kranzusch et al. 2013), RNase L (PDB code: 4OAV) (Han et al. 2014) and STING (PDB codes: 4LOH, 4QXP, and 4KSY) (Gao et al. 2013b; Gao et al. 2014; Zhang et al. 2013) were derived from the Protein Data Bank (PDB) (*http://www.pdb.org*, last accessed October 31, 2014); the human OAS2 model was obtained from the Protein Model Portal (code: P29728). Structure superimposition and sites mapping were performed using PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.2 Schrödinger, LLC).

Table 1

Likelihood Ratio Test Statistics for Models of Variable Selective Pressure among Sites (codon frequency model:F3x4)

Gene/Selection Model	*N* Species	−2ΔLn *L*	*P* Value	% of Sites (average *dN/dS*)
OAS1				
M1a versus M2a				
Primates	17	56.91	4.40×10^{-13}	20.0 (2.9)
Chiroptera	7	32.49	8.77×10^{-8}	7.0 (6.4)
M7 versus M8				
Primates	17	62.14	3.22×10^{-14}	23.6 (2.7)
Chiroptera	7	33.10	6.50×10^{-8}	7.0 (6.3)
OAS2				
M1a versus M2a				
Primates	16	63.19	1.89×10^{-14}	7.9 (3.3)
M7 versus M8				
Primates	16	89.44	3.78×10^{-20}	12.6 (2.7)
MB21D1				
M1a versus M2a				
Primates	16	37.05	8.99×10^{-9}	8.9 (3.2)
Chiroptera	6	12.49	0.002	35.1 (2.0)
M7 versus M8				
Primates	16	45.89	1.08×10^{-10}	11.4 (2.9)
Chiroptera	6	13.10	0.001	35.0 (2.0)
RNASEL				
M1a versus M2a				
Primates	21	41.32	1.06×10^{-9}	6.0 (3.2)
Chiroptera	7	44.19	2.53×10^{-10}	9.8 (4.2)
M7 versus M8				
Primates	21	56.06	6.70×10^{-13}	9.5 (2.6)
Chiroptera	7	44.13	2.61×10^{-10}	10.3 (4.1)
TMEM173				
M1a versus M2a				
Primates	17	6.62	0.04	3.3 (2.5)
Chiroptera	7	32.91	7.14×10^{-8}	16.9 (4.1)
M7 versus M8				
Primates	17	7.44	0.02	11.3 (1.8)
Chiroptera	7	33.32	5.81×10^{-8}	16.8 (4.1)

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

Results

Adaptive evolution in Primates

We analyzed the evolutionary history of *OAS* genes (including the enzymatically inactive

OASL), *MB21D1*, and *TMEM173* in Primates (supplementary table S1). Although Jin et al. (Jin et al. 2012) previously described adaptive evolution of *RNASEL* in Primates, we included the gene to allow comparison with bats and mapping of selected sites on the 3 dimensional structure, which has recently been solved (see below). Coding sequences for available primate species were retrieved from public databases; the tree shrew sequence was also included as an outgroup (supplementary table S2). Direct sequencing of *Gorilla gorilla* DNA was used to fill-in gaps in the coding sequence of *MB21D1*.

DNA alignments were generated using RevTrans (Wernersson and Pedersen 2003) and screened for the presence of recombination using GARD (genetic algorithm recombination detection) (Kosakovsky Pond et al. 2006). No breakpoint was detected for any gene.

The average nonsynonymous/synonymous substitution rate ratio (dN/dS, also referred to as ω) was calculated using the single-likelihood ancestor counting (SLAC) method (Kosakovsky Pond and Frost 2005). In analogy to most mammalian genes (Lindblad-Toh et al. 2011), dN/dS was always lower than 1

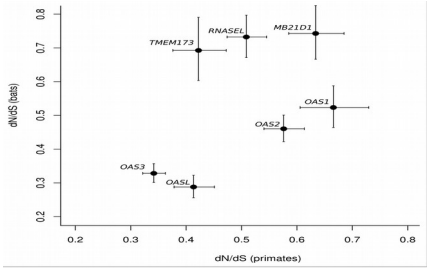


Figure 1. Plot of dN/dS values (with 95% confidence intervals) calculated for *OAS* family genes, *MB21D1*, *RNASEL*, and *TMEM173* in primates and bats.

Figure 1), indicating purifying selection as the major force shaping diversity at these genes in Primates. This finding does not exclude that localized positive selection acts on specific sites or domains. To test this possibility we applied likelihood ratio tests (LRT) implemented in the *codeml* program (Yang 1997; Yang 2007).

Under different codon frequency models, two nected in favor of the positive selection models (M2a and M8) for *OAS1*, *OAS2*, *MB21D1*, *TMEM173*, and *RNASEL* (Table 1). No evidence of positive selection was detected for *OAS3* and *OASL*. We next applied the Bayes Empirical Bayes (BEB) analysis (Anisimova et al. 2002; Yang et al. 2005) and the Mixed Effects Model of Evolution (MEME) (Murrell et al. 2012) to identify specific sites targeted by positive

(selection in these genes; only sites detected using both methods were considered (Figure 2, Table 1).

Finally, we extended our analysis to explore possible variations in selective pressure across primate lineages. To this aim, we tested whether models that allow dN/dS to vary along branches had significant better fit than models that assume one same dN/dS across the entire phylogeny (Yang and Nielsen 1998). This hypothesis was verified for *OAS1*, *MB21D1*, and *RNASEL* (supplementary table S3). We thus used the branch site-random effects likelihood (BS-REL) method (Kosakovsky Pond et al. 2011) to analyze selection along specific lineages. BS-REL identified two branches for *OAS1*, three for *MB21D1* and two for *RNASEL* (Figure 2, supplementary table S3). These were cross-validated using *codeml* (branch-site LRT models) (Zhang et al. 2005), with application of false discovery rate (FDR) correction, as suggested (Anisimova and Yang 2007). The analysis did not confirm the O*AS1* branches detected by BS-REL. Conversely for the *MD21D1* gene all the three branches were validated and positively selected sites were identified for the Hominidae and Homininae lineages (Figure 3A, supplementary table S3). Finally, for *RNASEL* only the Tibetan macaque branch was confirmed but no positively selected sites were found (supplementary fig. S1).

Positive selection in primate lineages

To gain insight into the more recent selective events in Primates, we applied a population genetics-phylogenetics approach to study the evolution of *OAS* genes, *MB21D1*, *TMEM173*, and *RNASEL* in the human, chimpanzee, and gorilla lineages. Specifically, we applied gammaMap (Wilson et al. 2011) that jointly uses intra-species variation and inter-specific

diversity to estimate the distribution of selection coefficients (γ) along coding regions. For humans we exploited data from the 1000 Genomes Pilot Project (1000G) for Europeans (CEU), Yoruba (YRI), and Chinese plus Japanese (CHBJPT) (1000 Genomes Project Consortium et al. 2010). For chimpanzees and gorillas, we used SNP information from 25 and 27 individuals, respectively (Prado-Martinez et al. 2013).We also used gammaMap to identify specific codons evolving under positive selection (defined as those having a cumulative probability >0.75 of $\gamma \geq 1$) in each lineage.

Results indicated that in the three species, most genes evolved under different degrees of purifying selection, with the exclusion of *OAS1*, which showed a preponderance of sites with γ values in the 5 to 10 range (i.e. moderately beneficial) (Figure 3B). The distribution of γ values for *OAS1* was similar in humans, chimpanzees, and gorillas. In general, γ distributions at the seven genes were comparable in the three species with the exclusion of *OAS2* in humans and *TMEM173* in chimpanzees, which showed stronger constraint compared to the other two species (Figure 3B). We also detected sites targeted by positive selection at most genes. In *OAS1* some positively selected sites were shared by two or even three (codon 54) species (Table 2). nterestingly, most of these *OAS1* sites were previously shown to define two major haplotype clades that segregate in chimpanzees and are maintained by long-standing balancing selection (Ferguson et al. 2012) (Table 2). The most recent common ancestor of the two haplotype I clades was estimated to predate the human/chimpanzee/gorilla split (Ferguson et al. 2012). Because the ancient balanced haplotypes carry several coding variants and because some of these were also detected as positively selected sites in the BEB analyses (Table 2), we re-ran the PAML site models after masking these sites in the human, chimpanzee, and gorilla sequences.

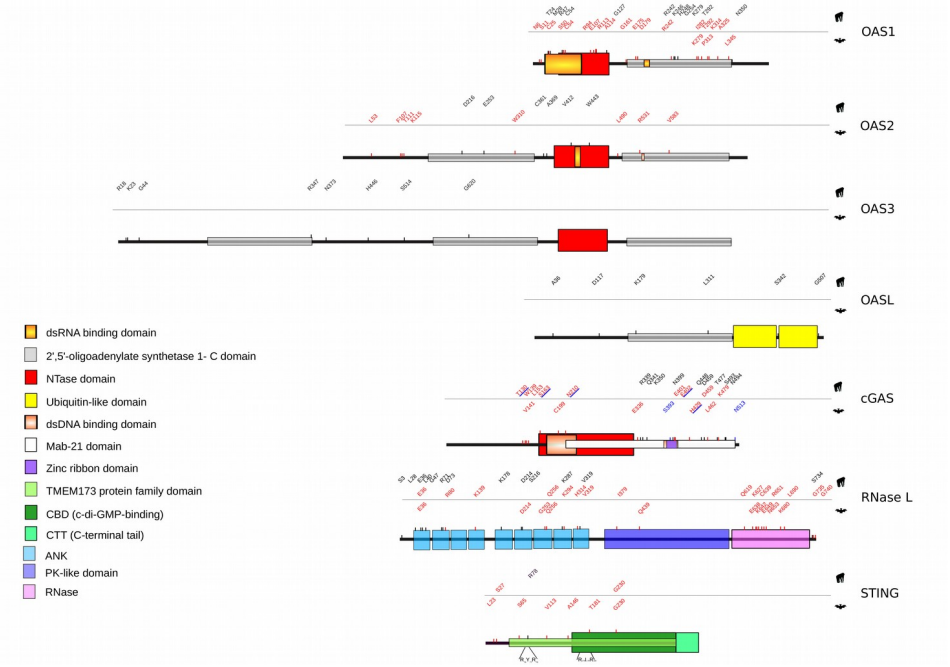


Figure 2. Schematic representation of the domain structure of OASs family members, cGAS, RNase L and STING. Domains are color-coded as reported in the legend (left). The position of positively selected sites is shown and colour-coded as follows: red, positively selected sites in the primate or bat phylogenies; blue, lineage-specific positively selected sites; black, positively selected sites in the human, chimpanzee or gorilla lineages.

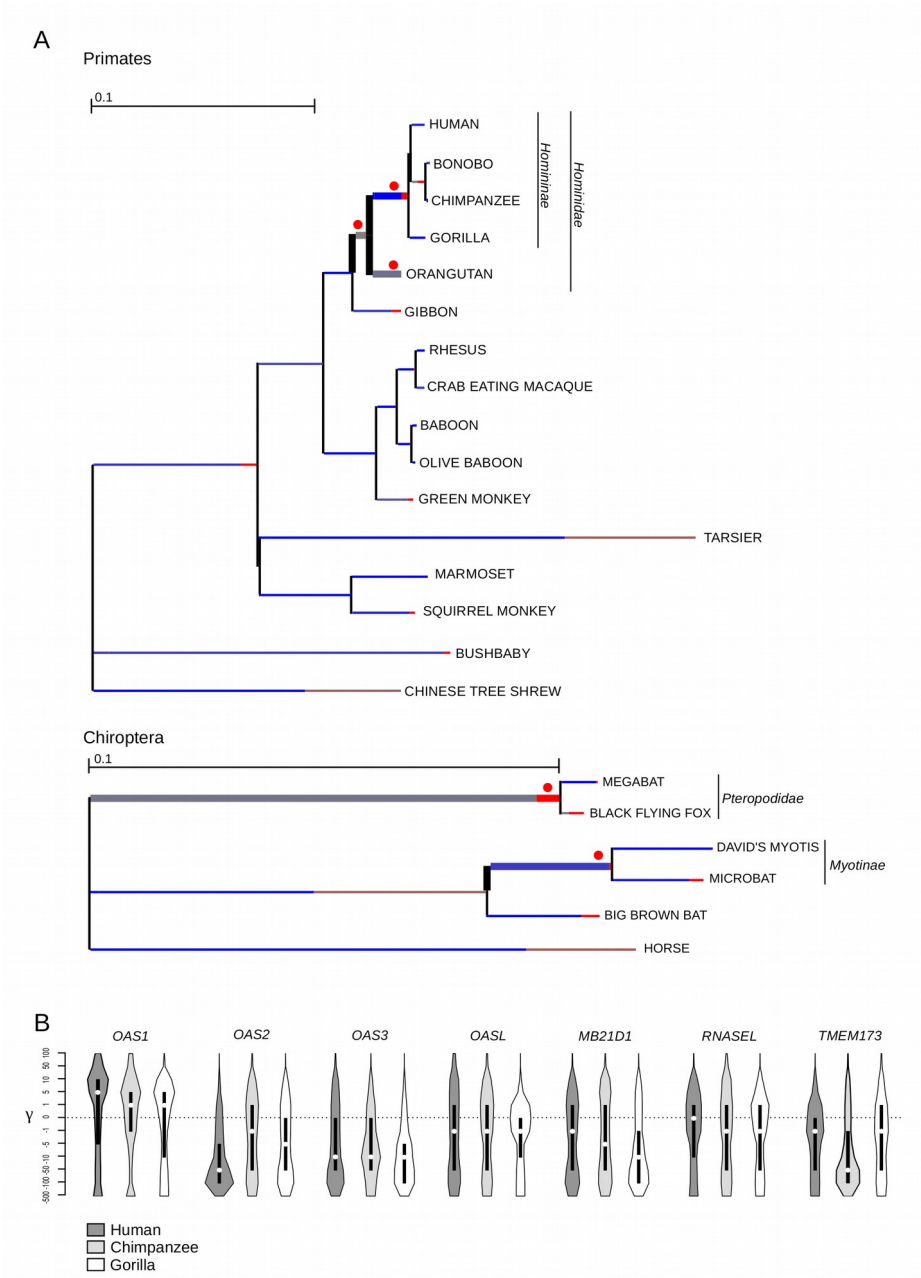


Figure 3.A. Branch-site analysis of positive selection for *MB21D1* gene (cGAS) in Primates and Chiroptera. Branch lengths are scaled to the expected number of substitutions per nucleotide, and branch colors indicate the strength of selection (ω). Red, positive selection (ω > 5); blue, purifying selection (ω = 0); gray, neutral evolution (ω = 1). The proportion of each color represents the fraction of the sequence undergoing the corresponding positive class of selection. Thick branches indicate statistical support for evolution under episodic diversifying selection as determined by BS-REL. Red dots denote branches that were also detected to be under selection using the PAML branch-site models. **B.** Violin plot of selection coefficients (median, white dot; interquartile range, black bar). Selection coefficients (γ) are classified as strongly beneficial (100, 50), moderately beneficial (10, 5), weakly beneficial (1), neutral (0), weakly deleterious (−1), moderately deleterious (−5, −10), strongly deleterious (−50, −100), and inviable (−500).

Fully significant results were obtained in all LRTs (data not shown), indicating that the positive selection signal at *OAS1* is not merely accounted for by the long-standing balancing selection event in hominids.Two positively selected sites shared between humans and gorillas were also detected at *MB21D1* (Figure 2, Table 2).

Positive selection in Chiroptera

As a comparison to Primates and given the role of these mammals as virus reservoirs, we analyzed the evolution of genes in the *OAS-RNASEL* and *MB21D1-TMEM173* axes in bats. Specifically, we obtained coding sequences for at least six bat species from public databases and we included the horse sequence as an outgroup (supplementary table S2). As for Primates, the average dN/dS substitution rate ratio, calculated using SLAC (Kosakovsky Pond and Frost 2005), was in all cases lower than 1. Comparison with Primates revealed a good correspondence in dN/dS across the seven genes, except for*TMEM173* and *RNASEL*, which showed comparatively higher values in bats than in primates (Figure 1). Application of the codeml site models indicated the action of positive selection for *OAS1*, *MB21D1*, *TMEM173*, and *RNASEL* (Table 1). BEB and MEME analyses identified positively selected sites in the four genes (Figure 2).

Variations in selective pressure among bat lineages was detected for *OAS1* and *MB21D1* (supplementary table S3); BS-REL identified the horse branch for *OAS1*, the Myotinae and Pteropodidae branches for both genes (supplementary fig. S1A, Figure 3A, supplementary table S3). With the exception of Myotinae branch for *OAS1*, c*odeml* analysis with FDR correction confirmed all branches, but BEB and MEME analyses detected lineage-specific positively selected sites in *MB21D1* only (Figure 2, supplementary table S3).

Parallel evolution at *OAS1, OAS2,* and *cGAS*

We identified several selected sites in *OAS1*; as previously noted (Ferguson et al. 2012), some residues (e.g T24, M28, R47, C54) defining the two haplotypes maintained by long-standing balancing selection are located along the so-called "spine" helix (helix α3)(Figure 4). This helix is central for human OAS1 function: it acts as a platform for nucleic acid binding and undergoes a dsRNA-induced structural switch (Donovan et al. 2013; Hornung et al. 2014). Additional sites (S11 and S50) positively selected in the whole primate phylogeny are located on this helix (Figure 4).

Comparison of the OAS1 three-dimensional structures with the OAS2 model predicted that OAS1 residue C25 corresponds to A369 in

Positive selection targets functional sites in *TMEM173* and *RNASEL*

In *TMEM173* we found six positively selected sites in bats; these sites mainly localize in functional regions of the protein. L23, S65 and V113 are in the transmembrane regions of the receptor, which are important for protein dimerization (Sun et al. 2009). T181 immediately flanks the second ER retention signal in the protein sequence (Figure 2). Interestingly, we also found R78, that is part of the first RXR retention minimal motif (Sun et al. 2009), as positively selected in the chimpanzee lineage. Other primates display amino acids different from arginine at this position, suggesting variable localization of STING in distinct species (supplementary fig. S2).

Position 230 was identified as positively selected both in the primate and in the bat phylogenies. This site is also polymorphic in human populations (G230A, rs78233829). Residue 230 is located in a loop forming the lid region that clamps onto the cyclic di-nucleotide binding pocket of the receptor; mutations of this residue affect the conformation of the protein C-terminal domain and also the binding to cyclic dinucleotides, as well as to pharmacological mimetic drugs (Gao et al. 2013b; Yi et al. 2013; Gao et al. 2014) (Figure 6).

Finally, position 146 (positively selected in bats) immediately flanks a residue (V147, human sequence) that was recently shown to determine the constitutive activation of STING, irrespective of cGAMP stimulation, when mutated to leucine in humans (Liu et al. 2014).

The evolutionary history of *RNASEL* in 2014) (Figure 6).

Finally, position 146 (positively selected in bats) immediately flanks a residue (V147,

human sequence) that was recently shown to determine the constitutive activation of STING, irrespective of cGAMP stimulation, when mutated to leucine in humans (Liu et al. 2014).

The evolutionary history of *RNASEL* in Primates had previously been analyzed (Jin et al. 2012). Herein we confirmed most sites reported by Jin and co-workers (Jin et al. 2012) and detected few more sites, possibly as a result of increased species number. Most positively selected sites detected by gammaMap or BEB/MEME localize to the ankyrin domain, whereas residues 379 and 439 (positively selected in primates and bats, respectively) lie in the ATP binding pocket of the kinase-like domain (Figure 7A). Although this domain lacks the phosphotransfer activity, nucleotide binding is maintained and required for the assembly of a functional RNase dimer (Huang et al. 2014). In bats we also found a positively selected site at position 680, within a positively charged residue patch ($_{677}$KHKKMKLK$_{684}$, human sequence) that possibly interacts with the acidic ankyrin domain (Tanaka et al. 2004) (Figure 7B). The interaction is thought to inhibit RNase L activity in absence of 2'-5' poliadenylates. In Chiroptera this position is mainly occupied by hydrophobic residues such as valine and tryptophan (Figure 7B). Finally, positively selected sites were detected in the RNase domain: most of these (E638, C639, K642, E649, R651, N653) are part of a α-helix/loop element (HLE) that creates the substrate-binding pockets (Figure 7B). Deletions in the HLE modulate the substrate preference of human RNase L (Korennykh et al. 2009; Han et al. 2014).

Discussion

Infections account for about 66% and 72% of deaths among wild chimpanzee and extant human traditional societies, respectively (Finch 2010); these figures underscore the relevance of infectious agents as powerful selective forces during the evolutionary history of Primates and, most likely, of other mammals. In fact, genetic data revealed that, among environmental factors, pathogens represented the strongest selective pressure for humans (Fumagalli et al. 2011) and several reports used inter- or intra-species diversity data to describe widespread adaptive evolution at immune response loci (Barreiro and Quintana-Murci 2010; Daugherty and Malik 2012; Quintana-Murci and Clark 2013). Specific selective events act to increase the host resistance against one or more pathogens and ample evidence indicates that the selective pressure exerted by past infections contributed to shaping the susceptibility to present-day pathogens (Kaiser et al. 2007; Emerman and Malik 2010). Also, it has previously been suggested that, by generating diversity, selection may induce species-specific differences in the response to pharmacological compounds (e.g. vaccine adjuvants), suggesting caution when extrapolating results obtained in model organisms (Werling et al. 2009; Forni et al. 2013). For these reasons, evolutionary analyses of immuneresponse genes may provide valuable information on the molecular determinants underlying species-specific infection susceptibility and may clarify the differential response to natural or synthetic molecules.

Herein we performed evolutionary analysis in primates and bats. These latter were included because of the exceptional wide range of viruses they host without developing evident pathology. Only six bat species are presently available for analysis, possibly resulting in low accuracy and power in positive selection tests (Anisimova et al. 2001). Whereas we limited the false positive rate by using two different methods to declare a site as positively selected, we may have failed to detect some true positives. Indeed, fewer selected sites were generally detected in Chiroptera than in Primates. Taking this limitation into account, we note that several selected sites were identified in the bat *TMEM173* and *RNASEL* genes, suggesting that selective pressure in these mammals was comparatively stronger for downstream effectors than for PRRs, in accordance with the higher average dN/dS values (Figure 2). In line with the tenet that natural selection targets functionally relevant residues, position 230 in *TMEM173* was found to be positively selected in both the primate and bat phylogenies. This site lies in the flexible loop that acts as a lid above the cyclic di-nucleotide binding pocket of the receptor (Figure 6). Substitutions at this site greatly affect the response to natural ligands and to mimetic drugs, such as DMXAA (Yi et al. 2013; Gao et al. 2014). In humans, positions 230 and 232 are polymorphic (G230A and

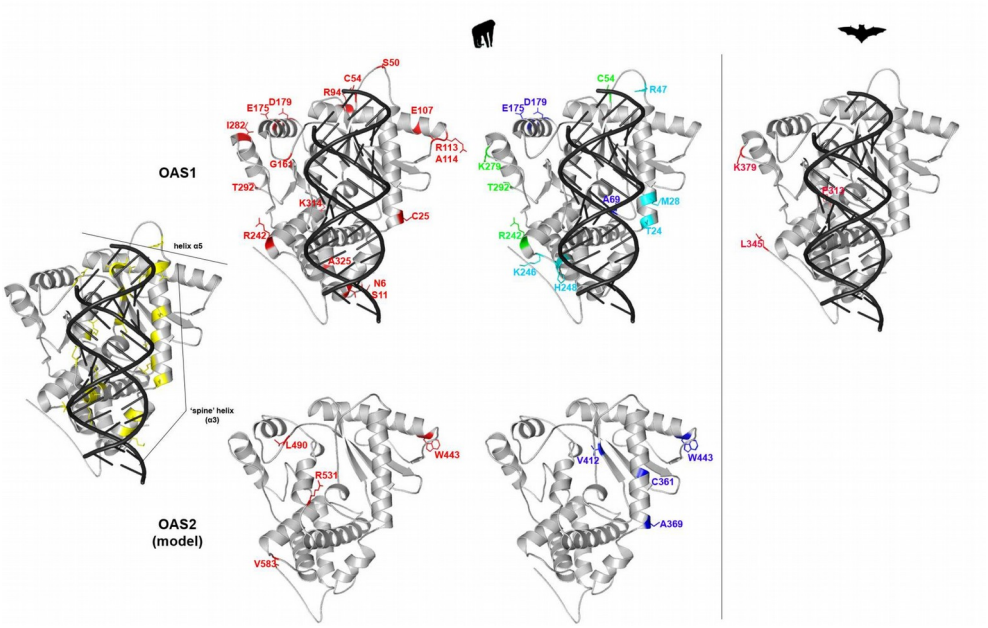


Figure 4. Positively selected sites mapped onto the human OAS1 (PDB code: 4IG8) and OAS2 (model, code: P29728) structures. Color codes are as follows: yellow, residues involved in dsRNA binding mapped onto OAS1; red, positively selected sites in the whole primate or bat phylogenies; cyan, positively selected sites in the chimpanzee lineage; blue, positively selected sites in the gorilla lineage; green, positively selected sites in more than one lineage among human, chimpanzee and gorilla.

H232R). Different alleles at these sites affect STING binding specificity for different substrates, including the canonical 3'-5' cyclic dinucleotides, known to be synthesized by bacteria, and the noncanonical [G(2'-5')pA(3'-5')p] cyclic dinucleotide, that contains a single 2'-5' phosphodiester bond and is produced by mammalian cGAS (Ablasser et al. 2013; Diner et al. 2013; Gao et al. 2013a; Zhang et al. 2013). Whereas G230 displays a substrate specificity restricted to the noncanonical dinucleotides, the G230A substitution enhances signal transduction at very low concentrations of canonical dinucleotides, because the flexibility of the loop is increased and favors the structural changes that occur upon ligand binding (Yi et al. 2013). Furthermore, even though H232R was demonstrated to be critical for the responsiveness to canonical dinucleotides, the coupled substitution G230A is required to restore a complete enzyme activation on these substrates (Diner and Vance 2014). Different aminoacid residues at position 230 were also shown to be responsible for the species-specific differences in the induction of the type I interferon pathway in response to DMXAA in human and mouse (Gao et al. 2014). Indeed, this mimetic drug showed promising antitumor effects in mice, but failed in human clinical trials because the human protein does not bind to or signal in response to DMXAA (Gao et al. 2014). Functional studies (Gao et al. 2014) nevertheless demonstrated that the substitution of Gly with Ile at position 230 results in the gain of function of human STING for DMXAA recognition.

These observations suggest that binding affinity for natural ligands drove the evolution of the STING binding crevice and eventually resulted in species-specific response to a synthetic compound. In this respect, it is worth noting that the variability of position 230 in primates should be taken into account in the design of DMXAA derivatives for the development of human antitumor and antiviral applications. These same considerations apply to the proposed use of STING-stimulating cyclic nucleotides as vaccine adjuvants (Dubensky et al. 2013).

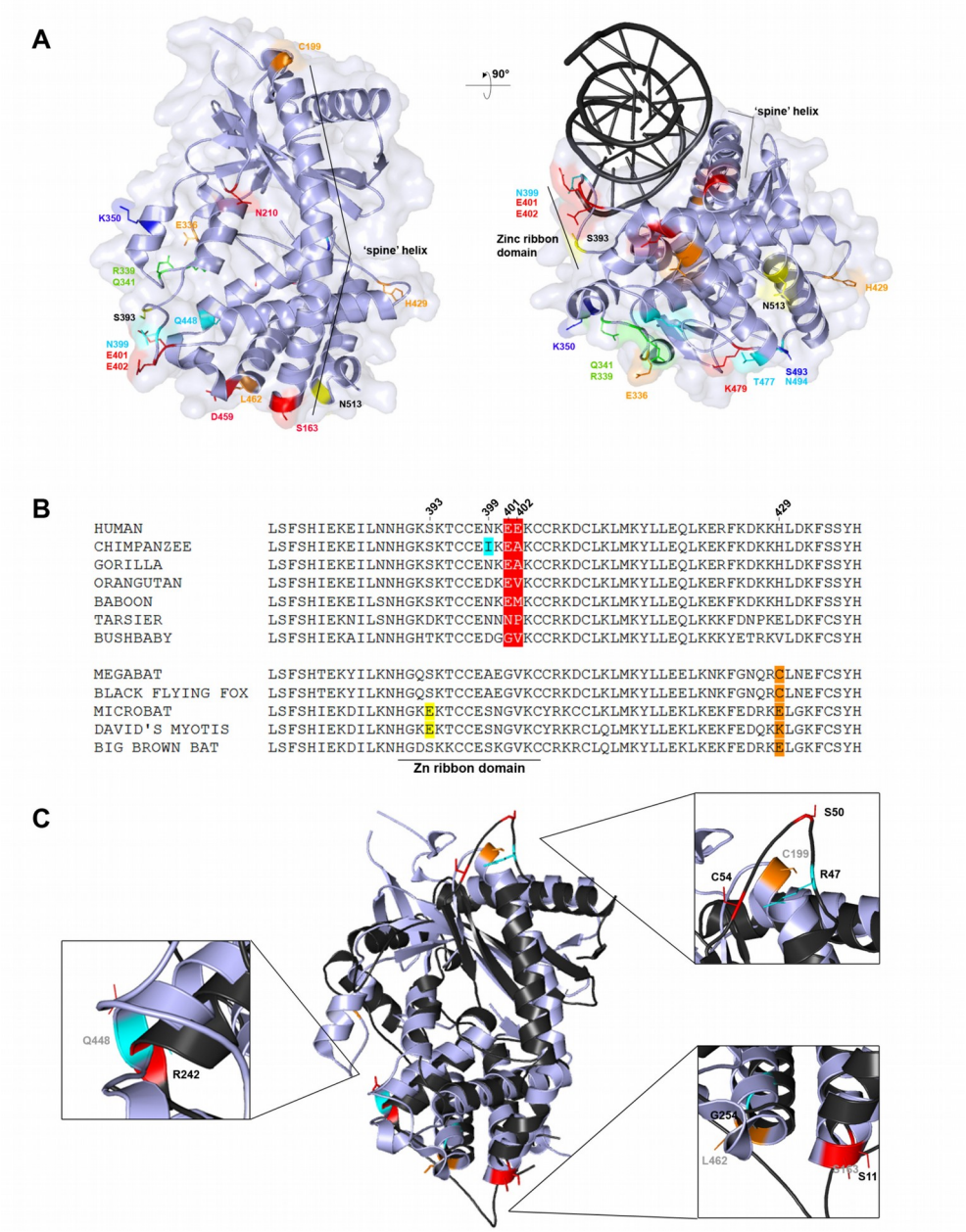


Figure 5.A. Positively selected sites mapped onto the human cGAS structure (PDB code: 4O67) . Color codes are as follows: red, positively selected sites in the primate phylogeny; orange, positively selected sites in the bat phylogeny; yellow, lineage-specific sites; cyan, positively selected sites in the chimpanzee lineage; blue, positively selected sites in the gorilla lineage; green, positively selected sites in more than one lineage among human, chimpanzee and gorilla. The cGAS-dsDNA complex was obtained by superimposing the human cGAS structure (PDB code: 4O67) with the porcine cGAS-dsDNA complex. The porcine cGAS structure is omitted. **B.** Multiple alignment of cGAS amino acids 377-437 (a portion of the sequence encompassing the zinc ribbon domain) for a few of representative primates and bats species. **C.** Superimposition of the structure of the cGAS (PDB code: 4O67, light grey) and OAS1 (PDB code: 4IG8, black). Enlargements highlight positively selected sites located in the corresponding regions of the two different enzymes. Color codes are as in A.

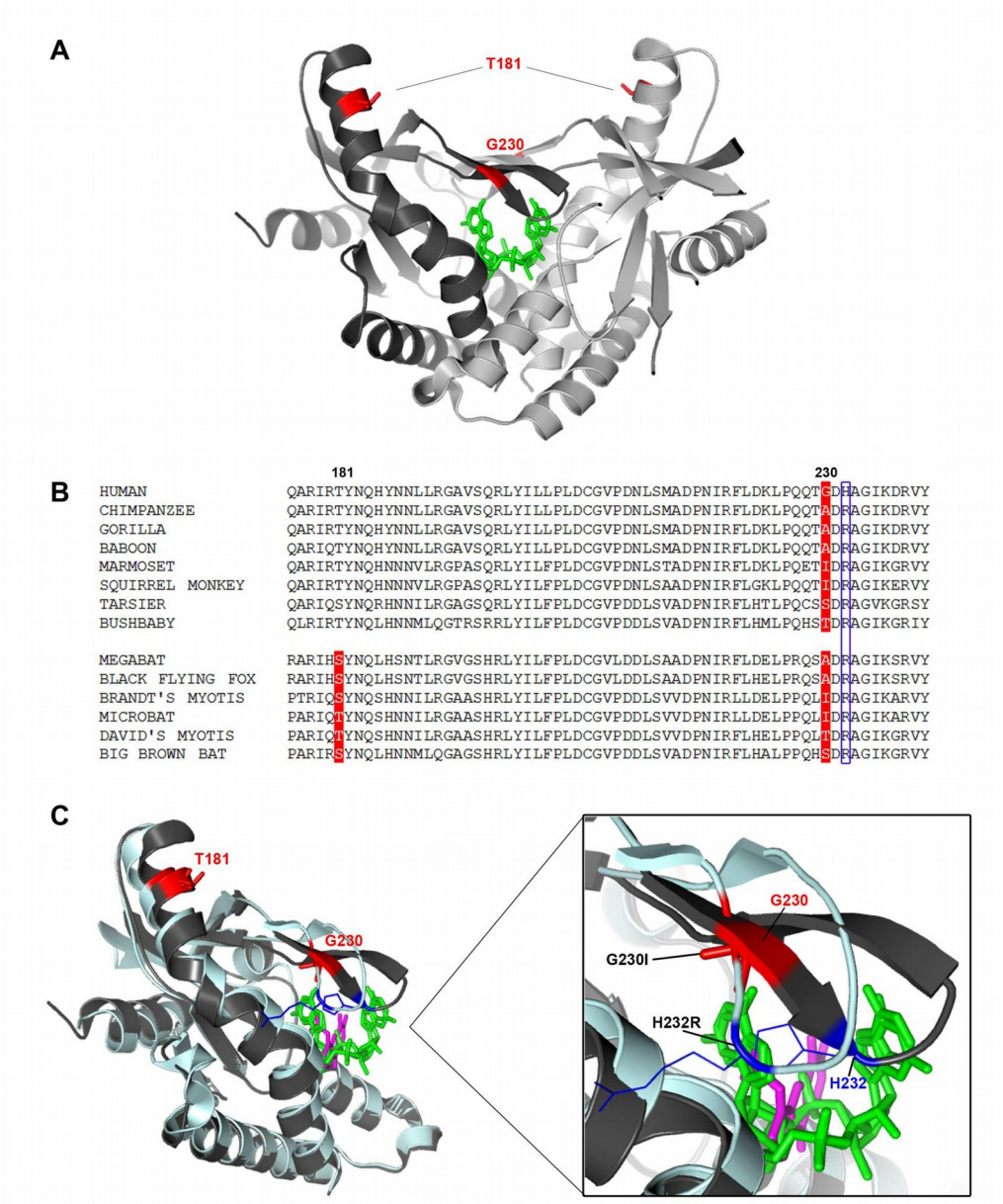


Figure 6.A. Positively selected sites mapped onto the human STING dimeric structure in complex with [G(2',5')pA(3',5')p] (green) (PDB code: 4LOH). The two monomers are colored in dark and light gray. Positively selected sites in both orders are in red. **B.** Multiple alignment of cGAS amino acids 176-240 for a few of representative primate and bat species. Positively selected in primates and/or bats are in red; position 232 is boxed in blue. **C**. Superimposition of the structure of the wt STING monomer (dark grey) in complex with [G(2',5')pA(3',5')p]) (green) (PDB code: 4LOH) and the STING double mutant (G230I, H232R) (pale cyan) in complex with DMXAA (magenta) (PDB code: 4QXP). The different conformation of the loop covering the dNTPs binding site is enlarged. [G(2',5')pA(3',5')p] and DMXAA are represented sticks.

Adding complexity to STING evolution, we also noted that positive selection in chimpanzee drove the loss of the N-terminal ER-retention signals in STING. Although the C-terminal motif ($_{178}$RIR$_{180}$) was shown to be more important for ER retention, mutagenesis of the N-terminal signal ($_{78}$RYR$_{80}$) resulted in a decreased ER localization of the protein (Sun et al. 2009). Analysis in primates revealed that additional species lack the $_{78}$RXR$_{80}$ signal. In bats, with the exclusion of the big brown bat, which has an intact second motif ($_{178}$RIR$_{180}$), no RXR motif is present in STING, nor is any of the other two motifs (KKXX and H/KDEL) usually associated with ER localization (Sun et al. 2009). The mitochondrial localization of human STING initially reported by Zhong et al (Zhong et al. 2009) has remained controversial (Ishikawa and Barber 2008; Burdette and Vance 2013). Recently, it has been suggested that the protein localizes to mitochondria-associated membranes (MAMs), where its interaction with MAVS and RIG-I occurs (Scott 2009; Horner et al. 2011). It will be interesting to assess whether species-specific differences in TMEM173 localization exist and how these affect immune response and interaction with viral-encoded inhibitors (Burdette and Vance 2013).

Similarly to TMEM173, positively selected sites were identified in functional domains of RNase L. Several positively selected sites (E638, K642, E649, and N653 in Chiroptera; C639 and R651 in Primates) localize to the short HLE element, which constitutes the binding-pocket for RNA and modulates the preference of the enzyme for ssRNA molecules or stem loops (Korennykh et al. 2009; Han et al. 2014) (Figure 7). Experiments in cell lines indicated that human RNase L cleaves HCV RNA predominately at UA and UU dinucleotides within loops of predicted stem-loop structures (Han and Barton 2002). More recently, a phylogenetically conserved RNA structure in the open reading frame of poliovirus (and other group C enteroviruses) was found to function as a competitive inhibitor of RNase L (Han et al. 2007). Specific stem loops motifs were found to be important for the inhibitory activity and to account for unusual resistance of poliovirus to RNase L-mediated cleavage (Townsend et al. 2008). Thus, positively selected sites in the HLE represent excellent candidates as modulators of RNase L cleavage rate or susceptibility to inhibitors.

The RNase L ankyrin repeats domain was strongly targeted by selection, as well. Most selected sites lie in the loop between the two antiparallel alpha-helices and in the outer helix of the ankyrin module; even though none is directly involved in the 2'-5' oligoadenylates binding, they could potentially mediate the dimerization process or the interaction with other proteins. Intriguingly, the ankyrin domain of murine RNase L is the molecular target of L*, a protein of Theiler's Virus, a neurotropic picornavirus. The interaction between RNase L and L* is strictly species-specific: the viral protein is unable to inhibit RNase L of non-murine origin (Sorgeloos et al. 2013). Although our analysis did not include rodents, these observations indicate that the ankyrin repeat domain may be a target of virus-encoded inhibitors.

Overall, these data suggest that the selective pressure acting on STING and RNase L is mainly related to the modulation of molecular recognition and, possibly, to the escape from viral inhibitors. It will be interesting to evaluate whether the positively selected sites we detected in bats contribute to the exceptional adaptation of these mammals to different viral pathogens. Analysis of the PRRs underscored major signatures of adaptive evolution for cGAS in both Chiroptera and Primates, whereas OAS1 was strongly targeted by selection in Primates, and much more weakly in bats. We found several positively selected sites to be located in the relatively short cGAS-specific zinc ribbon domain. This structure is thought to act as a molecular ruler and to endow cGAS with the ability of binding the B-form but not A-form of nucleic acids, which is instead recognized by OAS1 (Civril et al. 2013; Gao et al. 2013a; Kranzusch et al. 2013). Thus, positive selection in this domain may act to hone in the function of this protruding loop, which is responsible both for nucleic acid recognition and for protein dimerization, two essential steps for full enzyme activation (Li et al. 2013; Zhang et al. 2014). Human cGAS and OAS1 recognize nucleic acids through sequence-independent interactions to the minor groove, mainly mediated by a positively charged platform on the protein surface (Hornung et al. 2014).

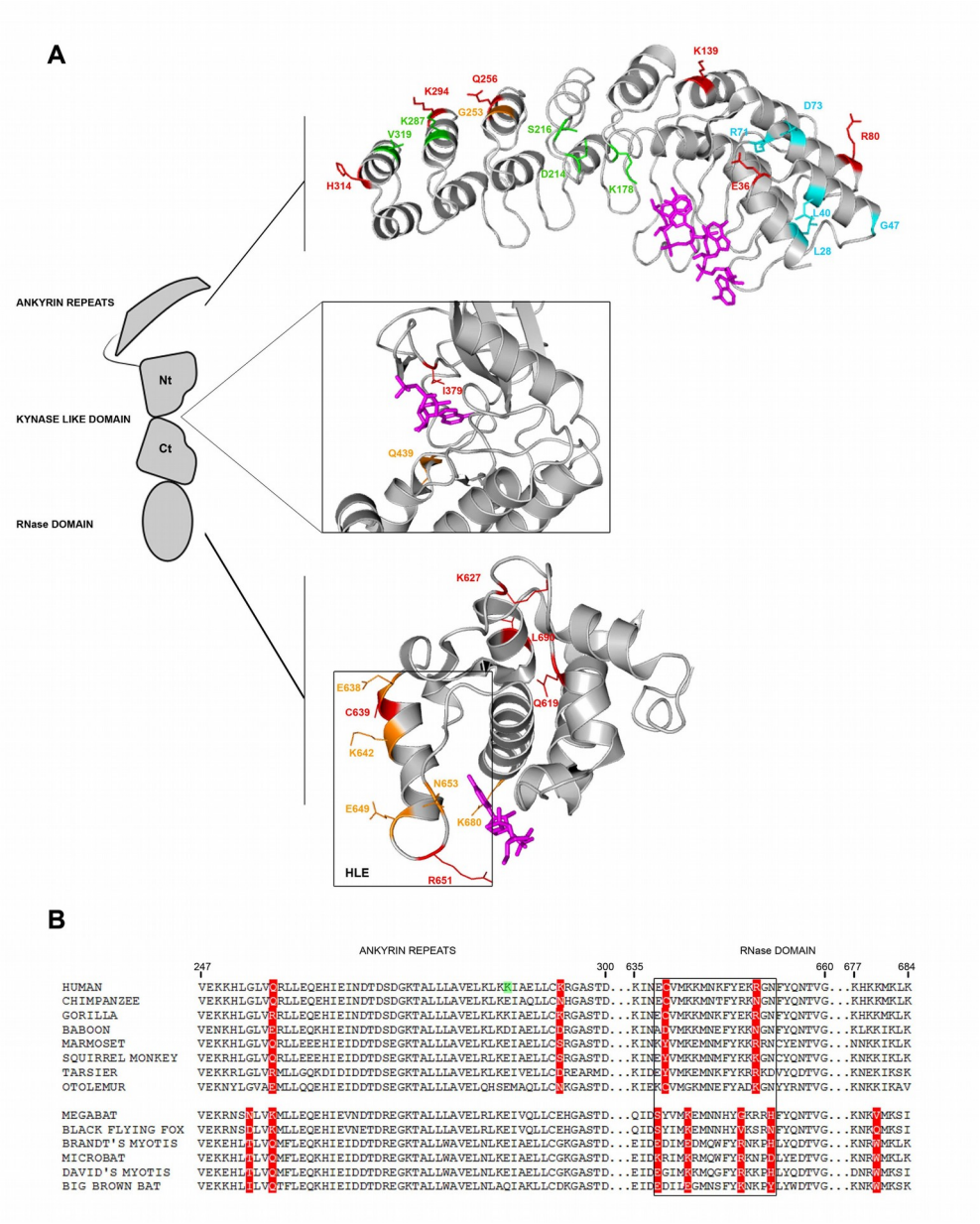


Figure 7.A. Positively selected sites mapped onto the human RNase L structure (PDB code: 4OAV). Enlargements show the ankyrin repeats in complex with 2-5pppA$_7$ (magenta), the ATP binding site of theprotein kinase-like domain in complex with AMP-PCP (magenta) and the RNase domain in complex with RNA (two sugar-phosphate groups and one pyrimidine nucleobase solved, magenta). Color codes are as follows: red, positively selected sites in the primate phylogeny; orange, positively selected sites in the bat phylogeny; cyan, positively selected sites in the chimpanzee lineage; green, positively selected sites in the human lineage. The HLE is boxed in black. **B.** Multiple alignment of RNase L amino acids 247-684 (a portion of the sequence encompassing the HLE, black boxed) for a few of representative primates and bats species. Positively selected sites in primates and bats are in red, the positive selected site in the human lineage is in green.

Table 2
Positively Selected Sites in the Human, Chimpanzee, and Gorilla Lineages

Gene	Lineage	Codon	Ancestral AA	Derived AA	Pr[a]	DAF[b]	Other Methods[c]
OAS1							
	Human	54[d]	Arg	Cys	0.868	1	MEME–BEB
		127	Asp	Gly	0.928	1	
		279[d]	Glu	Lys	0.885	1	BEB
		292[d]	Arg	Thr	0.976	1	MEME–BEB
		350	Asp	Asn	0.963	1	
	Chimpanzee	24[d]	Thr	Lys	0.988	1	
		28[d]	Met	Lys	0.990	1	BEB
		47[d]	Arg	Gln	0.982	1	
		54[d]	Arg	His	0.972	1	MEME–BEB
		242[d]	Arg	Gln	0.834	0.62	MEME–BEB
		246[d]	Lys	Glu	0.845	0.60	BEB
		248[d]	His	Asp	0.847	0.60	BEB
		254[d]	Gly	Glu	0.842	0.61	
		292[d]	Arg	Thr/Glu	0.941	0.62/0.38	MEME–BEB
		355	Trp	Stop[e]	0.775	0.62	
	Gorilla	54[d]	Arg	Cys	0.818	1	MEME–BEB
		69	Thr	Ala	0.823	1	BEB
		127	Asp	Gly	0.846	1	
		175	Glu	Lys	0.940	1	MEME–BEB
		179	Asp	Tyr	0.936	1	MEME–BEB
		242[d]	Arg	Gln	0.807	1	MEME–BEB
		279[d]	Glu	Lys	0.850	1	BEB
OAS2							
	Chimpanzee	216	Asp	Asn	0.834	1	
		253	Glu	Lys	0.828	1	BEB
	Gorilla	361	Cys	Phe	0.869	1	BEB
		369	Ala	Thr	0.867	1	
		412	Val	Ile	0.799	1	
		443	Trp	Ser	0.767	1	MEME–BEB
OAS3							
	Human	446	Arg	His	0.828	1	
		514	Gly	Ser	0.860	1	MEME–BEB
		620	Arg	Gly	0.837	1	
	Chimpanzee	347	Arg	Cys	0.794	1	MEME–BEB
		373	Asn	Ser	0.773	0.96	MEME–BEB
	Gorilla	18	Arg	Ser	0.870	1	
		23	Lys	Thr	0.873	1	
		44	Gly	Ala	0.818	1	
OASL							
	Human	179	Glu	Lys	0.857	1	
		311	His	Leu	0.857	1	
		432	Pro	Ser	0.856	1	
	Chimpanzee	36	Ala	Thr	0.799	1	
		117	Asp	Asn	0.759	1	
		507	Gly	Arg	0.839	1	
MB21D1							
	Human	339	Pro	Arg	0.909	1	MEME
		341	Lys	Gln	0.910	1	
	Chimpanzee	399	Asn	Ile	0.854	1	
		448	Gln	Glu	0.910	1	
		477	Thr	Ile	0.912	1	
		494	Asn	Asp	0.894	1	
	Gorilla	339	Pro	Arg	0.931	1	MEME
		341	Lys	Gln	0.932	1	

(continued)

Table 2 Continued

Gene	Lineage	Codon	Ancestral AA	Derived AA	Pr[a]	DAF[b]	Other Methods[c]
		350	Lys	Arg	0.900	1	BEB
		459	Asp	Gln	0.898	1	BEB
		493	Ser	Arg	0.774	1	
RNASEL							
	Human	3	Thr	Ser	0.823	1	
		178	Glu	Lys	0.816	1	BEB
		214	His	Asp	0.928	1	BEB
		216	Arg	Ser	0.928	1	MEME
		287	Glu	Lys	0.802	1	MEME
		319	Phe	Val	0.781	1	MEME–BEB
	Chimpanzee	28	Leu	Ser	0.987	1	
		36	Glu	Gly	0.997	1	MEME–BEB
		40	Leu	Gln	0.998	1	MEME
		47	Gly	Asp	0.995	1	BEB
		71	Arg	Lys	0.985	1	MEME
		73	Asp	Glu	0.982	1	MEME
		734	Ser	Cys	0.806	1	
TMEM173							
	Chimpanzee	78	Arg	Trp	0.759	1	

[a]Posterior probability of $\gamma \geq 1$ as detected by gammaMap.
[b]Derived allele frequency.
[c]Other methods that identified the same codon as positively selected.
[d]Site described as long-term balancing selection target (see text).
[e]To perform GammaMap analyses, the STOP codon was substituted with a different codon.

A long alpha helix, called 'spine' helix, opposite to the active site crevice, is the major structural component of the platform (Hornung et al. 2014). For an efficient activation in vitro, human OAS1 and cGAS require dsRNA molecules >17 bp long (Donovan et al. 2013) and dsDNA molecules >20 bp (Kranzusch et al. 2013), respectively. Intriguingly, we found positively selected sites at the double ends of the spine helix of both proteins, as shown by the superimposition of thethree-dimensional structures. As cGAS and OAS1 use double-stranded acid topology to distinguish between DNA and RNA and for specific self-activation upon binding, domains directly involved in nucleic acid recognition may have evolved adaptively to respond to specific PAMPs or to optimize enzyme activation. A similar evolutionary scenario has been recently proposed for positively selected sites in the pincer region of RIG-I, another PRR (Lemos de Matos et al. 2013; Cagliani et al. 2014a; Rawling et al. 2015). Thus, the selective pressure acting on *OAS* and *MB21D1* genes may be related to PAMP recognition and to the specific mechanism of enzyme activation, which envisages a conformational change. This hypothesis is strengthened by the observation that natural selection often targeted residues located in the same spatial position in different proteins. In this respect, we should add that the duplication events that originated the *OAS* gene family occurred before the radiation of mammalian lineages, although more recent expansions occurred in rodents (Kumar et al. 2000). Consequently, *OAS* family genes are quite divergent in sequence and the possibility that gene conversion between paralogs contributes substantially to their evolution has previously been dismissed (Ferguson et al. 2012). Likewise, gene conversion events between the *OAS1* and *MB21D1* coding regions are extremely unlikely, as the two genes display limited sequence identity, despite the extensive structural and functional similarities. Thus, the several instances of corresponding positions which were targeted by positive selection in *OAS1* and *OAS2*, as well as *OAS1* and *MB21D1*, should be regarded as independent events resulting from selection.

In terms of function and structure, it is worth noting that recent analyses (Kranzusch et al. 2014; Zhu et al. 2014) indicated that cGAS is homologous to bacterial enzymes that synthesize 3'-5' cGAMP, revealing an evolutionary connection across distinct kingdoms. Based on these structural similarities Kranzusch and coworkers (Kranzusch et al. 2014) showed that single aminoacid replacements around the cGAS binding site alter the enzyme's linkage specificity. Whereas the selected sites in STING may affect the specificity for products with different phosphodiester bonds, we did not detect positively selected sites in or near

the cGAS active site, suggesting that the major selective pressure acting on the enzyme was not related to changes in STING ligand specificity. Evolutionary analyses on additional species will be required, though, to address the potential of co-evolution for cGAS product specificity and STING binding preferences (Kranzusch et al. 2014).

The combined analysis of intra-species polymorphism and between-species divergence allows detection of positive selection targets in one species and provides information on the distribution of selective coefficients along the whole gene regions. A previous study of *TLR* gene evolution in humans and great apes revealed a stronger effect of purifying selection in chimpanzees and gorillas compared to humans (Quach et al. 2013). We analyzed these same species and did not detect a similar trend. Nonetheless, in that previous work, the major difference among species was accounted for by TLRs that recognize bacterial PAMPs, whereas the genes we analyzed herein are mainly devoted to antiviral response. In general, the distribution of selection coefficients were relatively similar among the three species, with the exception of *OAS2* and *TMEM173*, which showed a marked preponderance of selectively constrained codons in humans and chimpanzees, respectively.

An interesting observation emerging from the population genetics-phylogenetics analysis is that results were consistent with the long-standing balancing selection scenario that was previously described at the OAS1 locus in hominids (Ferguson et al. 2012). In fact, we identified shared selected sites in the three species, with some of them polymorphic in chimpanzees, but fixed in humans and gorillas. This observation makes perfect sense if, as shown by Ferguson et al. (Ferguson et al. 2012), two or more haplotypes originated before the split of great apes and were driven to fixation or maintained in the population as a result of selective forces. We also detected fixed positively selected sites shared between gorillas and humans in *MB21D1*, possibly suggesting a similar scenario as in *OAS1*.

Overall, our population genetics-phylogenetics analysis identified several sites which were targeted by positive selection in distinct great ape lineages; these represent extremely promising candidates as modulators of infection susceptibility in these species. Whereas some of these sites are located in protein regions with clear functional characterization (e.g. at the nucleic acid binding interface, at intracellular trafficking signals), the significance of other selected residues remains elusive. As suggested for the balanced polymorphisms in *OAS1* (Ferguson et al. 2012), selection may act on regions that play a role in protein folding and stability. Also, it will be interesting to investigate whether the diverse evolutionary histories for OAS1 and STING in distinct great ape species resulted from the selective pressure exerted by of one or more pathogens.

Acknowledgments

CP is supported by a fellowship of the Doctorate School of Molecular and Translational Medicine, University of Milan.

The *MB21D1* exons 3 and 5 coding sequence for *Gorilla gorilla* has beensubmitted to GenBank (provisional ID: KP085619).

References

1000 Genomes Project Consortium, et al. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061-1073.

Ablasser A., et al. 2013. cGAS produces a 2'-5'-linked cyclic dinucleotide second messenger that activates STING. Nature 498:380-384.

Anisimova M., and Z. Yang. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol. Biol. Evol. 24:1219-1228.

Anisimova M., J. P. Bielawski, and Z. Yang. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol. Biol. Evol. 18:1585-1592.

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

Areal H., J. Abrantes, and P. J. Esteves. 2011. Signatures of positive selection in Toll-like receptor (TLR) genes in mammals. BMC Evol.

Biol. 11:368-2148-11-368.

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

Burdette D. L., and R. E. Vance. 2013. STING and the innate immune response to nucleic acids in the cytosol. Nat. Immunol. 14:19-26.

Cagliani R., et al. 2014a. RIG-I-like receptors evolved adaptively in mammals, with parallel evolution at LGP2 and RIG-I. J. Mol. Biol. 426:1351-1365.

Cagliani R., et al. 2014b. Ancient and recent selective pressures shaped genetic diversity at AIM2-like nucleic acid sensors. Genome Biol. Evol. 6:830-845.

Calisher C. H., J. E. Childs, H. E. Field, K. V. Holmes, and T. Schountz. 2006. Bats: important reservoir hosts of emerging viruses. Clin. Microbiol. Rev. 19:531-545.

Capella-Gutierrez S., J. M. Silla-Martinez, and T. Gabaldon. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972-1973.

Civril F., et al. 2013. Structural mechanism of cytosolic DNA sensing by cGAS. Nature 498:332-337.

Cotten M., et al. 2013. Full-genome deep sequencing and phylogenetic analysis of novel human betacoronavirus. Emerg. Infect. Dis. 19:736-42B.

Daugherty M. D., and H. S. Malik. 2012. Rules of engagement: molecular insights from host-virus arms races. Annu. Rev. Genet. 46:677-700.

Delport W., A. F. Poon, S. D. Frost, and S. L. Kosakovsky Pond. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 26:2455-2457.

Diner E. J., and R. E. Vance. 2014. Taking the STING out of cytosolic DNA sensing. Trends Immunol. 35:1-2.

Diner E. J., et al. 2013. The innate immune DNA sensor cGAS produces a noncanonical cyclic dinucleotide that activates human STING. Cell. Rep. 3:1355-1361.

Donovan J., M. Dufner, and A. Korennykh. 2013. Structural basis for cytosolic double-stranded RNA surveillance by human oligoadenylate synthetase 1. Proc. Natl. Acad. Sci. U. S. A. 110:1652-1657.

Drexler J. F., et al. 2012. Bats host major mammalian paramyxoviruses. Nat. Commun. 3:796.

Dubensky T. W.,Jr, D. B. Kanne, and M. L. Leong. 2013. Rationale, progress and development of vaccines utilizing STING-activating cyclic dinucleotide adjuvants. Ther. Adv. Vaccines 1:131-143.

Emerman M., and H. S. Malik. 2010. Paleovirology--modern consequences of ancient viruses. PLoS Biol. 8:e1000301.

Ferguson W., S. Dvora, R. W. Fikes, A. C. Stone, and S. Boissinot. 2012. Long-term balancing selection at the antiviral gene OAS1 in Central African chimpanzees. Mol. Biol. Evol. 29:1093-1103.

Field H., B. McCall, and J. Barrett. 1999. Australian bat lyssavirus infection in a captive juvenile black flying fox. Emerg. Infect. Dis. 5:438-440.

Finch C. E. 2010. Evolution in health and medicine Sackler colloquium: Evolution of the human lifespan and diseases of aging: roles of infection, inflammation, and nutrition. Proc. Natl. Acad. Sci. U. S. A. 107 Suppl 1:1718-1724.

Forni D., et al. 2013. A 175 million year history of T cell regulatory molecules reveals widespread selection, with adaptive evolution of disease alleles. Immunity 38:1129-1141.

Fumagalli M., et al. 2011. Signatures of Environmental Genetic Adaptation Pinpoint

Pathogens as the Main Selective Pressure through Human Evolution. PLoS Genet. 7:e1002355.

Gao P., et al. 2014. Binding-Pocket and Lid-Region Substitutions Render Human STING Sensitive to the Species-Specific Drug DMXAA. Cell. Rep. 8:1668-1676.

Gao P., et al. 2013a. Cyclic [G(2',5')pA(3',5')p] is the metazoan second messenger produced by DNA-activated cyclic GMP-AMP synthase. Cell 153:1094-1107.

Gao P., et al. 2013b. Structure-function analysis of STING activation by c[G(2',5')pA(3',5')p] and targeting by antiviral DMXAA. Cell 154:748-762.

Guindon S., F. Delsuc, J. F. Dufayard, and O. Gascuel. 2009. Estimating maximum likelihood phylogenies with PhyML. Methods Mol. Biol. 537:113-137.

Han J. Q., and D. J. Barton. 2002. Activation and evasion of the antiviral 2'-5' oligoadenylate synthetase/ribonuclease L pathway by hepatitis C virus mRNA. RNA 8:512-525.

Han J. Q., et al. 2007. A phylogenetically conserved RNA structure in the poliovirus open reading frame inhibits the antiviral endoribonuclease RNase L. J. Virol. 81:5561-5572.

Han Y., et al. 2014. Structure of human RNase L reveals the basis for regulated RNA decay in the IFN response. Science 343:1244-1248.

Horner S. M., H. M. Liu, H. S. Park, J. Briley, and M. Gale Jr. 2011. Mitochondrial-associated endoplasmic reticulum membranes (MAM) form innate immune synapses and are targeted by hepatitis C virus. Proc. Natl. Acad. Sci. U. S. A. 108:14590-14595.

Hornung V., R. Hartmann, A. Ablasser, and K. P. Hopfner. 2014. OAS proteins and cGAS: unifying concepts in sensing and responding to cytosolic nucleic acids. Nat. Rev. Immunol. 14:521-528.

Hovanessian A. G., R. E. Brown, and I. M. Kerr. 1977. Synthesis of low molecular weight inhibitor of protein synthesis with enzyme from interferon-treated cells. Nature 268:537-540.

Huang H., et al. 2014. Dimeric structure of pseudokinase RNase L bound to 2-5A reveals a basis for interferon-induced antiviral activity. Mol. Cell 53:221-234.

Ishikawa H., and G. N. Barber. 2008. STING is an endoplasmic reticulum adaptor that facilitates innate immune signalling. Nature 455:674-678.

Jin W., D. D. Wu, X. Zhang, D. M. Irwin, and Y. P. Zhang. 2012. Positive selection on the gene RNASEL: correlation between patterns of evolution and function. Mol. Biol. Evol. 29:3161-3168.

Jones K. E., et al. 2008. Global trends in emerging infectious diseases. Nature 451:990-993.

Kaiser S. M., H. S. Malik, and M. Emerman. 2007. Restriction of an extinct retrovirus by the human TRIM5alpha antiviral protein. Science 316:1756-1758.

Kerr I. M., and R. E. Brown. 1978. pppA2'p5'A2'p5'A: an inhibitor of protein synthesis synthesized with an enzyme fraction from interferon-treated cells. Proc. Natl. Acad. Sci. U. S. A. 75:256-260.

Korennykh A. V., et al. 2009. The unfolded protein response signals through high-order assembly of Ire1. Nature 457:687-693.

Kosakovsky Pond S. L., and S. D. Frost. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. 22:1208-1222.

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

Kosakovsky Pond S. L., et al. 2011. A random effects branch-site model for detecting episodic diversifying selection. Mol. Biol. Evol. 28:3033-3043.

Kranzusch P. J., A. S. Lee, J. M. Berger, and J. A. Doudna. 2013. Structure of human cGAS reveals a conserved family of second-messenger enzymes in innate immunity. Cell. Rep. 3:1362-1368.

Kranzusch P. J., et al. 2014. Structure-guided reprogramming of human cGAS dinucleotide linkage specificity. Cell 158:1011-1021.

Kumar S., C. Mitnik, G. Valente, and G. Floyd-Smith. 2000. Expansion and molecular evolution of the interferon-induced 2'-5' oligoadenylate synthetase gene family. Mol. Biol. Evol. 17:738-750.

Lemos de Matos A., G. McFadden, and P. J. Esteves. 2013. Positive evolutionary selection on the RIG-I-like receptor genes in mammals. PLoS One 8:e81864.

Li X., et al. 2013. Cyclic GMP-AMP synthase is activated by double-stranded DNA-induced oligomerization. Immunity 39:1019-1031.

Lindblad-Toh K., et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478:476-482.

Liu Y., et al. 2014. Activated STING in a vascular and pulmonary syndrome. N. Engl. J. Med. 371:507-518.

Murrell B., et al. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 8:e1002764.

Prado-Martinez J., et al. 2013. Great ape genetic diversity and population history. Nature 499:471-475.

Quach H., et al. 2013. Different selective pressures shape the evolution of Toll-like receptors in human and African great ape populations. Hum. Mol. Genet. 22:4829-4840.

Quan P. L., et al. 2013. Bats are a major natural reservoir for hepaciviruses and pegiviruses. Proc. Natl. Acad. Sci. U. S. A. 110:8194-8199.

Quintana-Murci L., and A. G. Clark. 2013. Population genetic tools for dissecting innate immunity in humans. Nat. Rev. Immunol. 13:280-293.

Rawling D. C., A. S. Kohlway, D. Luo, S. C. Ding, and A. M. Pyle. 2015. The RIG-I ATPase core has evolved a functional requirement for allosteric stabilization by the Pincer domain. Nucleic Acids Res. 42:11601-11611.

Scott I. 2009. Degradation of RIG-I following cytomegalovirus infection is independent of apoptosis. Microbes Infect. 11:973-979.

Sorgeloos F., B. K. Jha, R. H. Silverman, and T. Michiels. 2013. Evasion of antiviral innate immunity by Theiler's virus L* protein through direct inhibition of RNase L. PLoS Pathog. 9:e1003474.

Sun L., J. Wu, F. Du, X. Chen, and Z. J. Chen. 2013. Cyclic GMP-AMP synthase is a cytosolic DNA sensor that activates the type I interferon pathway. Science 339:786-791.

Sun W., et al. 2009. ERIS, an endoplasmic reticulum IFN stimulator, activates innate immune signaling through dimerization. Proc. Natl. Acad. Sci. U. S. A. 106:8653-8658.

Tanaka N., et al. 2004. Structural basis for recognition of 2',5'-linked oligoadenylates by human ribonuclease L. EMBO J. 23:3929-3938.

Tenthorey J. L., E. M. Kofoed, M. D. Daugherty, H. S. Malik, and R. E. Vance. 2014. Molecular basis for specific recognition of bacterial ligands by NAIP/NLRC4 inflammasomes. Mol. Cell 54:17-29.

Tong S., et al. 2012. A distinct lineage of influenza A virus from bats. Proc. Natl. Acad. Sci. U. S. A. 109:4269-4274.

Townsend H. L., B. K. Jha, R. H. Silverman, and D. J. Barton. 2008. A putative loop E motif and an H-H kissing loop interaction are conserved and functional features in a group C enterovirus RNA that inhibits ribonuclease L. RNA Biol. 5:263-272.

Varki A. 2000. A chimpanzee genome project

is a biomedical imperative. Genome Res. 10:1065-1070.

Wang L. F., P. J. Walker, and L. L. Poon. 2011. Mass extinctions, biodiversity and mitochondrial function: are bats 'special' as reservoirs for emerging viruses? Curr. Opin. Virol. 1:649-657.

Werling D., O. C. Jann, V. Offord, E. J. Glass, and T. J. Coffey. 2009. Variation matters: TLR structure and species-specific pathogen recognition. Trends Immunol. 30:124-130.

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

Willer D. O., et al. 2012. Experimental infection of Cynomolgus Macaques (Macaca fascicularis) with human varicella-zoster virus. J. Virol. 86:3626-3634.

Wilson D. J., R. D. Hernandez, P. Andolfatto, and M. Przeworski. 2011. A population genetics-phylogenetics approach to inferring natural selection in coding sequences. PLoS Genet. 7:e1002395.

Wlasiuk G., and M. W. Nachman. 2010. Adaptation and constraint at Toll-like receptors in primates. Mol. Biol. Evol. 27:2172-2186.

Wynne J. W., and L. F. Wang. 2013. Bats and viruses: friend or foe? PLoS Pathog. 9:e1003651.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555-556.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586-1591.

Yang Z., and R. Nielsen. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. 46:409-418.

Yang Z., W. S. Wong, and R. Nielsen. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22:1107-1118.

Yi G., et al. 2013. Single nucleotide polymorphisms of human STING can affect innate immune response to cyclic dinucleotides. PLoS One 8:e77846.

Zhang G., et al. 2013. Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. Science 339:456-460.

Zhang J., R. Nielsen, and Z. Yang. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22:2472-2479.

Zhang X., et al. 2013. Cyclic GMP-AMP containing mixed phosphodiester linkages is an endogenous high-affinity ligand for STING. Mol. Cell 51:226-235.

Zhang X., et al. 2014. The cytosolic DNA sensor cGAS forms an oligomeric complex with DNA and undergoes switch-like conformational changes in the activation loop. Cell. Rep. 6:421-430.

Zhong B., et al. 2009. The ubiquitin ligase RNF5 regulates antiviral responses by mediating degradation of the adaptor protein MITA. Immunity 30:397-407.

Zhu D., et al. 2014. Structural biochemistry of a Vibrio cholerae dinucleotide cyclase reveals cyclase activity regulation by folates. Mol. Cell 55:931-937.

3.2.1.2 Diverse selective regimes shape genetic diversity at *ADAR* genes and at their coding targets

Recent technological advances are providing an increasingly detailed picture of the extent, regulation and location of A-to-I editing events in the genome of humans and other species. More than one million A-to-I editing sites have currently been described in humans and RNA editing has been implicated in a number of processes and diseases [130-132]. Despite this wealth of knowledge, evolutionary studies have been lagging behind.

In this work I aimed at providing a comprehensive overview of the evolutionary history of the three mammalian genes that code for ADAR family members and which operate A-to-I RNA editing. I also provide evolutionary analysis at their coding targets.

To this aim I used a variety of datasets and applied state-of-the-art approaches to analyze the evolution of *ADAR*, *ADARB1* and *ADARB2* at the inter- and intra-species level. I next used available information on human editing sites in coding regions, as well as a score of conservation across mammals (I referred to RADAR database, http://rnaedit.com/ [133], and Genomic Evolutionary Rate Profiling (GERP) score, respectively), to study the evolution of recoding sites.

Results indicated that, over diverse time periods, diversity at the three *ADAR* genes was mainly shaped by purifying selection, as described for most mammalian genes. Nonetheless, likelihood ratio tests [13, 75] revealed that ADAR evolved adaptively in primates, with strongest selection in the unique N-terminal domain of the interferon-inducible isoform (ADARp150). Positively selected residues in the human lineage were also detected in the ADAR deaminase domain (adjacent to residues that cause Aicardi-Goutieres Syndrome when mutated) and in the second RNA binding domain of both ADARB1 and ADARB2. In particular, the corresponding

residue was independently selected in the two genes.

Population genetics analysis in humans revealed that the three *ADAR* family genes were targeted by selection during the recent history of human populations, as well. By integrating different positive selection tests, ENCODE annotation data, as well as large-scale information on eQTLs, I show that distinct variants in the three genes increased in frequency as a result of local selective pressures. Most selected variants are located within regulatory regions and some are in linkage disequilibrium with eQTLs in monocytes.

Data herein indicate that *ADAR* family genes were target by positive selection in primates, in the human lineage, and during the recent history of human populations. These findings and a wealth of previous data suggest a central role for these enzymes in physiological and pathological processes [131, 132, 134], a role at least partially mediated by the specific editing of coding sites. I thus set out to determine the evolutionary history of coding editing sites.

Recently, a study in macaques indicated that purifying selection is the major force acting at editing sites and at their flanking positions [135]. In partial agreement, a previous analysis had suggested that editing sites are less conserved across primates than their flanks, but that the overall region carrying the editing site is more conserved than control sequences [136]. Herein, I used a score of conservation across mammals and I focused on editing sites located in coding regions by separately analyzing editing events that originate synonymous and nonsynonymous substitution, to account for underlying variation in sequence conservation in coding sequences. In contrast to previous reports, my findings indicate that the editing site is less conserved than its flanks, which, in turn, are more variable than control positions randomly drawn from the edited coding

regions. This effect is observed at shared (among mammals) and non-shared nonsynonymous sites, as well as at positions that entail synonymous substitutions when edited. The possible reasons for these discrepancies are manifold. In previously works, the overwhelming majority of analyzed editing sites was accounted for by non-coding sites. As for the macaque editome data, the authors analyzed fewer than 30 editing sites in coding regions shared among macaques, chimpanzees and humans; sequence conservation was measured in terms of human-macaque percentage identity or dS [135]. Also, the authors did not consider the effect of editing on the protein sequence (whether or not recoding occurs) and did not use a comparison with control sequences. As suggested, the lower conservation I observed at the editing site might reflect fixation of the edited form in some mammals or correction of G-to-A mutations through editing [136]. Nonetheless, these possibilities do not explain why flanking positions are less conserved than control sites. I suggest that most editing events (at both synonymous and nonsynonymous sites) are slightly deleterious and are therefore counter-selected within regions that are poorly tolerant to change. Nevertheless, I show that a minority of recoding events occurs at highly conserved positions and possibly represents the functional fraction. These events are enriched in pathways related to HIV-1 infection and to epidermis/hair development. Interestingly, these processes are related to ADAR/ADARB1 known functions. In fact, ADAR and ADARB1 have been involved in HIV-1 infection [137-140]. Also, mutations in *ADAR* are responsible for dyschromatosis symmetrica hereditaria (DSH), a pigmentary skin disease. More recently, hair anomalies have also been described in patients with DHS [141]. Consistently, a conditional mouse model that lacks ADAR expression in the epidermis shows fur loss and skin pathology [142].

In summary, these data indicate that both *ADAR* family genes and their targets evolved under variable selective regimes, including purifying and positive selection. I suggest that pressures related to immune response were major drivers of evolution for *ADAR* genes and, possibly, some of their targets. These analyses do not support the previous suggestion whereby A-to-I RNA editing contributed to the development of higher brain functions in humans [143, 144].

Personal contribution to the work: I designed the study with my co-workers, I performed research and I analyzed data. I also produced tables and figures for the manuscript.

Diverse selective regimes shape genetic diversity at *ADAR* genes and at their coding targets

Diego Forni[1]*, Alessandra Mozzi[1]*, Chiara Pontremoli[1]*, Jacopo Vertemara[1], Uberto Pozzoli[1], Mara Biasin[2], Nereo Bresolin[1,3], Mario Clerici[4,5], Rachele Cagliani[1], Manuela Sironi[1]

[1] Bioinformatics, Scientific Institute IRCCS E. MEDEA, 23842 Bosisio Parini, Italy.
[2] Department of Biomedical and Clinical Sciences, University of Milan, 20157 Milan, Italy.
[3] Dino Ferrari Centre, Department of Physiopathology and Transplantation, University of Milan, Fondazione Ca' Granda IRCCS Ospedale Maggiore Policlinico, 20122 Milan, Italy.
[4] Chair of Immunology, Department of Physiopathology and Transplantation, University of Milan, 20090 Milan, Italy.
[5] Don C. Gnocchi Foundation ONLUS, IRCCS, 20148 Milan, Italy.

* these authors equally contributed to this work

Corresponding author: Manuela Sironi, PhD, Bioinformatics - Scientific Institute IRCCS E. MEDEA, 23842 Bosisio Parini, Italy. Tel: +39-031877915; Fax:+39-031877499; e-mail: manuela.sironi@bp.lnf.it

Abstract

A-to-I RNA editing operated by ADAR enzymes is extremely common in mammals. Several editing events in coding regions have pivotal physiological roles and affect protein sequence (recoding events) or function. We analyzed the evolutionary history of the three *ADAR* family genes and of their coding targets. Evolutionary analysis indicated that *ADAR* evolved adaptively in primates, with the strongest selection in the unique N-terminal domain of the interferon-inducible isoform. Positively selected residues in the human lineage were also detected in the ADAR deaminase domain and in the RNA binding domains of ADARB1 and ADARB2. During the recent history of human populations distinct variants in the three genes increased in frequency as a result of local selective pressures. Most selected variants are located within regulatory regions and some are in linkage disequilibrium with eQTLs in monocytes. Finally, analysis of conservation scores of coding editing sites indicated that editing events are counter-selected within regions that are poorly tolerant to change. Nevertheless, a minority of recoding events occurs at highly conserved positions and possibly represents the functional fraction. These events are enriched in pathways related to HIV-1 infection and to epidermis/hair development. Thus, both *ADAR* genes and their targets evolved under variable selective regimes, including purifying and positive selection. Pressures related to immune response likely represented major drivers of evolution for *ADAR* genes. As for their coding targets, we suggest that most editing events are slightly deleterious, although a minority may be beneficial and contribute to antiviral response and skin homeostasis.

Keywords
ADAR, A-to-I editing, positive selection, ADAR editing sites, evolutionary analysis

Introduction

RNA editing, defined as the post-transcriptional modification of RNA molecules not including splicing, capping, and polyadenylation, is a widespread phenomenon in several living organisms. In metazoans, the most common RNA editing event is the adenosine to inosine (A-to-I) conversion operated by ADAR (adenosine deaminases acting on RNA) enzymes mainly on dsRNA substrates.[1] Recent

estimates suggest that ~1.6 million editing sites exist in the human genome.[2]

Mammalian genomes encode three *ADAR* genes: the catalytically active *ADAR* and *ADARB1*, plus *ADARB2*, thought to be inactive and to serve a regulatory role.[3] *ADARB2* expression is brain-specific, whereas the other

two *ADAR* genes are transcribed in many tissues.[1] A unique feature of *ADAR* is the presence of an interferon (IFN)-inducible promoter that drives expression of a full-lenght ADARp150 protein; the constitutive, non IFN-responsive promoter determines the synthesis of a shorter N-terminally truncated ADARp110 product. In line with its IFN-inducible properties, ADAR was shown play a role in antiviral responses.[4] Mutations in *ADAR* are responsible for two different genetic diseases: Aicardi-Goutières Syndrome (AGS) and dyschromatosis symmetrica hereditaria (DHS).[5] This latter is a pigmentary skin disease, whereas AGS is an autoinflammatory conditions mainly affecting the brain and the skin. AGS patients carrying *ADAR* mutations display up-regulation of IFN stimulated genes, suggesting that the gene acts as a suppressor of IFN responses.[6]

A-to-I RNA editing is thought to be involved in several physiological and pathological processes. Inosine is recognized as guanosine by the translation and splicing machineries. Therefore, editing events can result in a wide range of effects and may affect protein sequence (recoding events) and function when they occur in coding regions. Compared to other mammals, primates exhibit much higher levels of transcriptome editing, mainly as a result of their genome being rich in *Alu* sequences, which represent preferential editing sites by virtue of their propensity to form double-stranded RNA structures.[7] Thus, the overwhelming majority of editing events occurs in non-coding repetitive sequences. Nonetheless, several A-to-I conversions in coding regions have a pivotal physiological role. Among these, the best studied examples include brain-specific ion channels and neurotransmitter receptors. For instance, editing at a single site (known as the Q/R site) in the ionotropic glutamate receptor subunit GLUR2 alters Ca^{2+} permeability and is essential for normal brain development. Indeed, *Adarb1*$^{-/-}$ mice suffer from epileptic seizures and die several weeks after birth. The phenotype is rescued by introduction of a transgene that allows expression of the *Glur2* edited form.[8] Whereas the GLUR2 Q/R editing event is shared by humans and rodents, conserved mammalian editing sites are a small minority.[9]

In humans and other primates most editing events occur in the brain and, in analogy to *GLUR2*, involve neuronal genes. Paz-Yaacov and coworkers also indicated that human-specific editable *Alu* insertions are enriched in genes related to neuronal functions or implicated in neurological diseases[10]. This observation, together with the higher editing levels in the brain of humans compared to chimpanzees and macaques, led some authors to suggest that A-to-I RNA editing contributed to the development of higher brain functions.[8,10]

From an evolutionary perspective, RNA editing is an extremely interesting phenomenon. In analogy to alternative splicing, editing can provide transcriptome variability and, as noted, it might allow variation at sites that would otherwise be inaccessible to mutation, which instead imposes high fitness costs.[11] Nevertheless, recent data have indicated that editing of coding sequences is generally nonadaptive in humans, although the presence of few beneficial recoding events was postulated.[12]

On this basis, we set out to perform an evolutionary analysis of *ADAR* family genes and of their coding targets.

Results

ADAR evolved adaptively in primates

To analyze the evolutionary history of *ADAR* genes (*ADAR*, *ADARB1*, and *ADARB2*) in primates, we obtained coding sequences for available species in public databases; the tree shrew sequence was also included (Table S1). The three DNA alignments were generated using RevTrans and screened for the presence of recombination breakpoints using GARD (genetic algorithm recombination detection).[13,14] No breakpoint was detected for any gene.

We next calculated the average nonsynonymous substitution/synonymous substitution rate ratio (dN/dS, also referred to as ω) using the single-likelihood ancestor counting (SLAC) method.[15] In all cases dN/dS

was lower than 1 (Table 1), indicating purifying selection as the major driving force in shaping diversity at *ADAR* genes in primates. This is not unusual, as most mammalian genes display variable levels of purifying selection at their coding regions.[16] A major effect of negative selection is not incompatible with positive selection acting on specific sites or domains. To assess whether positive selection acted on *ADAR* family members, we applied the likelihood ratio tests (LRT) implemented in the *codeml* program.[17,18] LRTs compare models of gene evolution that allow (NSsite models M2a and M8, positive selection models) or disallow (NSsite models M1a and M7, null models) a class of codons to evolve with dN/dS > 1. In the case of *ADAR*, but not of *ADARB1* and *ADARB2*, both neutral models were rejected in favor of the positive selection models; these results were confirmed using different models of codon frequency (Table 2). In order to identify specific ADAR sites targeted by positive selection, we applied the Bayes Empirical Bayes (BEB) analysis and the Mixed Effects Model of Evolution (MEME).[19-21] To be conservative, only sites detected using both methods were considered. We identified a total of eight positively selected sites. Interestingly, seven of these are located in the additional amino-terminal portion of the interferon-induced isoform of ADAR (p150) (Fig. 1). One of the positively selected sites (H129) maps to a nuclear exporting sequence (NES) (Fig. 1). The positively selected L25 residue immediately flanks a missense mutation identified in patients with DSH (Fig. 1).[5]We next extended our analysis to explore possible variations in selective pressure among primate lineages at ADAR. To this aim, we tested whether models that allow dN/dS to vary along branches had significant better fit than models that assume one same dN/dS across the entire phylogeny.[22] Because this hypothesis was verified, we used the branch site-random effects likelihood (BS-REL) method to analyze selection along specific lineages.[23] BS-REL identified three branches: green monkey, bonobo and tree shrew (Fig.S1). These were cross-validated using *codeml* (branch-site LRT models), with application of false discovery rate (FDR) correction, as suggested.[24,25] This analysis confirmed the bonobo branch only (Table S3), but detected no lineage-specific positively selected sites. We note that this is not unusual, as the simultaneous inference of both the site and the branch subject to diversifying selection is difficult;[21,24] thus, BEB analysis is accurate but has low power in this context.[24]

Positive selection of *ADAR* family genes in the human lineage

For humans, we exploited data from the 1000 Genomes Pilot Project (1000G) for Europeans (CEU), Yoruba (YRI), and Chinese plus Japanese (CHBJPT).[26] For chimpanzees, we used phased SNP information of 10 *Pan troglodytes verus.*[27] Ancestral sequences were reconstructed by parsimony from the human, chimpanzee, orangutan and macaque sequences. In line with the results obtained above, we observed a general preponderance of codons evolving under negative selection ($\gamma < 0$) for all genes. In particular, in both species *ADAR* was found to be less constrained than *ADARB1* and *ADARB2* (Fig. 2). ene regions. We thus used gammaMap, a

We next used gammaMap to identify specific codons evolving under positive selection in humans and chimpanzees. The combined analysis of intra-species polymorphism and between-species divergence may allow increased power to detect sites targeted by positive selection in one species. Moreover, this approach provides information on the distribution of selective effects along recently developed program that models intragenic variation in selection coefficients (γ), to study the evolution of *ADAR* family members in the human and chimpanzee lineages.

Table 1. Genomic position and average dN/dS for *ADAR* family genes.

Gene symbol[a]	Alias	Genomic location	Protein length (aa)	Average dN/dS (confidence intervals)
ADAR	*ADAR1*	Chr1:154,554,534–154,580,724	1226	0.289 (0.265, 0.314)
ADARB1	*ADAR2*	Chr21:46,494,493–46,646,478	741	0.081 (0.070, 0.093)
ADARB2	*ADAR3*	Chr10:1,223,253–1,779,670	739	0.104 (0.093, 0.116)

[a]Official gene symbol as approved by the HUGO Gene Nomenclature Committee (HGNC)

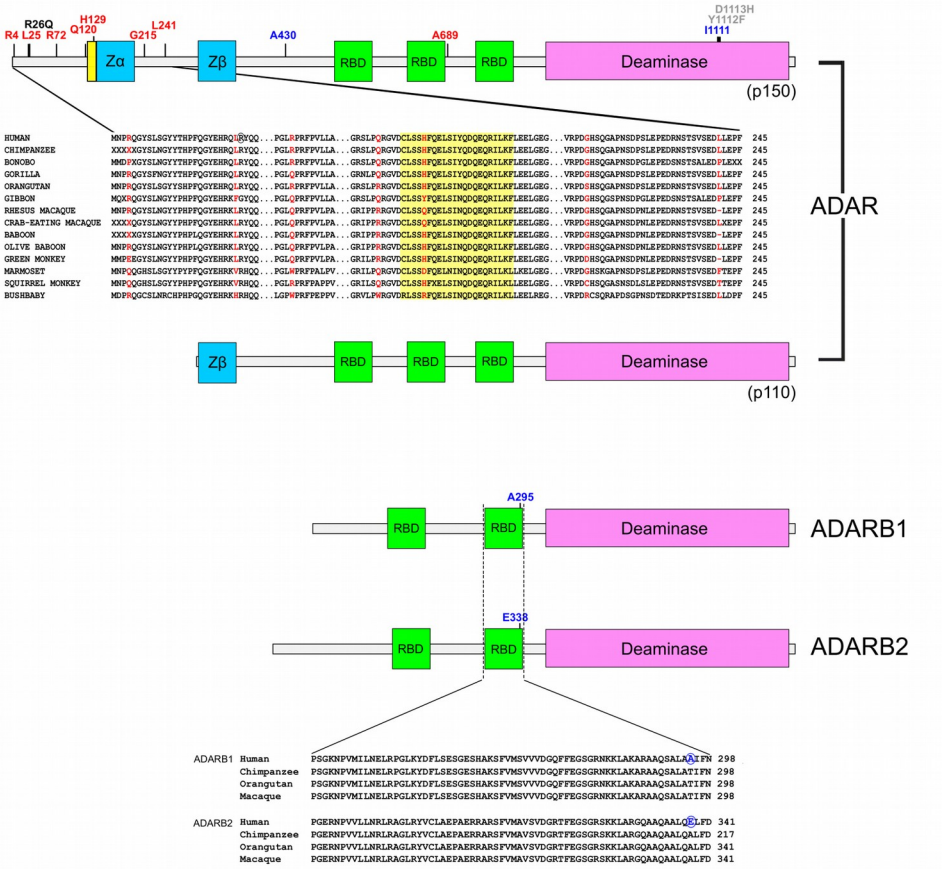


Figure 1. Adaptive evolution at *ADAR* genes in primates. Schematic representation of the domain structure of ADAR family members. Domains are color-coded: nuclear export signal (NES), yellow; Z-DNA binding domains, cyan; RNA binding motifs (RBD), green; deaminase domain, pink. The position of positively selected sites is shown together with sequence alignments for a few representative primates. Positively selected sites in primates and in the human lineage are shown in red and blue, respectively. Some missense mutations associated with AGS and DSH are shown in gray and black, respectively [5].

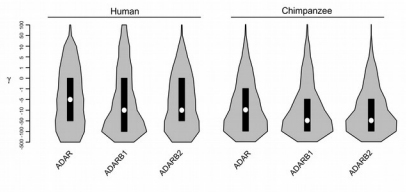


Figure 2. Analysis of selective pressure in the human and chimpanzee lineages. Violin plot of selection coefficients (median, white dot; interquartile range, black bar). Selection coefficients (γ) are classified as strongly beneficial (100, 50), moderately beneficial (10, 5), weakly beneficial (1), neutral (0), weakly deleterious (−1), moderately deleterious (−5, −10), strongly deleterious (−50, −100), and inviable (−500).

To be conservative, we declared a codon to be targeted by positive selection when the cumulative posterior probability of $\gamma \geq 1$ was > 0.75, as suggested.28 No positively selected codon was identified for *P. troglodytes*. For humans, two positively selected sites were identified in ADAR; one of these (I1111) is located within the deaminase domain and is immediately adjacent to two mutations (Y1112F and D1113H) responsible for AGS (Fig. 1 and Table S2).[5] One selected site was identified in both ADARB1 and ADARB2 (Table S2); the selected sites are in the second dsRNA binding (RBD) domain: alignment of ADARB1 and ADARB2 indicated that the

Table 2. Likelihood ratio test (LRT) statistics for models of variable selective pressure among sites and branches.

LRT model	Codon Frequency model	Degrees of freedom	-2ΔLnL[d]	*p* value	% of sites (average dN/dS)	Positively selected sites (BEB and MEME)
ADAR						
M1a vs M2a[a]	F3x4	2	22.24	1.48×10^{-5}	1.0% (6.0)	
	F61	2	19.07	7.21×10^{-5}	0.8% (6.2)	
M7 vs M8[b]	F3x4	2	27.37	1.14×10^{-6}	1.7% (4.8)	4R, 25L, 72R, 120Q, 129H, 215G, 241L, 689A
	F61	2	22.14	1.55×10^{-5}	1.3% (5.0)	
M0 vs M1[c]	F3x4	30	181.39	1.41×10^{-23}	-	-
	F61	30	174.01	3.18×10^{-22}	-	-

[a]M1a is a nearly neutral model that assumes one ω class between 0 and 1, and one class with ω=1; M2a (positive selection model) is the same as M1a plus an extra class of ω >1.
[b]M7 is a null model that assumes that 0<ω<1 is β distributed among sites; M8 (positive selection model) is the same as M7 but also includes an extra category of sites with ω>1.
[c]M0 and M1 are free-ratio models which assume all branches to have the same ω (M0) or allow each branch to have its own ω (M1).
[d]2ΔlnL: twice the difference of the natural logs of the maximum likelihood of the models being compared.

corresponding position is targeted by selection (Fig. 1).

Non-coding regulatory variants represent targets of positive selection in human populations

We next investigated whether natural selection acted on *ADAR* family members during the recent evolutionary history of human populations. The 1000G data for YRI, CEU, and CHBJPT were used to this purpose.

Integration of different tests can improve the power to detect selective sweeps and, importantly, allows identification of the causal adaptive variant(s).[29-31] We applied the DIND (Derived Intra-allelic Nucleotide Diversity) test, which is powerful in most derived allele frequency (DAF) ranges and less sensitive than iHS (Integrated Haplotype Score**)** to low genotype quality or low coverage (i.e. it is well suited for the 1000G data).[32,33] DIND results were combined with pairwise F_{ST} analyses, whereas DH was calculated in sliding-windows to account for local events and, for this reason, used as an a posteriori validation.[34] Statistical significance (in terms of percentile rank) for all tests was obtained by deriving empirical distributions from a control set of ~1000 genes (see Materials and Methods). We declared a variant to be selected if it displayed both a DIND and an F_{ST} percentile rank >0.95. DH was used to validate high-frequency sweeps, in line with the power profile of this test.[34]

We detected distinct selection signals in *ADAR*. In CEU and CHBJPT, the same variant (rs884618) had unusually high F_{ST} (YRI/CEU and YRI/CHBJPT comparisons) and represented a DIND outlier (Table 3). The

Table 3. Candidate targets of positive selection in human populations.

Gene	SNP ID	Derived allele	DAF[a] YRI	DAF[a] CEU	DAF[a] CHBJPT	DIND rank (population[b])	F_{ST} rank (comparison)
ADAR	rs2172708	A	0.25	0	0	0.97 (YRI)	>0.99 (YRI/CEU) >0.99 (YRI/CHBJPT)
	rs6677920	C	0.24	0	0	0.96 (YRI)	>0.99 (YRI/CEU) >0.99 (YRI/CHBJPT)
	rs9427095	G	0.24	0	0	0.95 (YRI)	>0.99 (YRI/CEU) >0.99 (YRI/CHBJPT)
	rs1542796	C	0.25	0	0	0.96 (YRI)	0.99 (YRI/CEU)>0.99 (YRI/CHBJPT)
	rs11806816	T	0.23	0	0	0.96 (YRI)	>0.99 (YRI/CEU) >0.99 (YRI/CHBJPT)
	rs884618	G	0.02	0.44	0.47	0.96 (CEU)0.96 (CHBJPT)	0.97 (YRI/CEU)0.98 (YRI/CHBJPT)
ADARB1	rs4819027	C	0	0.27	0.13	0.95 (CEU)0.95 (CHBJPT)	>0.99 (YRI/CEU)>0.99 (YRI/CHBJPT)
ADARB2	rs2820600	T	0.22	0	0	0.96 (YRI)	>0.99 (YRI/CEU)>0.99 (YRI/CHBJPT)
	rs2820599	G	0.22	0	0	0.96 (YRI)	>0.99 (YRI/CEU)>0.99 (YRI/CHBJPT)
	rs2805512	G	0.22	0.02	0	0.96 (YRI)	0.95 (YRI/CEU)>0.99 (YRI/CHBJPT)
	rs10903528	A	0.77	0.1	0.15	0.99 (YRI)	0.97 (YRI/CEU)
	rs60741147	T	0.78	0.98	0	0.98 (CEU)	0.95 (YRI/CEU)
	rs10794743	A	0.38	0.98	0.95	>0.99 (CEU)	>0.99 (YRI/CEU)
	rs4880500	C	0.36	0.98	0.96	>0.99 (CEU)	>0.99 (YRI/CEU)
	rs4880820	A	0	0.19	0.29	0.99 (CHBJPT)	>0.99 (YRI/CHBJPT)
	rs11597169	A	0.03	0.47	0.56	0.99 (CHBJPT)	0.97 (YRI/CHBJPT)
	rs11598750	C	0.04	0.46	0.55	>0.99 (CHBJPT)	0.95 (YRI/CHBJPT)

[a]Derived allele frequency;
[b]Population showing signatures of selection.

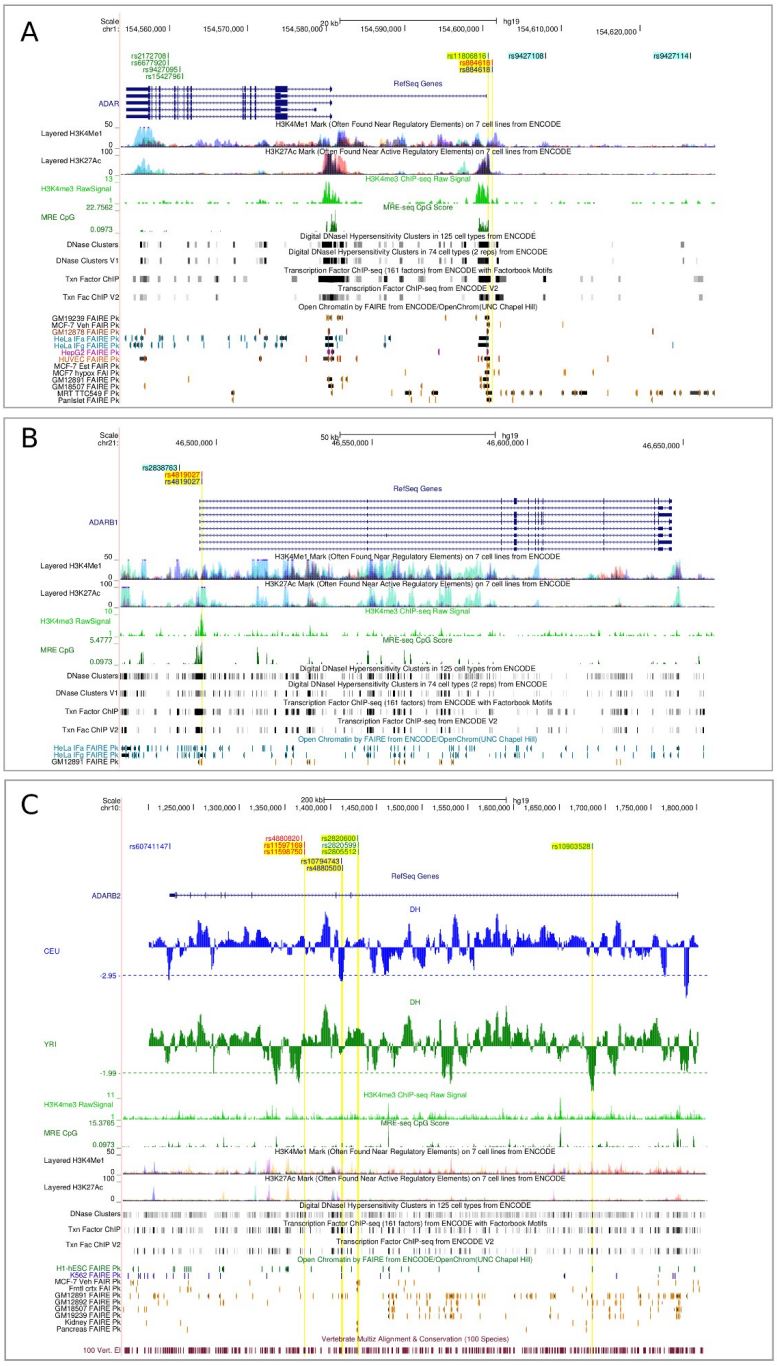


Figure 3. Location of the most likely seection targets in human populations. Candidate targets are shown for *ADAR* (A), *ADARB1* (B), *ADARB2* (C) within the UCSC Genome Browser view. Relevant annotation tracks are shown. For *ADARB2* a sliding-window analysis of DH is also shown in green (YRI) and blue (CEU). The horizontal dashed line represents the 5th percentile of DH. Variants in blue, red and green represent selection targets in CEU, CHBJPT, and YRI, respectively. Additional color codes are as follows: yellow highlight indicates SNPs mapping to regulatory elements; cyan indicates eQTL.

transcription factor binding sites have been annotated together with histone marks associated with regulatory elements and DNase hypersensitivity peaks (Fig. 3A). In CEU and CHBJPT, rs884618 is in full linkage disequilibrium (LD) (r^2=1 in both populations) with an eQTL SNP (rs9427108) in naive CD14+ monocytes;[35] the selected variant also shows LD (r^2= 0.78 and 1 in CEU and CHBJPT, respectively) with rs9427114, which acts as an interferon-dependent eQTL in monocytes (Fig. 3A and Table 3).[35] As for YRI, several linked variants (r^2>0.85) were found to be outliers in the DIND test and F_{ST} distributions (Fig. 3A, Table 3). One of these is

located close to rs884618 and falls within regulatory elements (Fig. 3A).

Similarly to *ADAR*, we detected one variant in *ADARB1* which represents the likely selection target in CEU and CHBJPT. Indeed, rs4819027 represented a DIND outlier in both populations and had very high F_{ST} in comparisons with YRI (Table 3). The SNP is located in the 5' portion of the first intron, where ENCODE data indicate the presence of regulatory elements (Fig. 3B). Interestingly, analysis of brain methylation and histone modification patterns showed that rs4819027 also maps to a region where unmethylated CpGs and H3K4me3 marks are located; these signals are associated with active transcription. Finally, rs4819027 is in moderate LD (r^2= 0.73 and 0.80 in CEU and CHBJPT, respectively) with rs2838763, an IFN-induced eQTL for *ADARB1* in monocytes.35s for *ADARB2*, 5 different selection signals were detected. In CEU, two high-frequency sweeps were detected at rs60741147 and rs10794743/rs4880500 (Table 3).

A rs60741147 and rs10794743/rs4880500 are in no LD (r^2=0) and in all cases variants mapped to DH valleys in CEU. rs60741147 is located within the 3' UTR, in a region extremely conserved in mammals (Fig. 3C). Likewise, rs10903528 was in a DH valley for YRI (Fig. 3C). In this population, a second lower frequency event was detected, which involved variants rs2820600/rs2820599/rs2805512 (in no LD with rs10903528). Finally, in CHBJPT three nearby variants were found to represent DIND and F_{ST} outliers (Table 3 and Fig. 3C). In most instances selected variants were found to map within ENCODE regulatory elements (Fig. 3C).

A minority of editing events occurs at highly conserved nonsynonymous sites

We next analyzed the evolutionary history of ADAR/ADARB1 editing sites located in human coding regions. To this aim, we retrieved A-to-I editing sites from the RADAR database (http://rnaedit.com/).[36] We limited our analysis to sites within coding regions and located outside of repetitive elements; editing sites corresponding to SNP positions (in the dbSNP137 database) were removed. The remaining sites were divided based on the change the editing causes (i.e. synonymous or nonsynonymous substitution). We also distinguished editing events based on their conservation in mammals. In particular, we designated as “shared” editing events that occur in at least one other species (among chimpanzee, macaque, and mouse; n shared sites=69); we refer to events that have been described in humans only as “non-shared” (n non-shared =664). Due to the small sample size (n shared=17), editing sites that entail synonymous substitutions were not separated based on their being shared or not.

The Genomic Evolutionary Rate Profiling (GERP) score, which measures the base-wise conservation across mammals, was next used to evaluate sequence conservation at the editing site and at flanking synonymous and nonsynonymous positions (4 codon extension at both sides).For non-shared events, editing sites at nonsynonymous positions were found to be significantly less conserved compared to flanking positions (Fig. 4A); the same occurred for editing sites at synonymous positions (Fig. 4B), but not at shared nonsynonymous editing sites (Fig. 4A, *p* values not shown), possibly due to the small sample size (n=52). As a further comparison, 1000 positions (reference sites), with flanking synonymous and nonsynonymous sites were randomly selected from the set of genes harboring the editing events. Analysis of GERP scores indicated that for both the synonymous and nonsynonymous editing events, the regions surrounding editing sites were significantly less conserved than the reference sites; this was observed at both shared and non-shared editing sites for nonsynonymous changes (Wilcoxon rank sum test, two-tailed, *p* values = 0.0064 and 2.2×10^{-16}, respectively; Fig. 4A) and at editing sites that cause synonymous changes (Wilcoxon rank sum test, two-tailed, *p* value= 0.0024, Fig. 4B).

In order to gain further insight, we compared GERP score distributions at editing sites to those deriving from 100 random samples of the same number of synonymous and nonsynonymous sites. In particular, site samples were randomly drawn from the same set of genes as those where the editing events occur (see Materials and Methods). Results confirmed a general shift of editing sites towards lower conservation scores (Fig. 4C). Nonetheless, for both shared and non-shared

editing events that cause nonsynonymous substitutions, the distribution was significantly wider than that of random samples, with a fraction of sites (and flanking positions) showing very high GERP scores (Fig. 4C). This was not observed for editing sites (and flanks) that cause synonymous changes.

To summarize, the editing site and its flanking positions are significantly less conserved than the average, both when the event causes a synonymous and a nonsynonymous substitution. Nonetheless, a fraction of editing events determine nonsynonymous substitution at highly conserved positions; this is not the case for events that cause synonymous changes and is not depended upon sharing of editing events among mammals.

We next used the WebGestalt tool to assess whether nonsynonymous edited sites showing low and high conservation scores impinge on specific pathways or biological processes.[37] Two related pathways, host interactions of HIV factors/HIV infection, were significantly enriched for genes that carry highly conserved editing sites (in the top 10% of GERP scores) compared to the overall set of edited genes (i.e. genes that carry at least one nonsynonymous editing site) (Table 4). In this same set, GO terms related to hair cycle and hair follicle/epidermis development were also enriched. No significant difference was observed for genes edited at poorly conserved positions.

Discussion

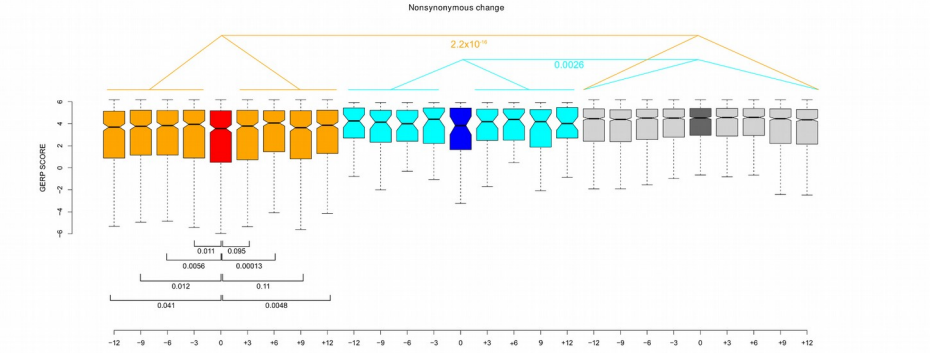
Recent technological advances are providing an increasingly detailed picture of the extent, regulation and location of A-to-I editing events in the genome of humans and other species. More than one million editing sites have currently been described in humans and RNA editing has been implicated in a number of processes and diseases.[2,5]

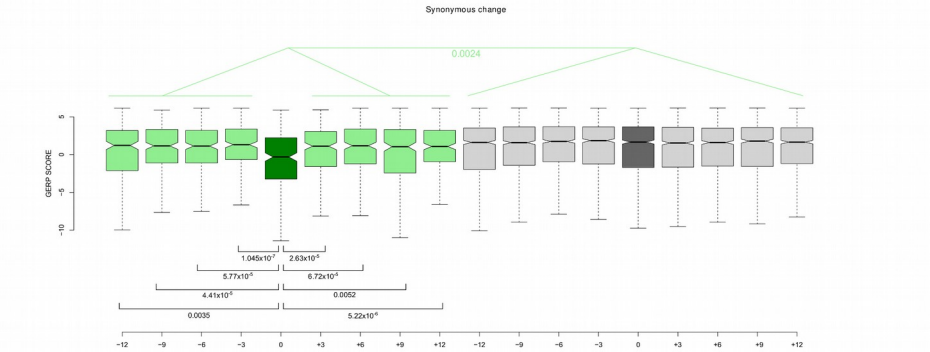
Results indicated that over diverse time periods, diversity at the three genes has been mainly shaped by purifying selection. We detected stronger constraint at ADARB1 and ADARB2 compared to ADAR, although this observation most probably reflects our using the long IFN-induced isoform of ADAR (encoding ADARp150) in both the SLAC and gammaMap analyses. Indeed, we show that ADAR evolved adaptively in primates and that most positively selected sites are located in the Zα domain-containing N-terminal portion specific to ADARp150. The Zα domain is functionally active (whereas Zβ is not) and can bind both dsDNA and dsRNA in a Z conformation.[38] Similar Z-DNA binding domains are generally found in proteins that participate in the interferon response pathway [39]. In line with its INF-responsiveness, ADAR has been suggested to play a role during viral infection, although both antiviral and proviral effects have been described.[4]

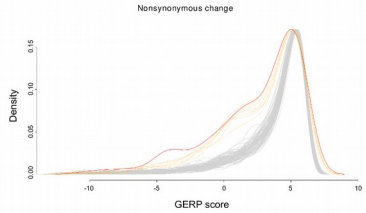
The N-terminal portion of ADARp150 also carries a NES that overlaps with the Zα domain and drives ADAR shuffling from the nucleus to the cytoplasm.[40,41] Thus, the N-terminal protein region widens the activity range of ADAR in terms of substrate recognition and cellular localization. One of the positively selected sites we identified is located within the NES, suggesting that it modulates the level or timing of nuclear-cytoplasmic transport. In turn, the cytoplasmic localization of ADARp150 might be relevant to viral detection and binding, as well as to stress response, as the protein localizes to stress granules. These structures form during viral infection or, more generally, during cell stress conditions, and also contain other editing enzymes such as APOBEC family members.[42] Thus, the variation pattern at ADAR in primates suggests that the selective pressures acting on the gene are related to its roles in immune or stress responses.

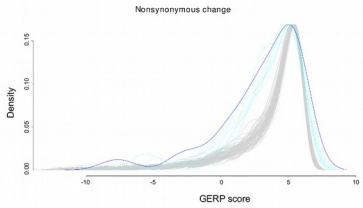
Conversely, the selection signal identified for the human lineage was located in the ADAR deaminase domain and positively selected sites were also detected for ADARB1 and ADARB2We note that,although gammaMap detected positively selected sites in humans but non in chimpanzees, the much larger sample size of human chromosomes compared to *P. troglodytes* might partially account for the different pattern in the two species.

The positively selected site in ADAR (I1111) is immediately adjacent to two positions that were found to be mutated in AGS patients (Y1112F and D1113H) (Fig. 1).[6] These residues lie along the dsRNA interaction surface and the AGS mutations did not alter the editing of a known ADAR substrate in an in vitro assay.[6] Thus, the pathogenic substitutions were hypothesized to act in a substrate- or cell type-specific manner. Clearly, this also represents an attractive possibility for the


Nonsynonymous change
2.2x10⁻¹⁶
0.0026
GERP SCORE
0.011
0.095
0.0056
0.00013
0.012
0.11
0.041
0.0048


Synonymous change
0.0024
GERP SCORE
1.045x10⁻⁷
2.63x10⁻⁵
5.77x10⁻⁵
6.72x10⁻⁵
4.41x10⁻⁵
0.0052
0.0035
5.22x10⁻⁶


Nonsynonymous change
Density
GERP score


Nonsynonymous change
Density
GERP score

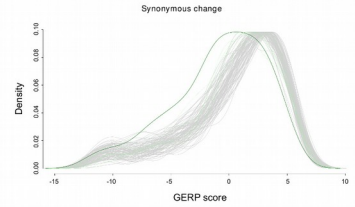

Synonymous change
Density
GERP score

Figure 4. Conservation at ADAR editing sites. (A) Box plot representation of GERP conservation scores for A-to-I editing events that cause nonsynonymous substitutions. Red and orange denote "non-shared" editing sites and their flanking sites, respectively; blue and cyan indicate "shared" editing sites and their flanking sites; dark gray indicates control nonsynonymous positions with their flanking codons in light gray (see text). Wilcoxon rank sum test (two-tailed) *p* values are also reported. (B) Box plot representation of GERP conservation scores for A-to-I editing events that cause synonymous substitutions. Dark and light green indicate editing sites and their flanking regions, respectively; dark gray indicates control synonymous positions, with light gray indicating their flanking sites. Wilcoxon rank sum test (two-tailed) *p* values are reported. (C) Distributions of GERP scores at editing-sites are reported for "non-shared" and "shared" nonsynonymous editing sites, as well as for synonymous editing sites. Color codes are as in the previous panels, with 100 random control distributions in gray. Flanking sites are represented with dashed lines.

positively selected site we detected in humans, as it might modulate editing at human-specific sites. nterestingly, we found the corresponding residue to be targeted by selection in human ADARB1 and ADARB2. The sites are located at the C-terminus of the highly conserved α2 helix structure of the second dsRNA binding domain (RBD2). Analysis of homologous domains indicated that C-terminal extensions of this helix can affect the RNA binding capability of the entire RBD domain [43].

I Furthermore RBD2 is the homologous of the third ADAR RBD, which was demonstrated to contribute to the corrected combination of a nuclear localization signal formed by two flexible fragments flanking the folded domain.[44] An interesting possibility is that the corresponding sites in the two proteins evolved in our species to modulate binding to a common interactor.

Population genetics analysis in humans revealed that the three *ADAR* genes were targeted by selection during the most recent history of human populations, as well. The approach we applied to detect selection is based on the integration of two tests, DIND and F_{ST}, which rely on distinct signatures left by selective sweeps, namely haplotype homozygosity and population genetic differentiation, respectively. As mentioned above, the combined use of distinct tests is expected to afford higher resolution in detecting the causal variant underlying the adaptive phenotype and to reduce the rate of false positive signals.[29-31] As for DH, which has more power than the original Fay and Wu's H statistic, it was used as a confirmatory test.[34] This choice was motivated by the difficulty of assessing statistical significance in sliding-window analyses, as multiple non-independent tests are performed. DH has very good power for high-frequency sweeps:indeed, the three *ADARB2* selected alleles at frequency >0.75 (rs60741147 and rs10794743 in CEU, and rs10903528 in YRI) were all located within DH valleys (Fig. 4C).[34]

DH reaches values lower than the 5th percentile in few relatively small regions along *ADARB2* (4 and 5 valleys in CEU and YRI, respectively) (Fig. 4C) and is based on a feature partially independent from population genetic differentiation and haplotype homozygosity. Thus, the DH results do provide support to the strategy we applied to detect selective events. Additional confirmation comes from the observation that most selected variants are located within regions with regulatory function, as assessed by ENCODE annotations.[45] In the case of *ADAR* and *ADARB1* the selection targets were also found to be in partial or full LD with previously described eQTLs. Overall, these data are in agreement with recent analyses indicating that regulation of gene expression is a major determinant of phenotypic variation in our species, as well as a common target of natural selection.[46,47]

In particular, the positively selected variants in *ADAR* and *ADARB1*, both shared by CEU and CHBJPT, are in linkage with eQTLs described in CD14+ monocytes, suggesting that a major selective pressure underlying adaptive evolution at these genes is accounted for by infectious agents. Interestingly, the positively selected *ADARB1* variant also maps to a region which likely regulates transcription in the brain. Given the central role of this enzyme in the editing of brain-specific genes, these data warrant further analysis of the modulatory effects of the two SNP alleles. *ADARB2* is also (and preferentially) expressed in the brain, but its biological role and regulation are poorly understood. We detected multiple sweep events at this gene, with diverse variants targeted in the same and in distinct populations, suggesting strong selective pressure.

Overall, data herein indicate that *ADAR* family genes were targeted by positive selection in primates, in the human lineage, and in the recent history of human populations. These findings and a wealth of previous data suggest a central role for these enzymes in physiological and pathological processes, a role at least partially mediated by the specific editing of coding sites with well-known effects.[3,5,48] These include the already cited GLUR2 Q/R site, several

nonsynonymous editing sites in HTR2C (serotonin receptor), an S/G recoding event in AZIN1 which predisposes to hepatocellular carcinoma, and an editing event in NEIL1 that alters the enzyme's specificity.[49,50] Beside these and a few more examples, though, the scenario of A-to-I editing in coding regions and its overall significance have remained elusive. We reasoned that further insight into this issue might be gained through an evolutionary analysis of ADAR/ADARB1 coding targets.

Recently, a study in macaques indicated that purifying selection is the major force acting at editing sites and at their flanking positions. [49] In partial agreement, a previous analysis had suggested that editing sites are less conserved across primates than their flanks, but that the overall region carrying the editing site is more conserved than control sequences.[51] Herein we used a score of conservation across mammals and we focused on editing sites located in coding regions by separately analyzing editing events that originate synonymous and nonsynonymous substitution, to account for underlying variation in sequence conservation in coding sequences. In contrast to previous reports, our findings indicate that the editing site is less conserved than its flanks which, in turn, are more variable than control positions randomly drawn from the edited coding regions. This effect is observed at shared and non-shared nonsynonymous sites, as well as at positions that entail synonymous substitutions when edited. The possible reasons for these discrepancies are manifold. With respect to Bahn and coworkers' data, the overwhelming majority of editing sites they analyzed was accounted for by non-coding sites. As for the macaque editome data, the authors analyzed fewer than 30 editing sites in coding regions shared among macaques, chimpanzees and humans; sequence conservation was measured in terms of human-macaque percentage identity or dS.[49] Also, the authors did not consider the effect of editing on the protein sequence (whether or not recoding occurs) and did not use a comparison with control sequences. As suggested, the lower conservation we observed at the editing site might reflect fixation of the edited form in some mammals or correction of G-to-A mutations through editing.[51] Nonetheless, these possibilities do not explain why flanking positions are less conserved than control sites. We suggest that most editing events are slightly deleterious and are therefore counter-selected within regions that are poorly tolerant to change and thus, by definition, highly conserved. This might be the case for both synonymous and nonsynonymous editing events, with the former excluded from regulatory regions (e.g. in splicing regulatory elements) and the latter counter-selected in constrained protein regions.

These conclusions are in agreement with a recent work showing that in humans the frequency and level of editing is lower at nonsynonymous than at synonymous sites, and that recoding events are rarer in essential genes or in genes subject to strong functional constraint (measured as dN/dS).[12] Based on these and other observations the authors proposed that most recoding events are deleterious byproducts, although few events might be functionally relevant and beneficial.[12] In fact, analysis of GERP score distributions revealed that a minority of recoding events does occur at highly conserved positions. These might represent the functional fraction. This hypothesis is supported by the identification of enriched pathways and process for genes that harbor recoding events at highly conserved positions. We note that the functionally relevant recoding events we mentioned above (in GLUR2, HTR2C, AZIN1, and NEIL1) all occur at highly conserved positions (GERP score > 4.5), but are not included in the enrichment analysis because they remain below the 90th percentile threshold (GERP score = 5.7) we set.

Interestingly, the significant pathways and processes we detected are related to ADAR/ADARB1 known functions. In HIV-1 infection both antiviral and proviral effects have been described for ADAR.[52-55] The three genes that contribute to the pathway (*PMSC4*, *ANAPC7*, and *NUPL2*) represent host factors for HIV-1 replication. Unlike restriction factors, which are specifically devoted to antiviral response and often fast-evolving, host factors carry out central physiological functions and are exploited by the virus for infection.[56] Therefore, genes coding for host factors usually evolve under purifying selection. In this respect editing might represent an advantage on the host side to allow some level of variability that may affect the viral replication process with a low overall fitness cost. It will be interesting to evaluate

whether editing at these genes can affect HIV-1 infection or replication efficiency.

Finally, genes harboring recoding events are involved in epidermis development and hair cycle/hair follicle development. As mentioned above, mutations in ADAR are responsible for dyschromatosis symmetrica hereditaria (DSH), a pigmentary skin disease. More recently, hair anomalies have also been described in patients with DHS.[57] Consistently, a conditional mouse model that lacks *ADAR* expression in the epidermis shows fur loss and skin pathology. In particular, epidermal necrosis and abnormal hair follicles were evident in these animals.[58] The pathogenic mechanism underlying DSH is poorly understood, but is thought to result from loss or decreased editing at specific target genes. Those we identified herein represent promising candidates.

In summary, our data indicate that both *ADAR* family genes and their targets evolved under variable selective regimes, including purifying and positive selection. We suggest that pressures related to immune response were major drivers of evolution for *ADAR* genes and, possibly, some of their targets. These analyses do not support nor dismiss the previous suggestion whereby A-to-I RNA editing contributed to the development of higher brain functions in humans.[8,10] Further analyses will be necessary to clarify this issue, although result herein do not reveal exceptionally fast evolution at *ADAR* genes in humans compared to other primates.

Materials and Methods

Evolutionary analysis in primates

Primate sequences for *ADAR*, *ADARB1* and *ADARB2* were retrieved from the Ensembl and NCBI databases (http://www.ensembl.org/index.html; http://www.ncbi.nlm.nih.gov/). All primate genes represented 1-to-1 orthologs of the human genes, as reported in the EnsemblCompara GeneTrees database (Tab. S1).[59] This information was not available for *Papio hamadryas*, *Macaca fascicularis*, and *Saimiri boliviensis*. BLAT search of the three *ADAR* gene coding sequences against the genome of these species (genome assemblies: Pham_1.0, MacFas_5.0, and SaiBol1.0) was performed using the Ensembl BLAST/BLAT utility; in all cases hits were consistent with the presence of a single ortholog, with no evidence of gene duplication.

A phylogenetic tree of the three ADAR family proteins was constructed using phyML with the best-fitting model (JTT plus gamma-distributed rates) generated by ProtTest 3.[60,61] The list of species and the phylogenetic tree are reported in Table S1 and in Figure S2, respectively.

For the analysis of positive selection, DNA alignments were performed using the RevTrans 2.0 utility, which uses the protein sequence alignment as a scaffold to construct the corresponding DNA multiple alignment.[13] This latter was checked and edited by hand to remove alignment uncertainties. Alignments were first screened for the presence of recombination breakpoints using GARD (Genetic Algorithm Recombination Detection);[14] the average nonsynonymous substitution/synonymous substitution rate ratio (dN/dS, also referred to as ω) was calculated using the single-likelihood ancestor counting (SLAC) method.[15]

The site models implemented in PAML have been developed to detect positive selection affecting only a few aminoacid residues in a protein. To detect selection, site models that allow (M2a, M8) or disallow (M1a, M7) a class of sites to evolve with ω >1 were fitted to the data using the F3x4 and the F61 codon frequency models. For these analyses we used trees generated by maximum-likelihood using the program PhyML.[18,60]

Positively selected sites were identified using the Bayes Empirical Bayes (BEB) analysis (with a cut-off of 0.90), which calculates the posterior probability that each codon is from the site class with ω >1 (under model M8).[19] A second method, the Mixed Effects Model of Evolution (MEME) (with the default cutoff of 0.1), which allows the distribution of ω to vary from site to site and from branch to branch at a site, was applied.[21]

To explore possible variations in selective pressure among different lineages, we applied the free-ratio models implemented in the PAML package: the M0 model assumes all branches to have the same ω, whereas M1 allows each branch to have its own ω.[17] The models are compared through likelihood-ratio tests (degree of freedom= total number of branches -1). In order to identify specific branches with a proportion of sites evolving with ω>1, we used

Table 4. Pathway and GO term enrichment analysis for genes carrying recoding events at highly conserved positions.

Pathway commons analysis			
Pathway Name	**N of Significant Genes[a]**	**Contributing genes**	**Corrected *p* value[b]**
Host Interactions of HIV factors	3	*NUPL2, ANAPC7, PSMC4*	0.0156
HIV Infection	3	*NUPL2, ANAPC7, PSMC4*	0.0156
KEGG pathway			
Pathway Name	**N of Significant Genes[a]**	**Contributing genes**	**Corrected *p* value[b]**
—	—	—	—
Gene ontlogy (GO)			
Term (GO ID)	**N of Significant Genes[a]**	**Contributing genes**	**Corrected *p* value[b]**
Hair cycle (GO:0042633)	5	*DYNC1H1, EGFR, INHBA, MYO5A, PSMC4*	0.0097
Hair cycle process (GO:0022405)	5	*DYNC1H1, EGFR, INHBA, MYO5A, PSMC4*	0.0097
Molting cycle process (GO:0022404)	5	*DYNC1H1, EGFR, INHBA, MYO5A, PSMC4*	0.0097
Molting cycle (GO:0042303)	5	*DYNC1H1, EGFR, INHBA, MYO5A, PSMC4*	0.0097
Epidermis development (GO:0008544)	5	*DYNC1H1, EGFR, INHBA, MYO5A, PSMC4*	0.0097
Hair follicle development (GO:0001942)	5	*DYNC1H1, EGFR, INHBA, MYO5A, PSMC4*	0.0097

[a]Number of genes in pathway/process
[b]Benjamini-Hochberg corrected *p* value

BS-REL with the PhyML-generated tree.[23] Branches identified using this approach were cross-validated with the branch-site likelihood ratio tests from PAML (the so-called modified model A and model MA1, "test 2").[24] A false discovery rate (FDR) correction was applied to account for multiple hypothesis testing (i.e. we corrected for the number of tested lineages), as suggested.[25] BEB analysis from MA (with a cut-off of 0.90) was used to identify sites that evolved under positive selection on specific lineages.[24]

GARD, SLAC, MEME and BS-REL analyses were performed through the DataMonkey server (http://www.datamonkey.org) or run locally (through HyPhy).[14,15, 21 ,62]

Population genetics-phylogenetics analysis

For gammaMap analysis [63], we assumed θ (neutral mutation rate per site), k (transitions/transversions ratio), and T (branch length) to vary among genes following log-normal distributions. For each gene we set the neutral frequencies of non-STOP codons (1/61) and the probability that adjacent codons share the same selection coefficient (p=0.02). For selection coefficients we considered a uniform Dirichlet distribution with the same prior weight for each selection class. For each gene we run 10,000 iterations with thinning interval of 10 iterations.

Population genetics analyses

Genotype data from the Pilot 1 phase of the 1000 Genomes Project were retrieved from the dedicated website, organized in a MySQL database, and analyzed according to selected regions/populations; these analyses were performed using the GeCo++and the libsequence libraries.[26, 64,65]

The pairwise F_{ST} and the DIND (Derived Intra-allelic Nucleotide Diversity) test were calculated for all SNPs mapping to *ADAR*, *ADARB1* and *ADARB2*, as well as for SNPs mapping to a control set of ~1,000 genes; these latter were used as a reference, as previously described.[66, 32, 30,31]

F_{ST} values are not independent from allele frequencies, so we binned variants based on their minor allele frequency (MAF, 50 classes) and calculated F_{ST} empirical distributions for each MAF class. The same procedure was applied for the DIND test; thus, we calculated statistical significance by obtaining an empirical distribution of DIND values for variants located within control genes; in particular, the DIND test was calculated using a constant number of 20 upstream and downstream flanking variants, as previously described.[30,31] DIND values for YRI, CEU and AS were binned in derived allele frequency (DAF) intervals (100 classes) and for each class the distributions were calculated. As suggested, for values of $i\pi_D=0$ we set the DIND value to the maximum obtained over the whole dataset plus 20. Only SNPs with both F_{ST} and DIND with a percentile rank >0.95 were considered as selection targets.[32]

DH was calculated in 5kb sliding windows moving with a step of 500 bp.[34,67] Sliding window analyses have an inherent multiple

testing problem that is difficult to correct because of the non-independence of windows. In order to partially account for this limitation, we calculated DH also for the control gene set, and the distribution of the statistic was obtained for the corresponding windows. This allowed calculation of the 5th percentile and the identification of regions below this threshold.
LD was calculated through the SNAP utility (http://www.broadinstitute.org/mpg/snap/). [68]

ADAR editing sites analysis
We retrieved A-to-I editing sites from the RADAR database (http://rnaedit.com/), limiting our search to sites within coding regions and located outside of repetitive elements.[36] Information concerning the presence of the same editing event in other species (chimpanzee, macaque, and mouse) was also based on RADAR annotations.
The Genomic Evolutionary Rate Profiling (GERP) scorewas obtained from UCSC tables (table name: GERP Scores for Mammalian Alignments) and used to evaluate conservation: positive scores represent a deficit in substitutions and indicate evolutionary constraint.[69] To generate 100 comparison distributions, nonsynonymous and synonymous positions were randomly drawn from the same genes harboring the editing events. In particular, distributions were generated by 100 resamplings of the same number of positions as the number of non-shared and shared editing events for nonsynonymous changes (n=443 and n=52, respectively), and for all editing sites for synonymous changes (n=238).

GO term and pathway enrichment
To evaluate whether low and high conserved nonsynonymous editing sites are involved in specific pathways or biological processes we used the WEB-based GEne SeT AnaLysis Toolkit.[37] Specifically, unique gene lists that carry highly or poorly conserved editing sites (in the 10% tails of GERP score distributions, Table S4) were used as queries; the background list was accounted for by the total set of genes carrying recoding events. The 87% of recoding events occurred in distinct genes (one event/gene). When the same gene carried more that one recoding event, it was counted only once in the relative list (for instance, if the same gene presented two recoding events at highly conserved positions, it was included only once in the highly conserved list). The minimum number of genes for a category was set to 3, and we applied a Benjamini and Hochberg correction for multiple testing. We queried for enrichment in GO categories, KEGG pathways and in the Pathway Commons database.

Acknowledgements

CP and DF are supported by fellowships of the Doctorate School of Molecular and Translational Medicine, University of Milan

Conflict of Interest Statements
The authors have no financial, personal, or professional interest related to this work.

References
1. Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem 2010;79:321-349; PMID:20192758.
2. Bazak L, Haviv A, Barak M, Jacob-Hirsch J, Deng P, Zhang R, Isaacs FJ, Rechavi G, Li JB, Eisenberg E, et al. A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. Genome Res 2014;24:365-376; PMID:24347612.
3. Savva YA, Rieder LE, Reenan RA. The ADAR protein family. Genome Biol 2012;13:252; PMID:23273215.
4. Samuel CE. ADARs: Viruses and innate immunity. Curr Top Microbiol Immunol 2012;353:163-195; PMID:21809195.
5. Slotkin W, Nishikura K. Adenosine-to-inosine RNA editing and human disease. Genome Med 2013;5:105; PMID:24289319.
6. Rice GI, Kasher PR, Forte GM, Mannion NM, Greenwood SM, Szynkiewicz M, Dickerson JE, Bhaskar SS, Zampini M, Briggs TA, et al. Mutations in ADAR1 cause aicardi-goutieres syndrome associated with a type I interferon signature. Nat Genet 2012;44:1243-1248; PMID:23001123.
7. Daniel C, Silberberg G, Behm M, Ohman M. Alu elements shape the primate transcriptome by cis-regulation of RNA editing. Genome Biol 2014;15:R28;

PMID:24485196.
8. Li JB, Church GM. Deciphering the functions and regulation of brain-enriched A-to-I RNA editing. Nat Neurosci 2013;16:1518-1522; PMID:24165678.
9. Pinto Y, Cohen HY, Levanon EY. Mammalian conserved ADAR targets comprise only a small fragment of the human editosome. Genome Biol 2014;15:R5; PMID:24393560.
10. Paz-Yaacov N, Levanon EY, Nevo E, Kinar Y, Harmelin A, Jacob-Hirsch J, Amariglio N, Eisenberg E, Rechavi G. Adenosine-to-inosine RNA editing shapes transcriptome diversity in primates. Proc Natl Acad Sci U S A 2010;107:12174-12179; PMID:20566853.
11. Gommans WM, Mullen SP, Maas S. RNA editing: A driving force for adaptive evolution? Bioessays 2009;31:1137-1145; PMID:19708020.
12. Xu G, Zhang J. Human coding RNA editing is generally nonadaptive. Proc Natl Acad Sci U S A 2014;111:3769-3774; PMID:24567376.
13. Wernersson R, Pedersen AG. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res 2003;31:3537-3539; PMID:12824361.
14. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol 2006;23:1891-1901; PMID:16818476.
15. Kosakovsky Pond SL, Frost SD. Not so different after all: A comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 2005;22:1208-1222; PMID:15703242.
16. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 2011;478:476-482; PMID:21993624.
17. Yang Z. PAML: A program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 1997;13:555-556; PMID:9367129.
18. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol 2007;24:1586-1591; PMID:17483113.
19. Anisimova M, Bielawski JP, Yang Z. Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol 2002;19:950-958; PMID:12032251.
20. Yang Z, Wong WS, Nielsen R. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol 2005;22:1107-1118; PMID:15689528.
21. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. Detecting individual sites subject to episodic diversifying selection. PLoS Genet 2012;8:e1002764; PMID:22807683.
22. Yang Z, Nielsen R. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 1998;46:409-418; PMID:9541535.
23. Kosakovsky Pond SL, Murrell B, Fourment M, Frost SD, Delport W, Scheffler K. A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol 2011;28:3033-3043; PMID:21670087.
24. Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 2005;22:2472-2479; PMID:16107592.
25. Anisimova M, Yang Z. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol 2007;24:1219-1228; PMID:17339634.
26. 1000 Genomes Project Consortium, Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature 2010;467:1061-1073; PMID:20981092.
27. Auton A, Fledel-Alon A, Pfeifer S, Venn O, Segurel L, Street T, Leffler EM, Bowden R, Aneas I, Broxholme J, et al. A fine-scale chimpanzee genetic map from population sequencing. Science 2012;336:193-198; PMID:22422862.
28. Quach H, Wilson D, Laval G, Patin E, Manry J, Guibert J, Barreiro LB, Nerrienet E, Verschoor E, Gessain A, et al. Different

selective pressures shape the evolution of toll-like receptors in human and african great ape populations. Hum Mol Genet 2013;22:4829-4840; PMID:23851028.
29. Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, Griesemer D, Karlsson EK, Wong SH, et al. Identifying recent adaptations in large-scale genomic data. Cell 2013;152:703-713; PMID:23415221.
30. Forni D, Cagliani R, Tresoldi C, Pozzoli U, De Gioia L, Filippi G, Riva S, Menozzi G, Colleoni M, Biasin M, et al. An evolutionary analysis of antigen processing and presentation across different timescales reveals pervasive selection. PLoS Genet 2014;10:e1004189; PMID:24675550.
31. Forni D, Cagliani R, Pozzoli U, Colleoni M, Riva S, Biasin M, Filippi G, De Gioia L, Gnudi F, Comi GP, et al. A 175 million year history of T cell regulatory molecules reveals widespread selection, with adaptive evolution of disease alleles. Immunity 2013;38:1129-1141; PMID:23707475.
32. Barreiro LB, Ben-Ali M, Quach H, Laval G, Patin E, Pickrell JK, Bouchier C, Tichit M, Neyrolles O, Gicquel B, et al. Evolutionary dynamics of human toll-like receptors and their different contributions to host defense. PLoS Genet 2009;5:e1000562; PMID:19609346.
33. Fagny M, Patin E, Enard D, Barreiro LB, Quintana-Murci L, Laval G. Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing datasets. Mol Biol Evol 2014PMID:24694833.
34. Zeng K, Fu YX, Shi S, Wu CI. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 2006;174:1431-1439; PMID:16951063.
35. Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, Jostins L, Plant K, Andrews R, McGee C, et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science 2014;343:1246949; PMID:24604202.
36. Ramaswami G, Li JB. RADAR: A rigorously annotated database of A-to-I RNA editing. Nucleic Acids Res 2014;42:D109-13; PMID:24163250.
37. Wang J, Duncan D, Shi Z, Zhang B. WEB-based GEne SeT AnaLysis toolkit (WebGestalt): Update 2013. Nucleic Acids Res 2013;41:W77-83; PMID:23703215.
38. Athanasiadis A, Placido D, Maas S, Brown BA,2nd, Lowenhaupt K, Rich A. The crystal structure of the zbeta domain of the RNA-editing enzyme ADAR1 reveals distinct conserved surfaces among Z-domains. J Mol Biol 2005;351:496-507; PMID:16023667.
39. Athanasiadis A. Zalpha-domains: At the intersection between RNA editing and innate immunity. Semin Cell Dev Biol 2012;23:275-280; PMID:22085847.
40. Poulsen H, Nilsson J, Damgaard CK, Egebjerg J, Kjems J. CRM1 mediates the export of ADAR1 through a nuclear export signal within the Z-DNA binding domain. Mol Cell Biol 2001;21:7862-7871; PMID:11604520.
41. Strehblow A, Hallegger M, Jantsch MF. Nucleocytoplasmic distribution of human RNA-editing enzyme ADAR1 is modulated by double-stranded RNA-binding domains, a leucine-rich export signal, and a putative dimerization domain. Mol Biol Cell 2002;13:3822-3835; PMID:12429827.
42. Anderson P, Kedersha N. RNA granules: Post-transcriptional and epigenetic modulators of gene expression. Nat Rev Mol Cell Biol 2009;10:430-436; PMID:19461665.
43. Masliah G, Barraud P, Allain FH. RNA recognition by double-stranded RNA binding domains: A matter of shape and sequence. Cell Mol Life Sci 2013;70:1875-1895; PMID:22918483.
44. Barraud P, Banerjee S, Mohamed WI, Jantsch MF, Allain FH. A bimodular nuclear localization signal assembled via an extended double-stranded RNA-binding domain acts as an RNA-sensing signal for transportin 1. Proc Natl Acad Sci U S A 2014;111:E1852-61; PMID:24753571.
45. ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the

human genome. Nature 2012;489:57-74; PMID:22955616.
46. Enard D, Messer PW, Petrov DA. Genome-wide signals of positive selection in human evolution. Genome Res 2014PMID:24619126.
47. Fraser HB. Gene expression drives local adaptation in humans. Genome Res 2013;23:1089-1096; PMID:23539138.
48. Tomaselli S, Locatelli F, Gallo A. The RNA editing enzymes ADARs: Mechanism of action and human disease. Cell Tissue Res 2014;356:527-532; PMID:24770896.
49. Chen JY, Peng Z, Zhang R, Yang XZ, Tan BC, Fang H, Liu CJ, Shi M, Ye ZQ, Zhang YE, et al. RNA editome in rhesus macaque shaped by purifying selection. PLoS Genet 2014;10:e1004274; PMID:24722121.
50. Yeo J, Goodman RA, Schirle NT, David SS, Beal PA. RNA editing changes the lesion specificity for the DNA repair enzyme NEIL1. Proc Natl Acad Sci U S A 2010;107:20715-20719; PMID:21068368.
51. Bahn JH, Lee JH, Li G, Greer C, Peng G, Xiao X. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. Genome Res 2012;22:142-150; PMID:21960545.
52. Doria M, Neri F, Gallo A, Farace MG, Michienzi A. Editing of HIV-1 RNA by the double-stranded RNA deaminase ADAR1 stimulates viral infection. Nucleic Acids Res 2009;37:5848-5858; PMID:19651874.
53. Doria M, Tomaselli S, Neri F, Ciafre SA, Farace MG, Michienzi A, Gallo A. ADAR2 editing enzyme is a novel human immunodeficiency virus-1 proviral factor. J Gen Virol 2011;92:1228-1232; PMID:21289159.
54. Phuphuakrat A, Kraiwong R, Boonarkart C, Lauhakirti D, Lee TH, Auewarakul P. Double-stranded RNA adenosine deaminases enhance expression of human immunodeficiency virus type 1 proteins. J Virol 2008;82:10864-10872; PMID:18753201.
55. Biswas N, Wang T, Ding M, Tumne A, Chen Y, Wang Q, Gupta P. ADAR1 is a novel multi targeted anti-HIV-1 cellular protein. Virology 2012;422:265-277; PMID:22104209.
56. Sawyer SL, Elde NC. A cross-species view on viruses. Curr Opin Virol 2012;2:561-568; PMID:22835485.
57. Kantaputra PN, Chinadet W, Ohazama A, Kono M. Dyschromatosis symmetrica hereditaria with long hair on the forearms, hypo/hyperpigmented hair, and dental anomalies: Report of a novel ADAR1 mutation. Am J Med Genet A 2012;158A:2258-2265; PMID:22821605.
58. Sharma R, Wang Y, Zhou P, Steinman RA, Wang Q. An essential role of RNA editing enzyme ADAR1 in mouse skin. J Dermatol Sci 2011;64:70-72; PMID:21788117.
59. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 2009;19:327-335; PMID:19029536.
60. Guindon S, Delsuc F, Dufayard JF, Gascuel O. Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol 2009;537:113-137; PMID:19378142.
61. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: Fast selection of best-fit models of protein evolution. Bioinformatics 2011;27:1164-1165; PMID:21335321.
62. Delport W, Poon AF, Frost SD, Kosakovsky Pond SL. Datamonkey 2010: A suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 2010;26:2455-2457; PMID:20671151.
63. Wilson DJ, Hernandez RD, Andolfatto P, Przeworski M. A population genetics-phylogenetics approach to inferring natural selection in coding sequences. PLoS Genet 2011;7:e1002395; PMID:22144911.
64. Cereda M, Sironi M, Cavalleri M, Pozzoli U. GeCo++: A C++ library for genomic features computation and annotation in the presence of variants. Bioinformatics 2011;27:1313-1315; PMID:21398667.
65. Thornton K. Libsequence: A C++ class library for evolutionary genetic analysis. Bioinformatics 2003;19:2325-2327; PMID:14630667.
66. Wright S. Genetical structure of populations. Nature 1950;166:247-249; PMID:15439261.
67. Fay JC, Wu CI. Hitchhiking under positive

darwinian selection. Genetics 2000;155:1405-1413; PMID:10880498.

68. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI. SNAP: A web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 2008;24:2938-2939; PMID:18974171.

69. Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 2005;15:901-913; PMID:15965027.

3.2.1.3 Evolutionary analysis identifies an MX2 haplotype associated with natural resistance to HIV-1 infection

The protein product of the human *MX2* (myxovirus resistance 2) gene was recently shown to act as a post-entry inhibitor of HIV-1 infection [145-147]. MX2 and its paralog MX1 (or MxA) belong to the dynamin-like large GTPase superfamily. Whereas MX1 can restrict a wide variety of viruses, MX2 is known to counteract retroviruses only [145-147]. Specifically, the human and macaque MX2 proteins efficiently restrict HIV-1 and other simian immunodeficiency viruses, but have a modest effect against retroviruses that infect non-primate species [146, 147]. This observation points to species-specific virus-host interactions that may result in an evolutionary arms races, a scenario previously described for MX1 [148].

I applied a multidisciplinary approach to study the evolutionary history of MX2, its structure, and the relevance of its variants for HIV-1 infection susceptibility in human populations.

Analysis of the evolutionary history of MX2 in placental mammals unveiled 11 sites subject to pervasive positive selection, most of which localized in a stalk domain loop (loop 4). The corresponding loop in *MX1* has been previously shown to harbor positively selected sites in primates and to confer antiviral specificity [148]. One of the positively selected sites identified in the stalk domain, corresponds to a polymorphic position in the pig coding sequence; aminoacid variation at this residue modulate the ability of the porcine MX2 protein to inhibit Vesicular Stomatitis virus (VSV) replication [149], supporting the hypothesis that positively selected sites determine antiviral activity.

Application of other statistical methods [14-16] to study lineage-specific selection and to identify selected sites for specific branches in the phylogeny indicated episodic selection in the primate and porcine lineages.

Modeling the 3D structure of MX2 using the known crystal structure of MX1 allowed further insight into the functional role of positively selected residues. This, and the use of I-TASSER [99], allowed a reliable *ab-initio* prediction of loop 4 structure, which includes two antiparallel alpha-helices forming a bump protruding from the stalk domain. Analysis of non-loop 4 selected sites in *MX2* and comparison with *MX1* selection targets [148] indicated that the same region of the stalk domain carries residues positively selected in *MX2* and in *MX1*. Interestingly, the MX2 primate-specific selection target in the GTPase domain is in spatial proximity to a site subject to diversifying selection in primate *MX1* genes.

I next applied a population genetics-phylogenetics approach to study the evolution of *MX1* and *MX2* in the human lineage. A general preponderance of codons evolving under negative selection was observed for both genes, with *MX1* showing stronger constraint. Thus, coding variants in the two genes did not represent positive selection targets during human evolutionary history.

The possible effect of natural selection on genetic diversity at *MX2* in human populations was next analyzed. To this purpose, I used an approach [55] that integrates information from the Human Genome Diversity Panel (HGDP) [150] and from the 1000 Genomes Project [86]. By applying tests based on population genetic differentiation (F_{ST}) and on haplotype homozygosity (Derived Intra-Allelic Nucleotide Diversity, DIND test, [93]), I show that distinct selective events have driven the frequency increase of two different haplotypes in populations of Asian and European ancestry. The Asian haplotype carries a susceptibility allele for melanoma [151]; the European haplotype is tagged by rs2074560, an intronic variant.

Because natural selection targets variants with a phenotypic effect, and given the role of MX2 as a HIV-1 restriction factor, I investigated whether

the allelic status at rs2074560 modulates susceptibility to HIV-1 infection. Most human subjects are susceptible to this virus, but a minority of individuals do not seroconvert despite multiple exposures (HIV-1 exposed seronegative individuals, HESN). I thus genotyped rs2074560 (G/A) in 3 independent case-control cohorts: 1) subjects from Spain who exposed themselves to the virus through injection drug use (all HCV positive, 106 HIV1-positive, 85 HESN); 2) italian heterosexual subjects who have a history of unprotected sex with their seropositive partners (88 HESN, 188 healthy controls, HC); 3) a small cohort of Spanish sexually-exposed women (n=37) and 188 Spanish HC. In all cohorts an excess of GG homozygotes was observed in HESN compared to controls. The results of the three association analyses were combined through a random effect meta-analysis, which revealed no heterogeneity among samples (Cochrane's Q *p* value= 0.66, I^2= 0) and yielded a *p* value of $1.55x10^{-4}$. Overall, these results indicate that the G allele of rs2074560 protects from HIV-1 with a recessive effect, irrespective of the infection route.

In line with the results outlined above, in vitro HIV-1 infection of PBMCs from 50 HESN subjects indicated that the G allele of rs2074560 is associated with significantly lower viral replication, as assesses by p24 antigen quantification (Kruskal-Wallis rank sum test, *p*= 0.034).

Finally, to assess the functional role of rs2074560 (or linked variants) I analyzed MX2 expression in response to IFN-α. Data from peripheral blood mononuclear cells from 45 healthy volunteers indicated that the G allele is associated with a significantly increased MX2 induction in response to IFN-α (one-way ANOVA, *p*= 0.015)

In summary, I analyzed the evolutionary history of *MX2* in mammals and in human populations and I exploited this information to identify a haplotype that modulates susceptibility to HIV-1 infection and transcript levels in

response to IFN-α. Thus, this work has general implications by showing that the analysis of selection patterns of antiviral response genes can provide valuable information on the allelic determinants of susceptibility to modern infections. The *MX2* variant reported herein represents one of the few human polymorphisms reliably associated with HIV-1 infection susceptibility. MX2 should therefore be regarded as an attractive target in preventative approaches to infection and possibly to favour clearance of latently HIV-1 infected cells.

Personal contribution to the work: I focused on the evolutionary analysis in mammals.

Evolutionary analysis identifies an *MX2* haplotype associated with natural resistance to HIV-1 infection

Manuela Sironi[1]*, Mara Biasin[2]*, Rachele Cagliani[1], Federica Gnudi[2], Irma Saulle[2], Salomè Ibba[2], Giulia Filippi[3], Sarah Yahyaei[2], Claudia Tresoldi[1], Stefania Riva[1], Daria Trabattoni[2], Luca De Gioia[3], Sergio Lo Caputo[4], Francesco Mazzotta[4], Diego Forni[1], Chiara Pontremoli[1], Juan Antonio Pineda[5], Uberto Pozzoli[1], Antonio Rivero-Juarez[6], Antonio Caruz[7], Mario Clerici[8,9]

[1] Scientific Institute IRCCS E. MEDEA, Bioinformatics, 23842 Bosisio Parini, Italy;
[2] Department of Biomedical and Clinical Sciences, University of Milan, 20157 Milan, Italy;
[3] Department of Biotechnology and Biosciences, University of Milan-Bicocca, 20126 Milan, Italy;
[4] S. Maria Annunziata Hospital, 50122 Florence, Italy;
[5] Infectious Diseases and Microbiology Clinical Unit. Valme Hospital, Seville, Spain;
[6] Maimonides Institut for Biomedical Research (IMIBIC)-Reina Sofia Universitary Hospital-University of Cordoba, Spain;
[7] Immunogenetics Unit, Department of Experimental Biology, University of Jaen, Jaen, Spain;
[8] Chair of Immunology, Department of Physiopathology and Transplantation, University of Milan, 20090 Milan, Italy;
[9] Don C. Gnocchi Foundation ONLUS, IRCCS, 20148 Milan, Italy.
* These authors equally contributed to this work

Address for correspondence: Manuela Sironi, Bioinformatics - Scientific Institute IRCCS E. MEDEA, 23842 Bosisio Parini, Italy. Tel. +39-031877915; Fax.+39-031877499; e-mail: manuela.sironi@bp.lnf.it

Abstract

The protein product of the *MX2* (myxovirus resistance 2) gene restricts HIV-1 and simian retroviruses. We demonstrate that *MX2* evolved adaptively in mammals with distinct sites representing selection targets in distinct branches; selection mainly involved residues in loop 4, previously shown to carry antiviral determinants. Modeling data indicated that positively selected sites form a continuous surface on loop 4, which folds into two antiparallel α-helices protruding from the stalk domain. A population genetics-phylogenetics approach indicated that the coding region of *MX2* mainly evolved under negative selection in the human lineage. Nonetheless, population genetic analyses demonstrated that natural selection operated on *MX2* during the recent history of human populations: distinct selective events drove the frequency increase of two haplotypes in populations of Asian and European ancestry. The Asian haplotype carries a susceptibility allele for melanoma*;* the European haplotype is tagged by rs2074560, an intronic variant. Analyses performed on three independent European cohorts of HIV-1 exposed seronegative individuals with different geographic origin and distinct exposure route showed that the ancestral (G) allele of rs2074560 protects from HIV-1 infection with a recessive effect (combined p value = 1.55×10^{-4}). The same allele is associated with lower in vitro HIV-1 replication and increases *MX2* expression levels in response to IFN-α. Data herein exploit evolutionary information to identify a novel host determinant of HIV-1 infection susceptibility.

Introduction

The protein product of the human *MX2* (myxovirus resistance 2, also known as *MxB*) gene was recently shown to act as an inhibitor of HIV-1 infection (Goujon et al. 2013; Kane et al. 2013; Liu et al. 2013b). MX2 blocks HIV-1 replication at a late post-entry step by decreasing nuclear accumulation and chromosomal integration of nascent viral DNA (Goujon et al. 2013; Kane et al. 2013; Liu et al. 2013b). HIV-1 capsid mutations that alter the

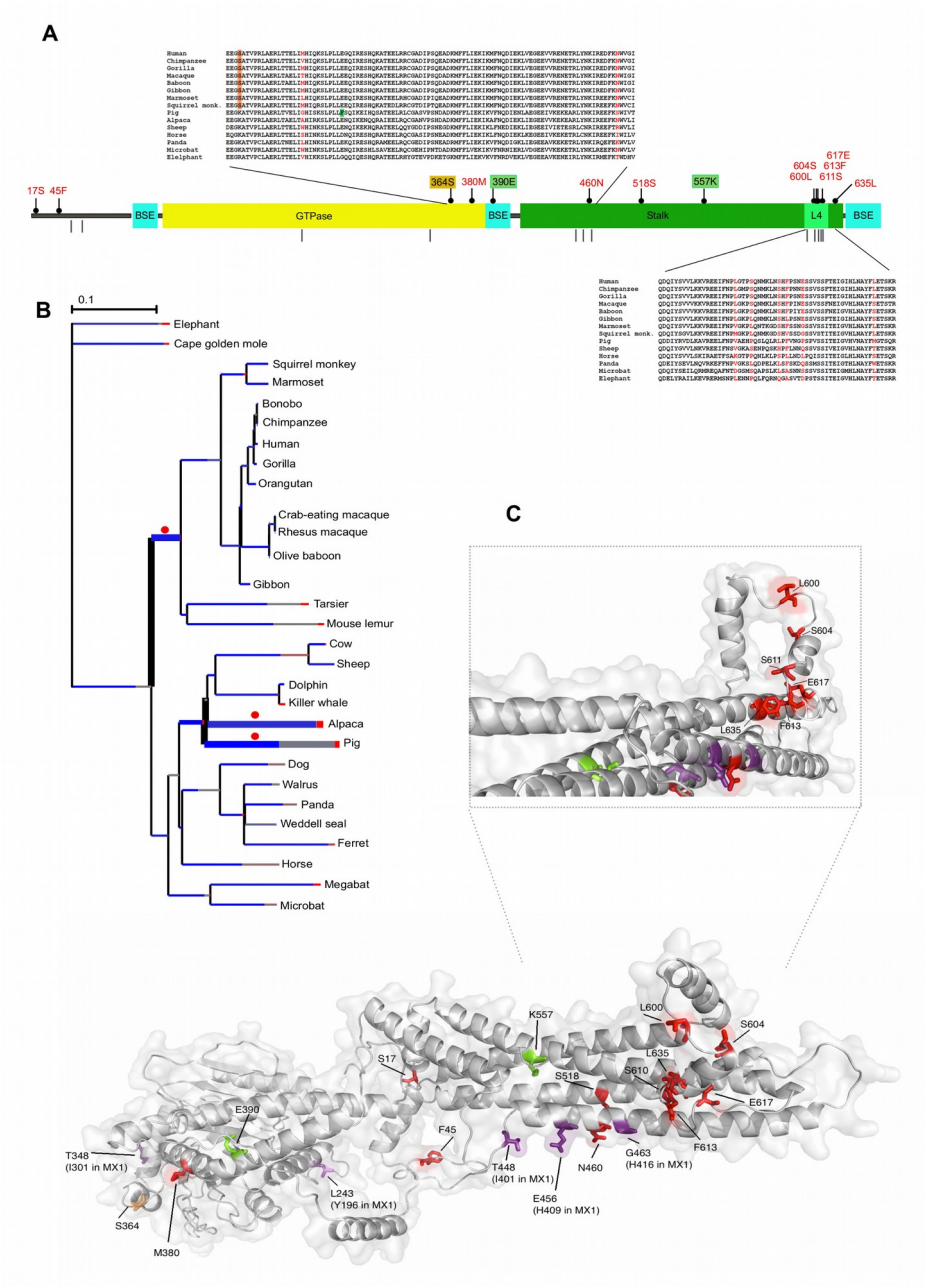


Figure 1. Adaptive evolution of *MX2* in mammals. (A) Schematic representation of the domain structure of *MX2* (cyan, bundle signaling element regions; yellow, GTPase domain, dark green, stalk domain with L4 in light green). The position of positively selected sites is shown together with sequence alignments for a few representative mammals. Positively selected sites are shown in red (sites identified in the whole phylogeny); primate- and pig-specific sites are in orange and green, respectively. Grey vertical bars below the domain cartoon denote positively selected sites in *MX1* (Mitchell et al. 2012). (B) Branch-site analysis of positive selection. Branch lengths are scaled to the expected number of substitutions per nucleotide, and branch colors indicate the strength of selection (ω). Red, positive selection (ω > 5); blue, purifying selection (ω = 0); grey, neutral evolution (ω = 1). The proportion of each color represents the fraction of the sequence undergoing the corresponding class of selection. Thick branches indicate statistical support for evolution under episodic

diversifying selection as determined by BS-REL. Red dots denote branches that were also detected to be under positive selection using the PAML branch-site models (after Bonferroni correction for multiple tests). (C) Structural model of the full-length MX2 protein predicted by I-TASSER. Positively selected sites identified in the whole phylogeny are in red, in orange and green primate- and pig-specific sites, respectively. Previously identified positively selected sites in the GTPase and stalk domains of MX1 are indicated in magenta. An enlargement of loop 4 and part of the stalk domain is also shown (with rotation).

nuclear import pathway used by the virus relieve the inhibition operated by MX2 (Goujon et al. 2013; Kane et al. 2013; Liu et al. 2013b). Thus, MX2 is thought to target the capsid and to interact with other cellular proteins such as peptidylprolyl isomerase A (cyclophilin A) to block HIV-1 nuclear import (Liu et al. 2013b).

MX2 and its paralog MX1 (or MxA) belong to the dynamin-like large GTPase superfamily and share a common structure consisting of an N-terminal GTPase domain and a C-terminal stalk connected by a bundle signaling element (Figure 1) (Haller and Kochs 2011). The protein products of the two genes display high identity but different cellular localization and diverse antiviral specificity. MX1 can restrict a wide variety of viruses including orthomyxoviruses (e.g. influenza and Thogoto), bunyaviruses (e.g La Crosse and rift valley viruses), and hepatitis B virus (Haller and Kochs 2011). Conversely, MX2 is known to counteract retroviruses only (Goujon et al. 2013; Kane et al. 2013; Liu et al. 2013b), and to exert a weak effect against vescicular stomatitis virus infection (Liu et al. 2012). Specifically, the human and macaque MX2 proteins efficiently restrict HIV-1 and other simian immunodeficiency viruses, but have a modest effect against retroviruses that infect non-primate species such as murine leukaemia virus and feline immunodeficiency virus (Goujon et al. 2013; Kane et al. 2013). This observation points to species-specific virus-host interactions that may result from evolutionary arms races (i.e. genetic conflicts between the host and the virus), a scenario previously described for *MX1*(Mitchell et al. 2012).

Herein we analyze the evolutionary history of *MX2* at the inter-and intra-specific level and use this information to identify a haplotype that associates with natural resistance to HIV-1 infection in humans.

Results

MX2 evolutionary history in mammals

To analyze the evolutionary history of *MX2* in placental mammals, we obtained coding sequence information for 29 species from public databases (Supplementary Tab. S1, Supplementary Material online). Rodent gene sequences were not included due to uncertain orthology (they likely arose by duplication from an ancestral *MX1*-like gene) (Verhelst, Hulpiau, Saelens 2013). The multiple sequence alignment was screened for recombination using GARD (Genetic Algorithm Recombination Detection) (Kosakovsky Pond et al. 2006), which detected no breakpoint. The average non-synonymous substitution/synonymous substitution rate (dN/dS, also known as ω) for *MX2* amounted to 0.39 (95% confidence intervals: 0.37, 0.41), indicating a major role for purifying selection (which leads to dN/dS values <1). However, whereas constraints on protein function and structure often result in purifying selection being the primary evolutionary force acting on gene regions, diversifying selection (dN/dS>1) might involve specific sites or domains. Indeed, this has previously been shown to be the case for primate *MX1* genes (Mitchell et al. 2012). To test this possibility we applied maximum-likelihood analyses implemented in the PAML (Phylogenetic Analysis by Maximum Likelihood) package (Yang 1997; Yang 2007). Specifically, we used the codeml program to compare models of gene evolution that allow (NSsite models M2a and M8, positive selection models) or disallow (NSsite models M1a and M7, null models) a class of codons to evolve with dN/dS >1. As reported in table 1, both null models were rejected in favor of the positive selection models; the same result was obtained using different codon frequency models (F3x4 and F61) (Table 1). Thus, *MX2* evolved adaptively in placental mammals.

To identify specific sites subject to positive selection, we applied the Bayes Empirical Bayes (BEB) method (with a cut-off of 0.90) (Anisimova, Bielawski, Yang 2002; Yang, Wong, Nielsen 2005). Because this approach may yield some false positives when a relatively large number of sequences is analyzed (Kosakovsky Pond and Frost 2005), we used the Mixed Effects Model of Evolution (MEME) (with the default cutoff of 0.1) (Murrell et al. 2012) as a second criterion. To be conservative, only sites detected by both methods were considered to be positively selected, although this may result in an

underestimation of the actual number of selected sites. These analyses allowed the identification of 11 positively selected sites in *MX2* (Figure 1A).

In order to explore possible variations in selective pressure among different lineages, we first tested whether a model that allows dN/dS to vary along branches (model M1) had significant better fit to the data than a model that assume one same dN/dS across the entire phylogeny (model M0) (Yang and Nielsen 1998). This was indeed the case (Table 1), indicating that different mammals experienced variable levels of selective pressure. We thus used the branch site-random effects likelihood (BS-REL) method (Kosakovsky Pond et al. 2011) to identify lineages on which a subset of sites have evolved under positive selection. BS-REL makes no a priori assumption about which lineages are more likely to represent selection targets; the method identified 3 branches (Figure 1B). These were cross-validated using the branch-site models implemented in PAML (Zhang, Nielsen, Yang 2005), which apply a likelihood ratio test to compare a model (MA) that allows positive selection on one or more lineages (foreground lineages) with a model (MA1) that does not allow such positive selection (Table 2). As suggested (Anisimova and Yang 2007), a Bonferroni correction was applied to these p values, as multiple hypotheses (3 in this case) are being tested on the same dataset. PAML analysis confirmed the three branches identified by BS-REL (Table 2, Figure 1B). The PAML branch-site models offer the possibility of identifying specific sites that evolved under positive selection in the foreground branches; this is achieved through implementation of a BEB analysis (Zhang, Nielsen, Yang 2005). Notably, the simultaneous inference of both the site and the branch subject to diversifying selection is difficult (Murrell et al. 2012; Zhang, Nielsen, Yang 2005); thus, BEB analysis is accurate but has low power (Zhang, Nielsen, Yang 2005). Also, because identifying sites subject to selection is more difficult than testing whether such sites exist, the branch-site test may provide statistical support for positive selection for the foreground branch(es), but BEB may fail to detect sites with a posterior probability >0.90 (Zhang, Nielsen, Yang 2005). This is the case of the alpaca lineage: we found evidence of episodic selection (Figure 1B, Table 2) but failed to identify the selected site(s). Conversely, BEB detected 1 site in the primate and 2 sites in the pig branches (Figure 1A). We reasoned that, because MEME was specifically developed to detect episodic positive selection (in addition to pervasive selection), lineage-specific BEB sites should have been identified by the MEME analysis we performed on the whole phylogeny. Indeed, the three sites identified by BEB were also detected by MEME (Figure 1A). These sites thus represent lineage-specific selection targets.

Structural modeling and analysis of MX2 selected sites

As indicated above, analyses allowed identification of 11 *MX2* sites subject to diversifying selection in the whole phylogeny; five of these are located in an unstructured loop (loop 4), where several positively selected sites were also detected in primate *MX1* genes (Mitchell et al. 2012) (Figure 1A). To gain further insight into the functional role of positively selected residues, we modeled the 3D structure of MX2 using the known crystal structure of MX1 (63% identity at the protein level) as a template. The use of I-TASSER (Roy, Kucukural, Zhang 2010) also allowed a reliable *ab-initio* prediction of loop 4 structure

Table 1. Likelihood Ratio Test (LRT) Statistics for Models of Variable Selective Pressure among Sites and among Branches.

LRT Model	Codon Frequency Model	Degree of Freedom	$-2\Delta\ln L$[a]	P Value[b]	Percentage of Sites with $\omega > 1$ (average dN/dS)
M1a versus M2a[c]	F3x4	2	67.83	1.86×10^{-15}	4.3% (2.79)
	F61	2	34.83	2.73×10^{-8}	3.4% (2.42)
M7 versus M8[d]	F3x4	2	86.04	2.08×10^{-19}	7.9% (1.98)
	F61	2	48.22	3.40×10^{-11}	10% (1.63)
M0 versus M1[e]	F3x4	56	110.56	1.91×10^{-5}	—
	F61	56	99.00	3.5×10^{-4}	—

[a]Twice the difference of the natural logs of the maximum likelihood of the models being compared.
[b]P value of rejecting the neutral models in favor of the positive selection models.
[c]M1a is a nearly neutral model that assumes one ω class between 0 and 1, and one class with $\omega = 1$; M2a (positive selection model) is the same as M1a plus an extra class of $\omega > 1$.
[d]M7 is a null model that assumes that $0 < \omega < 1$ is beta distributed among sites; M8 (positive selection model) is the same as M7 but also includes an extra category of sites with $\omega > 1$.
[e]M0 and M1 are free-ratio models which assume all branches to have the same ω (M0) or allow each branch to have its own ω (M1).

Table 2. Likelihood Ratio Test (LRT) Statistics for Models of Positive Selection on Specific Branches.

Foreground Branch[a]	Codon Frequency Model	Degree of Freedom	$-2\Delta\ln L$[b]	*P* Value[c]	Bonferroni Corrected *P* Value
Primates	F3x4	1	5.92	0.015	0.045
	F61	1	5.84	0.016	0.048
Alpaca	F3x4	1	31.52	1.97×10^{-8}	5.91×110^{-8}
	F61	1	30.90	2.7×110^{-8}	8.13×110^{-8}
Pig	F3x4	1	139.05	4.29×110^{-32}	1.29×110^{-31}
	F61	1	130.09	4.90×110^{-30}	1.20×110^{-27}

[a]MA and MA1 are branch-site models that assume four classes of sites: The MA model allows a proportion of codons to have $\omega \geq 1$ on the foreground branches (those to be tested for selection), whereas the MA1 model does not.
[b]$2\Delta\ln L$: Twice the difference of the natural logs of the maximum of the models being compared.
[c]P value of rejecting the neutral model in favor of the positive selection model.

(Supplementary Fig. S1, Supplementary Material online), which includes two antiparallel alpha-helices forming a bump protruding from the stalk domain. Positively selected sites in loop 4 form a continuous surface and are mainly located on unstructured loops (Figure 1). Analysis of non-loop 4 selected sites in MX2 and comparison with MX1 selection targets (Mitchell et al. 2012) indicated that the same region of the stalk domain carries the MX2 460N residue and three selected sites in MX1; also, the S518 residue is spatially close to these residues (Figure 1C). Interestingly, the MX2 primate-specific selection target in the GTPase domain (364S) is in spatial proximity to a site subject to diversifying selection in primate *MX1* genes.

Evolution of *MX1* and *MX2* in the human lineage

We next applied a recently developed population genetics-phylogenetics approach to study the evolution of *MX1* and *MX2* in the human lineage. Specifically, we ran the GammaMap program (Wilson et al. 2011) that jointly uses intra-specific variation and inter-specific diversity to estimate the distribution of fitness effects (DFE) (i.e. selection coefficients, γ) along coding regions. To this aim, we exploited data from the 1000 Genomes Pilot project deriving from the low-coverage whole genome sequencing of 179 individuals with different ancestry: Europeans (CEU), Yoruba from Nigeria (YRI), and Chinese plus Japanese (CHBJPT) (1000 Genomes Project Consortium et al. 2010). The ancestral sequence was reconstructed by parsimony from the human, chimpanzee, orangutan and macaque sequences. We first applied GammaMap to obtain the overall distribution of selection coefficients along *MX1* and *MX2*. A general preponderance of codons evolving under negative selection ($\gamma < 0$) was observed for both genes, with *MX1* showing stronger constraint (Figure 2). GammaMap allows

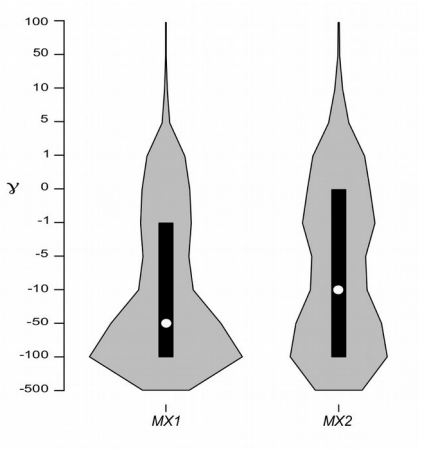


Figure 2. Lineage-specific selection and DFE analysis. Violin plot of selection coefficients (DFE) for *MX1* and *MX2* (median, white dot; interquartile range, black bar). Selection coefficients (γ) are classified as strongly beneficial (100, 50), moderately beneficial (10, 5), weakly beneficial (1), neutral (0), weakly deleterious (−1), moderately deleterious (−5, −10), strongly deleterious (−50, −100), and inviable (−500).

to identify specific codons evolving under positive selection.
We defined positively selected codons as those having a cumulative probability >0.80 of $\gamma \geq 1$. None was found in either *MX1* or *MX2*. Thus, coding variants in the two genes did not represent positive selection targets during human evolutionary history.

Population genetic analysis of *MX2* in humans

The results described above indicate that *MX2* has been a selection target throughout the evolutionary history of mammals. In the human

lineage, purifying selection had a major role in shaping *MX2* coding sequence diversity. Inspection of the 1000 Genomes Project data indicated that one single non-synonymous variant in *MX2* segregates in human populations at a frequency higher than 1% (rs56680307). This SNP is monomorphic or very rare in non-African populations, whereas it has a minor allele frequency (MAF) around 20% in Africa. We thus investigated whether natural selection also affected genetic diversity at *MX2* in human populations acting either on rs56680307 or on noncoding variants (as recent data have indicated that regulatory polymorphisms represent the bulk of selection targets in human populations (Forni et al. 2014; Grossman et al. 2013)). To this purpose, we initially exploited information from the Human Genome Diversity Panel (HGDP), which consists of ~650,000 single nucleotide polymorphisms (SNPs) genotyped in 52 populations distributed worldwide (Li et al. 2008). Using these data we calculated F_{ST} among continental groups.

F_{ST} is a measure of population genetic differentiation: under selective neutrality F_{ST} is mainly determined by demographic history and drift, but natural selection may drive allele frequencies to differ more than expected. Out of 22 HGDP variants in *MX2*, three, namely rs45430 (A/G), rs379839 (A/G), and rs2074560 (A/G), displayed an F_{ST} value higher than

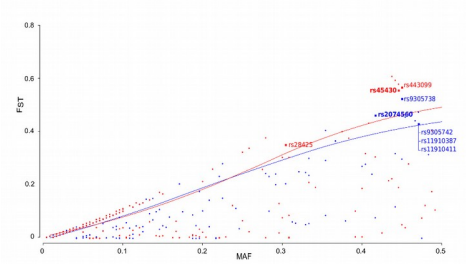


Figure 3. Population genetic differentiation. F_{ST} analysis of *MX2* variants for YRI-CEU (blue) and YRI-CHBJPT (red) comparisons. Lines represent the 95th percentiles in each MAF class. Solid squares denote variants identified in the HGDP SNP analysis (bold) or SNPs with a significant DIND test.

the 95th percentile calculated from the distribution of HGDP variants in the same minor allele frequency (MAF) class (see Materials and Methods) (Figure 3). Analysis of worldwide variation indicated that ancestral alleles for the three SNPs are almost fixed in African populations, whereas derived alleles reach high frequency outside Africa (Figure 4A). Thus, selective sweeps may have occurred during human colonization of Eurasia and the Americas.

To investigate this possibility further, we used the 1000 Genomes Pilot Project genotype data. Selected variants were searched for by integrating two approaches: one based on population genetic differentiation and the other on haplotype homozygosity. Thus, for all SNPs in *MX2* we calculated YRI-CEU and YRI-CHBJPT F_{ST}, as well as the Derived Intra-Allelic Nucleotide Diversity (DIND) test (Barreiro et al. 2009) in CEU and CHBJPT. The rationale behind the DIND test is that a derived allele under positive selection will rapidly increase in frequency in the population and will consequently display lower levels of nucleotide diversity at linked sites than expected. Thus, the ratio of intra-allelic diversity associated with the ancestral and derived alleles ($i\pi A/i\pi D$) is analyzed against the frequency of the derived allele (DAF); given a DAF interval, a high value of $i\pi A/i\pi D$ indicates that the neutral diversity associated with the derived allele is limited, suggesting positive selection. DIND has higher power than similar tests in most DAF ranges and is well suited for low-coverage data (Barreiro et al. 2009; Fagny et al. 2014). *T*he statistical significance of both tests was obtained by deriving empirical distributions of the same parameters calculated for a large number of randomly selected genes (see Materials and Methods for details).

In CEU 5 SNPs were outliers in the F_{ST} comparison (Figure 3) and had a significant DIND test (DAF: 0.75-0.85, DIND: 4.9-7.3). These variants include rs2074560 (derived allele: A) and, despite the high recombination rate in the region, define a homogeneous haplotype that carries derived alleles at several SNPs (Figure 4B). In CHBJPT, two variants had a significantly high F_{ST} value and a significant DIND test (rs28425 and rs443099, DAF: 0.57 and 0.82, DIND: 7.99 and 7.1, respectively) (Figure 3). The derived allele of rs443099 is carried by group of homogeneous haplotypes, a fraction of which also displays

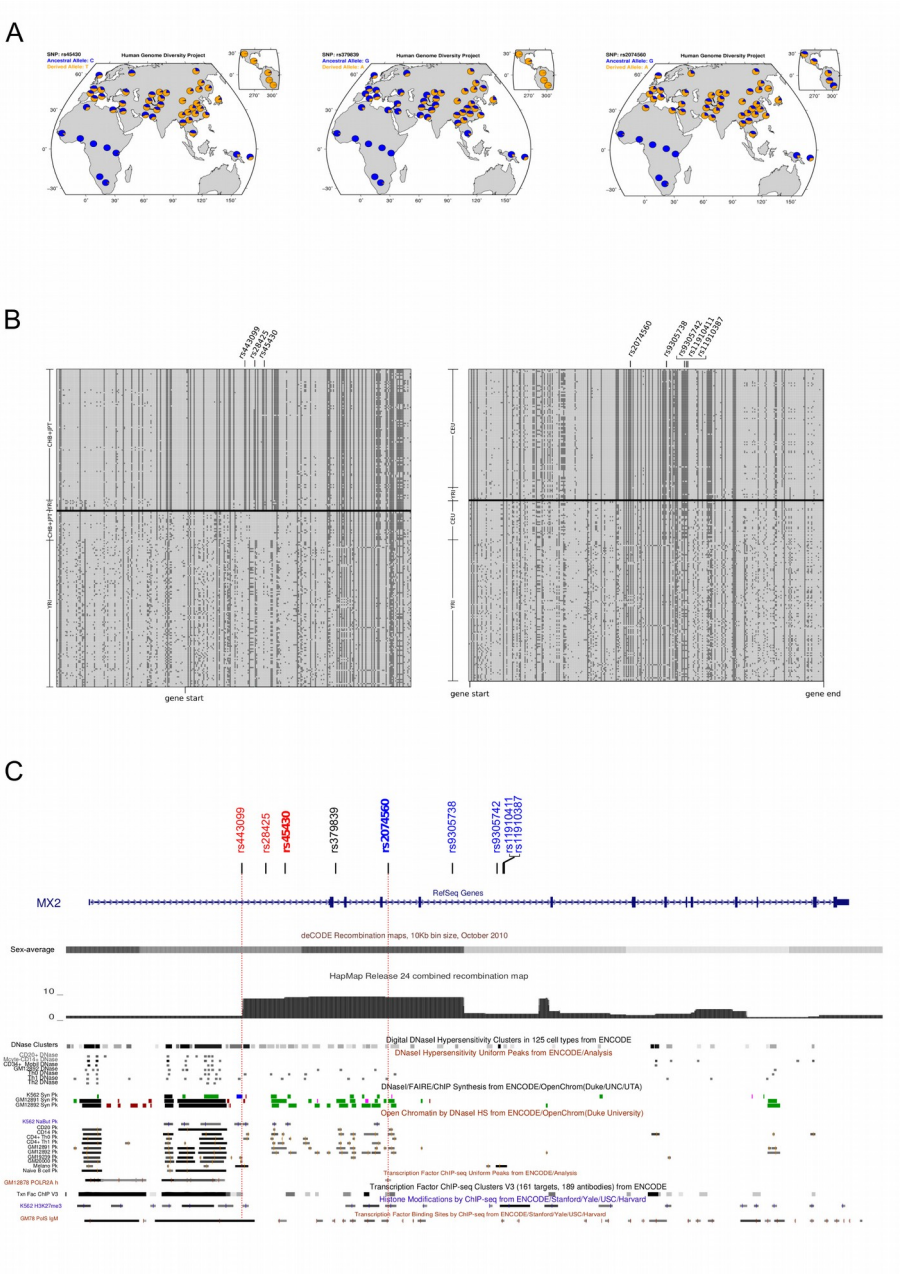


Figure 4. Natural selection at *MX2* in human populations. (A) Worldwide allele frequencies of the three HGDP variants showing outlier F_{ST} values. (B) Schematic representation of CHBJPT plus YRI (left) and CEU plus YRI (right) haplotypes in the regions surrounding rs443099 and rs2074560, respectively. Each line represents a haplotype, columns indicate polymorphic positions. Dark grey, derived alleles, light grey, ancestral alleles. The thick horizontal line separates haplotypes carrying the ancestral (bottom) and derived (top) allele at the selected variants (rs443099 and rs2074560 in the left and right panels, respectively). (C) Schematic representation of *MX2* within the UCSC Genome Browser view. The location of the selection targets, recombination rate (Kong et al. 2002), and relevant ENCODE tracks are also shown. SNP colour codes are as in Figure 3.

the derived allele for rs28425 (Figure 4B). The rs45430 SNP, which we identified as an outlier in the initial F_{ST} screen, has been associated with melanoma susceptibility in a genome-wide association study (GWAS); this variant is in tight LD (r^2=0.95, D'=1) with rs443099,
suggesting that the derived allele, associated with increased melanoma risk, has swept to high frequency outside Africa as a result of selection.
Overall, these results indicate that two distinct *MX2* haplotypes have been targeted by natural selection
in Asian and European populations.
Interestingly, ENCODE annotations (ENCODE Project Consortium et al. 2012) deriving from experiments in white blood cells or lymphoblastoid lineages indicated that the CEU and CHBJPT selected variants are located in a large region extending downstream the *MX2* transcription start site where several regulatory motifs are located, as assessed by analysis of DNase hypersensitive regions, transcription factor bindng sites and histone modifications (Figure 4C). C

***MX2* haplotype association with HIV-1 infection susceptibility**

Given the recently described role of MX2 as a restriction factor for HIV-1, we investigated whether the positively selected European haplotype modulates susceptibility to HIV-1 infection. Most human subjects are susceptible to this virus, but a minority of individuals do not seroconvert despite multiple exposures. We thus genotyped rs2074560 (G/A) in a cohort of 191 Spanish individuals who were exposed to the virus through injection drug use (IDU). All of them were HCV-positive, but 85 tested HIV-1 negative despite multiple exposures through needle sharing (HIV-1 exposed seronegative, IDU-HESN); the remaining subjects were HIV-1 positive (IDU-CTR). We observed a deviation from Hardy-Weinberg equilibrium (HWE) in IDU-HESN with an excess of homozygotes (HWE p value= 0.0047); IDU-CTR complied to HWE. A significant difference was observed in the genotypic distribution of rs2074560 in the two cohorts (Table 3), and the frequency of individuals homozygous for the G allele was significantly higher in IDU-HESN (20.0%) compared to IDU-CTR (7.5%). The odds ratio (OR) for a recessive model with the GG genotype being protective against HIV-1 infection was 3.06 (95% IC: 1.25-7.5, logistic regression, p= 0.014). In order to replicate this finding a second HESN population with different geographic origin and exposed to the virus through a different infection route was analyzed as well. Thus, rs2074560 was genotyped in a well characterized cohort of 88 heterosexual Italian subjects who have a history of unprotected sex with their seropositive partners (sex-exposed HESN, SexExp-HESN). As a control 188 Italian healthy controls (HC) were genotyped; both samples complied to HWE. Again, a significant difference was observed in the genotype distribution of rs2074560 (logistic regression, p= 0.014) (Table 3); similarly to what was observed in the Spanish sample, GG homozygotes were significantly over-represented in HESN (12.5%) compared to controls (3.4%). Thus, a recessive model yielded a significant association of the GG genotype with HIV-1 protection (logistic regression, p value= 0.005, OR: 4.33, 95% IC: 1.55-12.14). Further confirmation was sought in a third and smaller population of 37 SexExp-HESN from Spain. These subjects are women who exposed themselves repeatedly through unprotected intercourse with their HIV-1 infected partners. A sample of 180 Spanish HC was also analysed. The genotype proportions of rs2074560 complied to HWE in both samples. Again, GG homozygotes were much more abundant among SexExp-HESN (10.8%) than among HC (5.5%), thus supporting results obtained in the two other case-control cohorts (Table 3). As expected, the small sample size of the HESN population resulted in a non-significant p value due to lack of power. The results of the three association analyses were combined through a random effect meta-analysis, which revealed no heterogeneity among samples (Cochrane's Q p value= 0.66, I2= 0) and yielded a p value of 1.55x10-4 (Table 3). Overall, these results strongly suggest that the G allele of rs2074560 protects from HIV-1 with a recessive effect, irrespective of the infection route.

***MX2* haplotype association with in vitro viral infection**

Table 3. Association of rs2074560 with HIV-1 Infection Susceptibility.

Sample	Genotype Counts (GG/GA/AA)		$P_{genotypic}$ [a]	Genotype Counts (recessive)(GG/AG + AA)		$P_{recessive}$ [b]	OR (95% CI)[c]	MetaAnalysis $P_{recessive}$ and OR[d]
	HESN	CTR[e]		HESN	CTR			
IDU (Spain)	17/27/41	8/33/65	0.036	17/68	8/98	0.0143	3.06 (1.25–7.5)	
	HESN	HC		HESN	HC			1.55×10^{-4}
SexExp (Italy)	11/29/48	6/79/103	0.014	11/77	6/182	0.0052	4.33 (1.55–12.14)	OR: 3.12
SexExp (Spain)	4/15/18	10/78/92	0.5078	4/33	10/170	0.24	2.06 (0.61–6.97)	

[a]Logistic regression P value for a genotypic model.
[b]Logistic regression P value for a dominant model.
[c]Odds ratio for a recessive model with 95% CIs.
[d]Random-effect metaanalysis P value (recessive model) and OR.
[e]IDU HIV-1 positive (IDU-CTR in the text).

To verify whether rs2074560 affects HIV-1 replication in vitro, we performed an infection assays. Specifically, PBMCs from 50 Italian HESN subjects were cultured and infected with HIV-1_{Ba-L}. Viral replication was assessed after 3 days by measuring viral p24 levels produced by the infected cells. Results indicated that the G allele is associated with significantly lower p24 antigen levels (Kruskal-Wallis rank sum test, p= 0.034) (Figure 5).

Figure 5

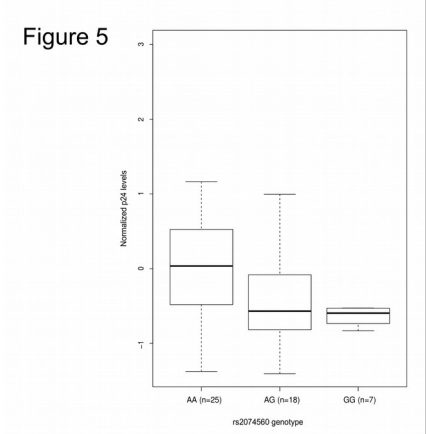


Figure 5. Box-and-whisker plot of p24 antigen. PBMCs from 50 Italian HESN subjects were infected with HIV-$1_{Ba-L;}$ the levels of viral p24 were measured after 3 days and normalized within experiments. Data are shown as a function of rs2074560 genotype in standard box-and-whisker plot representation (thick line: median; box: quartiles; whiskers: 1.5 x interquartile range).

MX2 expression in response to interferon treatment

Recent evidences have indicated that *MX2* expression is increased in response to IFN-α and that intra-individual variability may exist in the level of induction (Goujon et al. 2013; Kane et al. 2013). Thus, we assessed whether allelic status at rs2074560 influences *MX2* expression levels in response to IFN-α. To this aim, peripheral blood mononuclear cells (PBMC) from 45 healthy volunteers were stimulated with IFN-α and *MX2* mRNA induction was evaluated by real-time PCR. Results indicated that the G allele of rs2074560 is associated with a significantly increased *MX2* expression in response to IFN-α (Kruskal-Wallis rank sum test, p= 0.033) (Figure 6).

Discussion

The human MX2 protein has long been thought to lack antiviral activity and to serve cellular functions such as nucleo-cytoplasmic transport and cell-cycle progression (King, Raposo, Lemmon 2004). Recent evidences have challenged this view by showing that

Figure 6

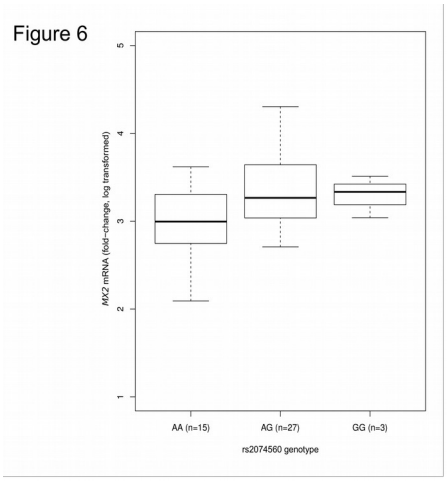


Figure 6. Box-and-whisker plot of *MX2* induction in response to IFN-α. Data derive from IFN-α stimulation of PBMCs deriving from 45 healthy volunteers. *MX2* transcript levels are shown as log-transformed fold-change expression from the unstimulated sample. Data are shown in standard box-and-whisker plot representation, as in Fig 4..

MX2 acts as a restriction factor against HIV-1 and other simian retroviruses (Goujon et al. 2013; Kane et al. 2013; Liu et al. 2013b). Herein we follow-up on these observations and demonstrate that *MX2* evolved adaptively in mammals and in human populations and that a polymorphism in the gene modulates both *MX2* expression in response to IFN-α and HIV-1 infection susceptibility.

Antiviral effectors are known to represent common targets of natural selection (Daugherty and Malik 2012), which most likely ensues from host-pathogen genetic conflicts. Thus, determinants of antiviral activity are often accounted for by rapidly evolving residues. In primate *MX1* genes most sites targeted by positive selection are located in loop 4 and explain species-specific difference in antiviral activity against orthomyxoviruses, although additional determinants of antiviral specificity must be accounted for by residues outside loop 4 (Mitchell et al. 2012). The natural selection pattern we observed at MX2 closely reflects the one described for MX1, with loop 4 strongly targeted by natural selection and the same helix surface in the stalk domain carrying three positively selected sites in MX1 and the positively selected 460 residue in MX2 (Mitchell et al. 2012). Interestingly, one of the positively selected sites we identified in MX2, S518 in spatial proximity to the 460 residue, corresponds to a polymorphic position in the pig coding sequence; aminoacid variation at this residue modulate the ability of the porcine MX2 protein to inhibit vescicular stomatitis virus replication (Sasaki et al. 2014), supporting the hypothesis that positively selected sites determine antiviral activity.

Application of two different methods indicated that lineage-specific selective events shaped diversity at primate *MX2* genes. Currently available tests have limited power to simultaneously detect the sites and the lineages subject to episodic positive selection (Murrell et al. 2012; Zhang, Nielsen, Yang 2005). Consequently, our analysis cannot be regarded as comprehensive in this respect, and additional sites evolving under episodic selection may have gone undetected. Indeed, we were unable to identify significant sites in the alpaca lineage, and more sites may represent selection targets in primates and pigs. Nonetheless, we consider thta the sites we did identify are supported by robust evidence, as they were validated by two distinct methods. Interestingly, one site evolving adaptively in primates is located in the MX2 GTPase domain, in spatial proximity to an MX1 site positively selected in primates (Mitchell et al. 2012). These observations strongly suggest that the GTPase domain might play a role as an antiviral determinant, although its precise function needs further experimental analysis, as MX2 mutants unable to bind or hydrolyse GTP retain their anti-HIV-1 activity (Goujon et al. 2013; Kane et al. 2013; Liu et al. 2013b). The role of loop 4 residues as antiviral determinants is more clearly established, as single aminoacid replacements alter the ability of MX1 to restrict Thogoto virus and influenza A virus (Mitchell et al. 2012). A crystallographic study indicated that MX1 forms ring-shaped oligomeric structures, but failed to solve the structure of loop 4, that was however predicted to face the internal surface of the ring (Gao et al. 2011). Our modeling data indicate that MX2 loop 4 protrudes from the stalk backbone and, were MX2 to adopt the same conformation as MX1 oligomers, might act as the foremost contact surface with the viral substrate (the capsid protein in the case of HIV-1 (Goujon et al. 2013; Kane et al. 2013; Liu et al. 2013b)).

Results herein also indicate a continuum in selective pressure acting in mammals and in human populations, and targeting *MX2*. Nonetheless, in line with a large-scale analysis of positive selection, which revealed most human selection targets to be located in noncoding regions (Grossman et al. 2013), positive selection acted on *MX2* SNPs with a likely regulatory role, whereas the coding sequence was mainly subject to selective constraint in the human lineage. Population genetic analyses indicate that two different haplotypes swept to high frequency in populations of non-African ancestry; in both cases the result is counter-intuitive as selection favoured the frequency increase of a melanoma susceptibility allele and of a polymorphism/haplotype associated with decreased IFN-α induction and augmented HIV-1 infection susceptibility. Previous analyses have indicated that risk alleles for human diseases may represent selection targets (Forni et al. 2013; Fumagalli et al. 2009;

Jostins et al. 2012; Raj et al. 2012; Zhernakova et al. 2010) and this possibly reflects either changes in environmental pressures or pleiotropic effects of the selected alleles. Thus, the variant we analysed may differently affect IFN-α response in cell types other than PBMCs. Indeed, rs2074560 is located in a region which likely modulates transcriptional activity in different cell types, as suggested by the presence of DNaseI hypersensitivity peaks and transcription factor binding sites. An alternative explanation is that increased levels of *MX2* induction in response to IFN-α is deleterious to some cellular process unrelated to antiviral response, an hypothesis previously proposed to explain the loss of APOBEC3H activity in populations of non-African ancestry (OhAinle et al. 2008). Conversely, our data indicate that the stronger response to IFN-α afforded by the rs2074560 ancestral allele confers protection in the context of HIV-1 infection. Indeed, we analysed three populations of HESN subjects with different geographic origin and distinct routes of exposure to the virus. Results were consistent in showing that homozygosity for the G allele at rs2074560 protects from infection. Because association results were replicated in the three independents cohorts, we consider that evidences are strong enough to establish a role for rs2074560 as a determinant of HIV-1 infection susceptibility in humans, despite the relatively small sample of the HESN populations. Moreover, the association results are supported by the *in vitro* infection assay, which indicated that the G allele significantly decreases viral replication. Recently, Lane and co-workers performed a genome-wide association study of resistance to HIV-1 infection in highly exposed uninfected hemophiliacs and detected no genome-wide significant association with the exclusion of the known *CCR5Δ32* allele (Lane et al. 2013); nonetheless, given the average minor allele frequency of rs2074560 in Europeans (around 0.25), its recessive effect, and the OR estimated herein (OR: 3.12), their study was underpowered to reach genome-wide significance for this variant.

A recent study of HIV-1-infected and seronegative subjects confirmed that *CCR5Δ32* homozygotes are protected from infection (OR=0.2), whereas no effect was detected for the *HLA-B*57:01* and *27:05 alleles, both associated with slow progression to AIDS (McLaren et al. 2013). The *CCR5Δ32* allele is almost exclusively found in Europe and West Asia, where its frequency follows a latitudinal cline (Novembre, Galvani, Slatkin 2005). In Italy and Spain, the estimated frequency of *CCR5Δ32* homozygotes is around 0.25-0.5% (Novembre, Galvani, Slatkin 2005), indicating that this allele makes a minor contribution to HIV-1 susceptibility at the population level. Additional polymorphisms that modulate the susceptibility to HIV-1 infection have been identified in recent years; all of them map to genes involved directly (e.g. *ERAP2*, *IRF1*, and *TLR3*) or indirectly (*FREM1* and *CYP7B1*) in immune response (Ball et al. 2007; Biasin et al. 2013; Limou et al. 2012; Luo et al. 2012; Sironi et al. 2012). Some of these loci showed a relatively strong effect, comparable to that estimated for the *MX2* variants we describe herein (Ball et al. 2007; Biasin et al. 2013; Limou et al. 2012; Luo et al. 2012). Nonetheless, most of these analyses, including those reported in our study, were performed on relatively small population samples, due to the intrinsic difficulty in characterizing and recruiting HESN cohorts. Thus, calculation of the effect size on lager samples will be necessary to provide more reliable estimations.

Another point that will require further investigation is the reason(s) why rs2074560, which modulates expression levels, confers protection with a recessive effect. One possibility is that the cellular amount of MX2 protein is rate-limiting for efficient HIV-1 restriction, so that a threshold needs to be reached to afford *in vivo* protection. This model might be appealing in light of the proposed model for MX1 antiviral activity, whereby oligomers enclose and sequester viral components (Gao et al. 2011; Reichelt et al. 2004). Another possibility is that, although in the association analysis infection susceptibility was treated as a dichotomous variable (i.e HESN/non-HESN status), it may be a continuous trait. Thus, the HESN status possibly represents an extreme phenotype and selection of HESN cohorts may result in enrichment for the most protective genotype. This hypothesis is in line with the results from the *in vitro* infection assay, whereby rs2074560 genotype was analysed against a continuous variable (p24 antigen levels); in this

experiment the effect of the G allele was also evident in heterozygous subjects, and the limited the number of GG homozygotes makes it difficult to determine whether an additive model applies to these data.

The observation that the *MX2* polymorphism that modulates susceptibility to HIV-1 infection also regulates MX2 expression in response to IFN-α is biologically interesting. *MX2* is part of the interferon-regulated genes, a vast and heterogeneous family of genes whose expression is modulated by IFN-α. Some of these genes, including *CH25H*, were recently shown to be endowed with potent antiviral properties (Liu et al. 2013a). It is thus tempting to envision a scenario in which an initial infection with HIV-1 of target cells that are GG homozygous would trigger the up-regulation of *MX2*; this in turn would block viral replication and impede nuclear accumulation and chromosomal integration of nascent viral DNA. The infection would be aborted and the immune response would be able to clear away the few cells that had been initially infected, preventing the infection to spread in an uncontrolled fashion. This could explain the puzzling presence of HIV-specific T helper cells and CTL in the absence of any detectable ongoing infection: a phenomenon that is observed in HESN not only during, but, transiently, also after cessation of exposure (Miyazawa et al. 2009).

In summary, building on the recently described role of MX2 as HIV-1 restriction factors, we exploited evolutionary information to identify a variant that confers natural resistance to HIV-1 infection. Results herein establish the role of MX2 as a central element of antiviral response in mammalian species and a possible target for therapeutic intervention in HIV-1 treatment and prevention.

Materials and Methods

Evolutionary analysis in mammals

Mammalian sequences for *MX2* were retrieved from the Ensembl and NCBI databases (http://www.ensembl.org/index.html; http://www.ncbi.nlm.nih.gov/). The list of species is reported in Supplementary Table S1, Supplementary Material online. DNA alignments were performed using the RevTrans 2.0 utility (Wernersson and Pedersen 2003), which uses the protein sequence alignment as a scaffold to construct the corresponding DNA multiple alignment. This latter was checked and edited by hand to remove alignment uncertainties. For PAML analyses (Yang 2007) we used trees generated by maximum-likelihood using the program PhyML (Guindon et al. 2009).

The site models implemented in PAML have been developed to detect positive selection affecting only a few aminoacid residues in a protein. To detect selection, site models that allow (M2a, M8) or disallow (M1a, M7) a class of sites to evolve with ω >1 were fitted to the data using the F3x4 and the F61 codon frequency model. Positively selected sites were identified using the Bayes Empirical Bayes (BEB) analysis (with a cut-off of 0.90), which calculates the posterior probability that each codon is from the site class of positive selection (under model M8) (Anisimova, Bielawski, Yang 2002). A second method, the Mixed Effects Model of Evolution (MEME) (with the default cutoff of 0.1) (Murrell et al. 2012) was applied to identify positively selected sites. MEME allows the distribution of ω to vary from site to site and from branch to branch at a site, therefore allowing the detection of both pervasive and episodic positive selection (Murrell et al. 2012).

To explore possible variations in selective pressure among different lineages, we applied the free-ratio models implemented in the PAML package: the M0 model assumes all branches to have the same ω, whereas M1 allows each branch to have its own ω (Yang and Nielsen 1998). The models are compared through likelihood-ratio tests (degree of freedom= total number of branches -1). In order to identify specific branches with a proportion of sites evolving with ω>1, we used BS-REL (Kosakovsky Pond et al. 2011) with the PhyML-generated tree. Branches identified using this approach were cross-validated with the branch-site likelihood ratio tests from PAML (the so-called modified model A and model MA1, “test 2”) (Zhang, Nielsen, Yang 2005). A false discovery rate correction was applied to account for multiple hypothesis testing (i.e. we corrected for the number of tested lineages), as suggested (Anisimova and Yang 2007). The advantage of this method is that it also implements a BEB analysis analogous to that described above to calculate

the posterior probabilities that each site belongs to the site class of positive selection on the foreground lineages. Thus, BEB allows identification of specific sites that evolve under positive selection on specific lineages, although it has limited statistical power (Zhang, Nielsen, Yang 2005). GARD (Kosakovsky Pond et al. 2006), MEME (Murrell et al. 2012), and BS-REL analyses were performed through the DataMonkey server (Delport et al. 2010) (http://www.datamonkey.org).

Structural model prediction and validation

Structural models of MX2 were initially predicted using 3 different methods: MODELLER (Eswar et al. 2006) with loop refinement, I-TASSER (Roy, Kucukural, Zhang 2010; Zhang 2008) with a defined template (MX1, PDB structure: 3SZR (Gao et al. 2011)) or I-TASSER without any template. The quality of each model was assessed with VADAR (Willard et al. 2003). The overall quality was estimated with respect to its geometry and energy (packaging defects, free energy of folding, core hydrophobic and charged residues) (Supplementary Tab. S2, Supplementary Material online). Secondary structures found in loops were validated through the use of PSIPRED (McGuffin, Bryson, Jones 2000) server (Supplementary Fig. S1, Supplementary Material online). According to these criteria the model produced by I-TASSER with MX1 as template was used for our analysis.

Population genetics-phylogenetics analysis

Data from the Pilot 1 phase of the 1000 Genomes Project were retrieved from the dedicated website (1000 Genomes Project Consortium et al. 2010). Low-coverage SNP genotypes were organized in a MySQL database. A set of programs was developed to retrieve genotypes from the database and to analyze them according to selected regions/populations. These programs were developed in C++ using the GeCo++ (Cereda et al. 2011), the libsequence (Thornton 2003), and the mysqlpp libraries. Coding sequence information was obtained for *MX1* and *MX2*. To analyze the DFE we used GammaMap (Wilson et al. 2011). We assumed θ (neutral mutation rate per site), k (transitions/transversions ratio), and T (branch length) to vary among genes following log-normal distributions. For each gene we set the neutral frequencies of non-STOP codons (1/61) and the probability that adjacent codons share the same selection coefficient (p=0.02). For selection coefficients we considered a uniform Dirichlet distribution with the same prior weight for each selection class. For each gene we run 10,000 iterations with thinning interval of 10 iterations.

Population genetic analysis in humans

HGDP genotype data derive from a previous work (Li et al. 2008). Atypical or duplicated samples and pairs of close relatives were removed (Rosenberg 2006). F_{ST} among continental groups was calculated for all HGDP SNPs. Because F_{ST} values are not independent from allele frequencies, variants were divided in 100 percentile classes based on MAF and F_{ST} distributions were calculated for each class; outliers were defined as variants with an F_{ST} higher than the 95th percentile in the distribution of SNPs in the same MAF class.

Data from the Pilot 1 phase of the 1000 Genomes Project were retrieved as described above. Genotype information was obtained for *MX2* and for 1,200 randomly selected RefSeq genes. Orthologous regions in the chimpanzee, orangutan or macaque genomes (outgroups) were retrieved using the liftOver tool (http://hgdownload.cse.ucsc.edu/); genes showing less than 80% human-outgroup aligning bases were discarded. This originated a final set of 987 genes (control set). F_{ST} (Wright 1950) and the DIND test (Barreiro et al. 2009) were calculated for all SNPs mapping to *MX2* and to the control gene sets. For F_{ST} variants were binned variants based on their MAF (100 classes) and the 95th percentile for each MAF class was calculated. As for the DIND test, it was originally developed for application to Sanger or high coverage sequencing data (Barreiro et al. 2009), so that statistical significance can be inferred through coalescent simulations. This is not the case for the 1000 Genomes Project data; thus, we calculated statistical significance by obtaining an empirical distribution of DIND-DAF value pairs for variants located within control genes. Specifically, DIND values were calculated for all SNPs using a constant number of 40 flanking variants (20 up- and down-stream). The distributions of DIND-DAF pairs for YRI, CEU and CHBJPT was binned in DAF intervals (100 classes, bin=0.01) and for each

class the 95th percentile was calculated. As suggested (Barreiro et al. 2009), for values of $i\pi_D = 0$ we set the DIND value to the maximum obtained over the whole dataset plus 20.

Subject cohorts, genotyping and statistical analysis

We recruited 191 males exposed to HIV-1 infection by injection drug use (IDU) and enrolled in prospective cohort studies in Spain (Valme Hospital, Sevilla) who had shared needles for >3 months. Concurrent markers of hepatitis C virus (HCV) infection were present in 100% of IDU subjects. Eighty-five of these subjects were HIV-1 negative (IDU-HESN), 106 were HIV-1 positive (IDU-CTR).

Thirty-eight Spanish HESN exposed to the virus through unprotected sexual intercourse (SexExp-HESN) were also recruited. These subjects are female partners of HIV-1 positive patients (without treatment and viremic, mean number of unprotected sexual intercourse per year: 110, mean number of years as sexual partners: 5, range 3-17). Healthy controls (HC, n=180) that were anonymous blood donors from The City of Jaen Hospital were also recruited. All these individuals were seronegative for both HIV-1 and HCV. All subjects were Spanish of Caucasian origin. The study was designed and performed according to the Helsinki declaration and was approved by the Ethics Committee of the participating hospitals and the University of Jaen. All patients and healthy blood donors provided written informed consent to participate in this study.

As for Italian SexExp-HESN, inclusion criteria were a history of multiple unprotected sexual episodes for more than 4 years at the time of the enrolment, with at least 3 episodes of at-risk intercourse within 4 months prior to study entry and an average of 30 (range, 18 to >100) reported unprotected sexual contacts per year (Miyazawa et al. 2009). Sex Exp-HESN and 188 healthy controls (HC) were recruited at the S. M. Annunziata Hospital, Florence; all of them were Italian of Caucasian origin. The study was reviewed and approved by the institutional review board of the S. M. Annunziata Hospital, Florence. Written informed consent was obtained from all subjects.

rs2074560 was genotyped through PCR amplification and direct sequencing (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), and inspected manually by two distinct operators. Genetic association analyses were performed by logistic regression and results from the three cohorts were combined using a random-effect meta-analysis; all analyses were performed using PLINK (Purcell et al. 2007).

HIV infection assay

PBMC from 50 HESN subjects were separated on lymphocyte separation medium (Organon Teknica, Malvern, Pa); 10×10^6 cells/mL were cultured for 2 days at 37°C and 5% CO_2 in RPMI 1640 containing FBS (20%), phytohemagglutinin (PHA) (7.5 μg/mL), and interleukin (IL)—2 (15 ng/mL). After viability assessment, 2.5×10^6 cells were resuspended in medium containing 1 ng of HIV-$1_{Ba\text{-}L}$ p24 viral input/10^6 PBMC and incubated for 3 h at 37°C. Cells were then washed and resuspended in 3 mL of complete medium with IL-2 (15 ng/mL). Cells were plated in 24-well tissue culture plates and incubated at 37°C and 5% CO_2. After 3 days supernatants were collected for p24 antigen ELISA analyses. Absolute levels of p24 were measured using the Alliance HIV-1 p24 ELISA Kit (PerkinElmer). PBMCs from the 50 subjects were split and infected in three independent experiments. To account for minor differences in virus titre, p24 levels were normalized within experiment. HIV-$1_{Ba\text{-}L}$ was provided through the EU programme EVA centre for AIDS Reagents NIBSC, UK.

IFN-α stimulation and *MX2* transcript quantification

Whole blood was collected from 45 healthy controls by venupuncture in Vacutainer tubes

containing EDTA (Becton Dickinson, NJ, USA), and PBMC were separated on lymphocyte separation medium (Organon Teknica, Malvern, Pa). Based on data derived from a kinetic study (data not shown), 5 x 10^5 freshly isolated PBMC were incubated for 3h with medium alone or 400 U/ml IFN-α (Sigma Aldrich).

RNA was extracted from cultured PBMC by using the acid guanidium thiocyanate–phenol–chloroform method. The RNA was dissolved in RNase-free water, and purified from genomic DNA with RNase-free DNase (RQ1 DNase, Promega, Madison, Wisconsin, USA). One microgram of RNA was reverse transcribed into first-strand cDNA in a 20-µl final volume containing 1 µM random hexanucleotide primers, 1 µM oligo dT and 200 U Moloney murine leukemia virus reverse transcriptase (Clontech, Palo Alto, California, USA). cDNA quantification for *MX2* and *GAPDH* was performed by a real-time PCR strategy (DNA Engine Opticon 2; MJ Research, Ramsey, USA). Reactions were performed using a SYBR Green PCR mix (5 prime, Gaithersburg, USA). Results were expressed as ΔΔCt and presented as ratios between the target gene and the *GAPDH* housekeeping mRNA.

Acknowledgments

D.F. and C.P. are supported by fellowships of the Doctorate School of Molecular Medicine, University of Milan.

Supplementary Information

Table S1. List of mammalian species.
Table S2. Model quality assessment.
Figure S1. Secondary structure prediction for loop 4.

References

1000 Genomes Project Consortium, Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061-1073.

Anisimova M, Yang Z. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol. Biol. Evol. 24:1219-1228.

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

Ball TB, Ji H, Kimani J, McLaren P, Marlin C, Hill AV, Plummer FA. 2007. Polymorphisms in IRF-1 associated with resistance to HIV-1 infection in highly exposed uninfected kenyan sex workers. Aids 21:1091-1101.

Barreiro LB, Ben-Ali M, Quach H, Laval G, Patin E, Pickrell JK, Bouchier C, Tichit M, Neyrolles O, Gicquel B et al. 2009. Evolutionary dynamics of human toll-like receptors and their different contributions to host defense. PLoS Genet. 5:e1000562.

Biasin M, Sironi M, Saulle I, de Luca M, la Rosa F, Cagliani R, Forni D, Agliardi C, lo Caputo S, Mazzotta F et al. 2013. Endoplasmic reticulum aminopeptidase 2 haplotypes play a role in modulating susceptibility to HIV infection. Aids 27:1697-1706.

Cereda M, Sironi M, Cavalleri M, Pozzoli U. 2011. GeCo++: A C++ library for genomic features computation and annotation in the presence of variants. Bioinformatics 27:1313-1315.

Daugherty MD, Malik HS. 2012. Rules of engagement: Molecular insights from host-virus arms races. Annu. Rev. Genet. 46:677-700.

Delport W, Poon AF, Frost SD, Kosakovsky Pond SL. 2010. Datamonkey 2010: A suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 26:2455-2457.

ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57-74.

Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY,

Pieper U, Sali A. 2006. Comparative protein structure modeling using modeller. Curr. Protoc. Bioinformatics Chapter 5:Unit 5.6.

Fagny M, Patin E, Enard D, Barreiro LB, Quintana-Murci L, Laval G. 2014. Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing datasets. Mol. Biol. Evol. .

Forni D, Cagliani R, Pozzoli U, Colleoni M, Riva S, Biasin M, Filippi G, De Gioia L, Gnudi F, Comi GP et al. 2013. A 175 million year history of T cell regulatory molecules reveals widespread selection, with adaptive evolution of disease alleles. Immunity 38:1129-1141.

Forni D, Cagliani R, Tresoldi C, Pozzoli U, De Gioia L, Filippi G, Riva S, Menozzi G, Colleoni M, Biasin M et al. 2014. An evolutionary analysis of antigen processing and presentation across different timescales reveals pervasive selection. PLoS Genet. 10:e1004189.

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

Gao S, von der Malsburg A, Dick A, Faelber K, Schroder GF, Haller O, Kochs G, Daumke O. 2011. Structure of myxovirus resistance protein a reveals intra- and intermolecular domain interactions required for the antiviral function. Immunity 35:514-525.

Goujon C, Moncorge O, Bauby H, Doyle T, Ward CC, Schaller T, Hue S, Barclay WS, Schulz R, Malim MH. 2013. Human MX2 is an interferon-induced post-entry inhibitor of HIV-1 infection. Nature .

Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, Griesemer D, Karlsson EK, Wong SH et al. 2013. Identifying recent adaptations in large-scale genomic data. Cell 152:703-713.

Guindon S, Delsuc F, Dufayard JF, Gascuel O. 2009. Estimating maximum likelihood phylogenies with PhyML. Methods Mol. Biol. 537:113-137.

Haller O, Kochs G. 2011. Human MxA protein: An interferon-induced dynamin-like GTPase with broad antiviral activity. J. Interferon Cytokine Res. 31:79-87.

Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, Lee JC, Schumm LP, Sharma Y, Anderson CA et al. 2012. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491:119-124.

Kane M, Yadav SS, Bitzegeio J, Kutluay SB, Zang T, Wilson SJ, Schoggins JW, Rice CM, Yamashita M, Hatziioannou T et al. 2013. MX2 is an interferon-induced inhibitor of HIV-1 infection.Nature

King MC, Raposo G, Lemmon MA. 2004. Inhibition of nuclear import and cell-cycle progression by mutated forms of the dynamin-like GTPase MxB. Proc. Natl. Acad. Sci. U. S. A. 101:8957-8962.

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.

Kosakovsky Pond SL, Frost SD. 2005. Not so different after all: A comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. 22:1208-1222.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. 2006. Automated phylogenetic detection of recombination using a genetic algorithm. Mol. Biol. Evol. 23:1891-1901.

Kosakovsky Pond SL, Murrell B, Fourment M, Frost SD, Delport W, Scheffler K. 2011. A random effects branch-site model for detecting episodic diversifying selection. Mol. Biol. Evol. 28:3033-3043.

Lane J, McLaren PJ, Dorrell L, Shianna KV, Stemke A, Pelak K, Moore S, Oldenburg J, Alvarez-Roman MT, Angelillo-Scherrer A et al. 2013. A genome-wide association study of resistance to HIV infection in highly exposed uninfected individuals with hemophilia A. Hum. Mol. Genet. 22:1903-1910.

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.

Limou S, Delaneau O, van Manen D, An P, Sezgin E, Le Clerc S, Coulonges C, Troyer JL, Veldink JH, van den Berg LH et al. 2012. Multicohort genomewide association study reveals a new signal of protection against HIV-1 acquisition. J. Infect. Dis. 205:1155-1162.

Liu SY, Sanchez DJ, Aliyari R, Lu S, Cheng G. 2012. Systematic identification of type I and type II interferon-induced antiviral factors. Proc. Natl. Acad. Sci. U. S. A. 109:4239-4244.

Liu SY, Aliyari R, Chikere K, Li G, Marsden MD, Smith JK, Pernet O, Guo H, Nusbaum R, Zack JA et al. 2013a. Interferon-inducible cholesterol-25-hydroxylase broadly inhibits viral entry by production of 25-hydroxycholesterol. Immunity 38:92-105.

Liu Z, Pan Q, Ding S, Qian J, Xu F, Zhou J, Cen S, Guo F, Liang C. 2013b. The interferon-inducible MxB protein inhibits HIV-1 infection. Cell. Host Microbe .

Luo M, Sainsbury J, Tuff J, Lacap PA, Yuan XY, Hirbod T, Kimani J, Wachihi C, Ramdahin S, Bielawny T et al. 2012. A genetic polymorphism of FREM1 is associated with resistance against HIV infection in the pumwani sex worker cohort. J. Virol. 86:11899-11905.

McGuffin LJ, Bryson K, Jones DT. 2000. The PSIPRED protein structure prediction server. Bioinformatics 16:404-405.

McLaren PJ, Coulonges C, Ripke S, van den Berg L, Buchbinder S, Carrington M, Cossarizza A, Dalmau J, Deeks SG, Delaneau O et al. 2013. Association study of common genetic variants and HIV-1 acquisition in 6,300 infected cases and 7,200 controls. PLoS Pathog. 9:e1003515.
Mitchell PS, Patzina C, Emerman M, Haller O, Malik HS, Kochs G. 2012. Evolution-guided identification of antiviral specificity determinants in the broadly acting interferon-induced innate immunity factor MxA. Cell. Host Microbe 12:598-604.

Miyazawa M, Lopalco L, Mazzotta F, Lo Caputo S, Veas F, Clerici M, ESN Study Group. 2009. The 'immunologic advantage' of HIV-exposed seronegative individuals. Aids 23:161-175.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 8:e1002764.

Novembre J, Galvani AP, Slatkin M. 2005. The geographic spread of the CCR5 Delta32 HIV-resistance allele. PLoS Biol. 3:e339.

OhAinle M, Kerns JA, Li MM, Malik HS, Emerman M. 2008. Antiretroelement activity of APOBEC3H was lost twice in recent human evolution. Cell. Host Microbe 4:249-259.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559-575.

Raj T, Shulman JM, Keenan BT, Chibnik LB, Evans DA, Bennett DA, Stranger BE, De Jager PL. 2012. Alzheimer disease susceptibility loci: Evidence for a protein network under natural selection. Am. J. Hum. Genet. 90:720-726.

Reichelt M, Stertz S, Krijnse-Locker J, Haller O, Kochs G. 2004. Missorting of LaCrosse virus nucleocapsid protein by the interferon-induced MxA GTPase involves smooth ER membranes. Traffic 5:772-784.

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.

Roy A, Kucukural A, Zhang Y. 2010. I-TASSER: A unified platform for automated protein structure and function prediction. Nat. Protoc. 5:725-738.

Sasaki K, Tungtrakoolsub P, Morozumi T, Uenishi H, Kawahara M, Watanabe T. 2014. A single nucleotide polymorphism of porcine MX2 gene provides antiviral activity against vesicular stomatitis virus. Immunogenetics 66:25-32.

Sironi M, Biasin M, Cagliani R, Forni D, De Luca M, Saulle I, Lo Caputo S, Mazzotta F, Macias J, Pineda JA et al. 2012. A common polymorphism in TLR3 confers natural resistance to HIV-1 infection. J. Immunol. 188:818-823.

Thornton K. 2003. Libsequence: A C++ class library for evolutionary genetic analysis. Bioinformatics 19:2325-2327.

Verhelst J, Hulpiau P, Saelens X. 2013. Mx proteins: Antiviral gatekeepers that restrain the uninvited. Microbiol. Mol. Biol. Rev. 77:551-566.

Wernersson R, Pedersen AG. 2003. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res. 31:3537-3539.

Willard L, Ranjan A, Zhang H, Monzavi H, Boyko RF, Sykes BD, Wishart DS. 2003. VADAR: A web server for quantitative evaluation of protein structure quality. Nucleic Acids Res. 31:3316-3319.

Wilson DJ, Hernandez RD, Andolfatto P, Przeworski M. 2011. A population genetics-phylogenetics approach to inferring natural selection in coding sequences. PLoS Genet. 7:e1002395.

Wright S. 1950. Genetical structure of populations. Nature 166:247-249.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586-1591.

Yang Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555-556.

Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. 46:409-418.

Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22:1107-1118.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22:2472-2479.

Zhang Y. 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9:40-2105-9-40.

Zhernakova A, Elbers CC, Ferwerda B, Romanos J, Trynka G, Dubois PC, de Kovel CG, Franke L, Oosting M, Barisani D et al. 2010. Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. Am. J. Hum. Genet. 86:970-977.

3.2.1.4 The complement system as an epitome of host-pathogen genetic conflicts

Because the complement system acts as a first line defense and integrates innate and adaptive immune responses, an extremely wide range of pathogens have developed complement evasion strategies. Indeed, pathogenic organisms as diverse as helminths and viruses encode molecules that inhibit or antagonize one or more steps of the activation cascade [152-154]. In this respect, the complement system is an epitome for the constant host-pathogen conflict that contrasts immune response with immune evasion [12].

Genetic conflicts have previously been described mainly in the context of host/virus interactions [155, 156], and few studies have integrated evolutionary analysis of both host and pathogen interacting partners. Moreover, the complement system is targeted by pathogens that naturally infect humans but not other primates (e.g. *Neisseria meningitidis* and *Neisseria gonorrhoeae*), offering the opportunity to dissect the evolutionary events underlying host specificity. Finally, microbial interactors of complement system components are often used for vaccine design.

Thus, the aims of this work were: i) to provide a comprehensive catalog of sites in complement system genes that were targeted by positive selection in primates and in the human lineage; ii) to test whether positive selection acted on bacterial interactors; iii) to integrate evolutionary analysis and docking studies with previous biochemical *in vitro* data to test general hypotheses on host-pathogen interactions.

I analyzed the evolutionary history of 40 complement genes in primates and in the human lineage using conservative likelihood ratio tests and a population genetics-phylogenetics approach [19]. I also applied omegaMap [80], a Bayesian method that simultaneously estimates recombination rate

and selection, to study 5 bacterial interactors of complement components. I finally integrated evolutionary analyses with literature data and performed protein-protein docking to evaluate the effect of positively selected sites.

Results show that host components targeted by several microbial/viral proteins evolved under the strongest selective pressure in primates; these include CFH, CD59, C4BPA, and CD55. At such genes, as well as at others (e.g. *CR2* and *VTN*), selection acted on residues that are located at the interaction interface with microbial/viral components. In CFH, for instance, the C-terminal CCP domains (19-20), which are bound by an array of pathogens, represented preferential selection targets (uniform random sampling, $p<10^{-5}$). In particular, three of the sites detected in these domains were previously shown to be bound by different pathogens belonging to distinct phyla [157], indicating very strong and long-standing selective pressures. One of these sites also represents a human-specific determinant for the binding of *N. gonorrhoeae*, as is the case for three positively selected sites in C4BPA [158]. Partially because of the resistance to human complement-mediated killing, natural infection with *N. gonorrhoeae* is restricted to our species [159]. Examples of selection-driven species-specific susceptibility to infection had previously been reported for viral pathogens [155, 156], but rarely for bacteria [160].

As for bacteria, I analyzed five genes encoding complement-interacting partners: two porins from *N. gonorrhoeae*, *PspC* (*Steptococcus pneumoniae)*, *OspE* (*Borrelia spp.*), and *CspZ* (*Borrelia spp.*). They were selected because the molecular determinants of the interaction with CFH or C4BP, the two most common targets, are known. All five genes were found to have evolved adaptively and in all instances positively selected sites were found to be involved in the binding of the host interactor. Theseresults reflect the expectations under a genetic conflict scenario whereby the

host's and the pathogen's genes evolve within binding avoidance-binding seeking dynamics. Protein-protein docking analyses supported this view. In fact, I exploited previous data to validate an *in silico* mutagenesis and docking strategy that was subsequently used to analyze the effect of selected sites. To this aim, a subset of host-pathogen protein-protein interactions was selected for the *in silico* docking studies based on the availability of the 3D structure of the complex. These analyses revealed that all positively selected sites we analyzed (both in the host and in the pathogen interacting partner) modulate binding.

The pathogens included in this study are extremely important from a medical perspective and account for substantial morbidity and mortality. Bacterial proteins that interact with complement components have been regarded as attractive vaccine candidates, because vaccine efficacy may be enhanced by virulence impairment [161]. The possibility that vaccines and, more generally, antimicrobial compounds, contribute to pathogen evolution ultimately resulting in cultural-based arms race scenarios has previously been envisaged [162, 163]. These data support this view: complement-interacting protein surfaces may be more than poised for the next round of drug-driven selection.

Personal contribution to the work: I performed the evolutionary analysis in primates and I produced figures and tables.

The mammalian complement system as an epitome of host-pathogen genetic conflicts

Rachele Cagliani[1*], Diego Forni[1], Giulia Filippi[2], Alessandra Mozzi[1], Luca De Gioia[2], Chiara Pontremoli[1], Uberto Pozzoli[1], Nereo Bresolin[1,3], Mario Clerici[4,5], Manuela Sironi[1]

[1] Bioinformatics, Scientific Institute IRCCS E. MEDEA, 23842 Bosisio Parini, Italy.
[2] Department of Biotechnology and Biosciences, University of Milan-Bicocca, 20126 Milan, Italy.
[3] Dino Ferrari Centre, Department of Physiopathology and Transplantation, University of Milan, Fondazione Ca' Granda IRCCS Ospedale Maggiore Policlinico, 20122 Milan, Italy.
[4] Department of Physiopathology and Transplantation, University of Milan, 20090 Milan, Italy.
[5] Don C. Gnocchi Foundation ONLUS, IRCCS, 20148 Milan, Italy.

Corresponding author: Rachele Cagliani, PhD, Bioinformatics - Scientific Institute IRCCS E.MEDEA, 23842 Bosisio Parini, Italy. Tel: +39-031877826; Fax:+39-031877499; e-mail: rachele.cagliani@bp.lnf.it.

Abstract

The complement system is an innate immunity effector mechanism; its action is antagonized by a wide array of pathogens and complement evasion determines the virulence of several infections. We investigated the evolutionary history of the complement system and of bacterial-encoded complement-interacting proteins. Complement components targeted by several pathogens evolved under strong selective pressure in primates, with selection acting on residues at the contact interface with microbial/viral proteins. Positively selected sites in CFH and C4BPA account for the human specificity of gonococcal infection. Bacterial interactors, evolved adaptively as well, with selected sites located at interaction surfaces with primate complement proteins. These results epitomize the expectation under a genetic conflict scenario whereby the host's and the pathogen's genes evolve within binding avoidance-binding seeking dynamics. *In silico* mutagenesis and protein-protein docking analyses supported this by showing that positively selected sites, both in the host's and in the pathogen's interacting partner, modulate binding.

Keywords: Complement system; host-pathogen genetic conflict; positive selection; human specific infections.

Introduction

The complement system is an integral arm of innate immunity that plays essential roles in the clearance of pathogenic invaders, in the maintenance of tissue integrity, and in the elicitation of inflammatory reactions. This system consists of several circulating and cell-surface-bound proteins that orchestrate a series of proteolytic cascades (Ricklin & Lambris 2013). Different triggers result in distinct pathways of complement activation. The classical pathway is activated by IgM- and IgG- including immunocomplexes, but also by endogenous molecules such as C reactive protein (CRP) or by direct binding of C1q to pathogen-associated molecular patterns (Fig. 1). The alternative pathway does not require specific activation, but is initiated by a spontaneous hydrolysis of unstable C3 component (Fig. 1). Finally, the lectin pathway is triggered by recognition of specific carbohydrate structures on the surface of foreign cells through ficolins and mannose-binding lectin (Fig. 1) (Ricklin & Lambris 2013). The three pathways converge on C3 and the first effect of the complement

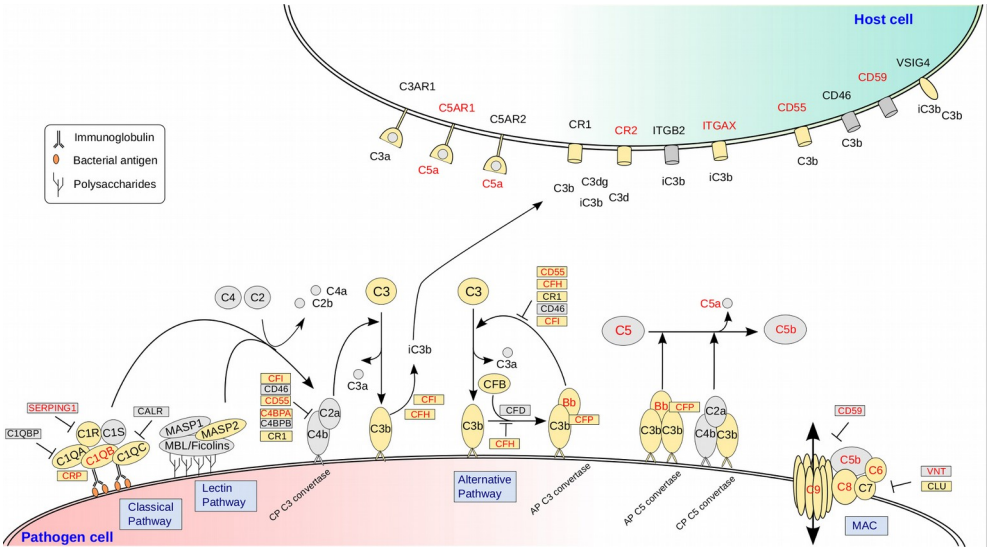


Figure 1. Positive selection at complement system genes The molecular components of the complement system are shown to provide a general overview of the extent of positive selection and to highlight the function of positively selected genes. Thus, the figure is not meant to show all molecules involved in the process or to convey mechanistic insights. Gene products are color-coded as follows: yellow, positive selection in the human lineage (at least one positively selected site in gammaMap analysis); red symbol, positive selection in primates.

cascade activation is the production of an active C3 convertase (Fig. 1). The alternative and the classical/lectin pathways form distinct C3 convertases which cleave C3 into the small C3a anaphylatoxinand C3b. Anaphylotoxins act as potent pro-inflammatory molecules at the systemic and local level and C3b deposited onto cell surfaces acts as an opsonin. Opsonization induces cell clearance by phagocytosis and triggers the assembly of C5 convertases. In turn, C5 cleavage generates C5a, another anaphylatoxin, and the C5b fragment. The final effector mechanism is mediated by the sequential addition of soluble components C6, C7, C8 and C9 to C5b and leads to the formation of the membrane attack complex (MAC), a lytic pore that can cause cell lysis (Ricklin & Lambris 2013) (Fig. 1). Deregulation of the complement system can cause tissue damage and inflammation resulting in the development of inflammatory and autoimmune disorders (Ricklin & Lambris 2013). Thus, the system is tightly regulated at different levels: complement regulators act to decrease the activity of the C3 convertase (i.e. CFH and C4BP), to modulate the activity of the C1 complex (SERPING1), and/or to inhibit MAC assembly (CD59, vitronectin and clusterin) (Ballanti *et al.* 2013) (Fig. 1).

An extremely wide range of pathogens have developed complement evasion strategies. Indeed,pathogenic organisms as diverse as helminths and viruses encode molecules that inhibit or antagonize one or more steps of the activation cascade (Lambris *et al.* 2008; Pangburn *et al.* 2008; Serruto *et al.* 2010). The most commonly used strategies include i) the recruitment of complement regulators (especially CFH, C4BP, CD55, and CD59); ii) the use of proteases to degrade complement components; iii) the direct inhibition of the complement cascade; and iv) the secretion of proteins that mimic complement regulators. On one hand, the same pathogen may be endowed with an arsenal of complement-targeting molecules; on the other hand, the same microbial-encoded protein may target multiple complement components (Lambris *et al.* 2008; Pangburn *et al.* 2008; Serruto *et al.* 2010). In this respect, the complement system is an epitome for the constant host-pathogen conflict that confronts immune response with immune evasion.

Because of these considerations, complement components and microbial-encoded antagonists

are expected to engage in a constant molecular "arms race" and to be targeted by natural selection (Red Queen hypothesis). Selective pressure is expected to be strongest at protein-protein interaction surfaces and to result in positive selection at codons that encode such surfaces (Sironi *et al.* 2015). Herein we set out to verify these predictions.

Genetic conflicts have previously been described mainly in the context of host/virus interactions (Sironi *et al.* 2015; Sawyer & Elde 2012; Daugherty & Malik 2012), and few studies have integrated evolutionary analysis of both host and pathogen interacting partners (Demogines *et al.* 2013; Barber & Elde 2014). Moreover, the complement system is targeted by pathogens that naturally infect humans but not other primates (e.g. *Neisseria meningitidis* and *Neisseria gonorrhoeae*), offering, as a result, the opportunity to dissect the evolutionary events underlying host-specificity. Thus, we performed an evolutionary analysis of the complement system and of some microbial interactors. We report extremely widespread positive selection in primates and we show that most positively selected sites are located at interaction surfaces. This was also true for pathogen-encoded proteins, with docking analysis supporting role of selected sites in the modulation of binding affinities.

Methods

Evolutionary analysis in primates

In this study we considered 44 genes encoding complement proteins on the basis of the Complement Map Database (CMAP) network (http://www.complement.us/cmap, "Main Map-Direct", "Complement Proteins"). Among the CMAP "Endogenous Molecules", we selected four more genes (*CD59*, *CLU*, *CRP*, and *VNT*) involved in the regulation of the system. Specifically, we focused on 4 regulators that directly interact with complement proteins as previously reported (Ricklin *et al.* 2010).

Primate coding sequences were retrieved from the Ensembl and NCBI databases (http://www.ensembl.org/index.html; http://www.ncbi.nlm.nih.gov/).

For all human genes, we checked whether primate genes represented 1-to-1 orthologs using the EnsemblCompara GeneTrees database (Vilella *et al.* 2009). This information was not available for *Pan paniscus*, *Macaca fascicularis*, and *Saimiri boliviensis;* a BLAST search of the gene coding sequences against the genome of these three species was performed using the NCBI BLAST utility. We removed single primate genes for which BLAST or EnsemblCompara hits were not consistent with the presence of a single ortholog. Eight genes (*C4A*, *C4B*, *CFHR1*, *CFHR4*, *CR1*, *CFB*, *FCN1*, and *FCN2*) were not included in the study due to extensive copy number variation in humans and to the impossibility of reliably identifying orthologs in primates. The list of genes, species, and the number of primate sequences analyzed for each gene are reported in Table S1 and Table S2, Supporting Information, respectively.

DNA alignments were performed with the RevTrans 2.0 utility (MAFFT v6.240 aligner) (Wernersson & Pedersen 2003) using the protein sequence alignment as a scaffold to construct the corresponding DNA multiple alignment. Alignments were checked and manual editing was used to correct few misalignments in proximity of small gaps. We next used GUIDANCE (Penn *et al.* 2010) to calculate confidence scores for all columns in the alignments. Scores were always higher that 0.90, indicating high reliability (Privman *et al.* 2012).

All alignments were screened for the presence of recombination breakpoints using GARD (Genetic Algorithm Recombination Detection) (Kosakovsky Pond *et al.* 2006).

To detect positive selection, we used the site models implemented in PAML (Yang 1997; Yang 2007) for whole gene alignments or independently for sub-regions defined in accordance with the recombination breakpoints. Specifically, we fitted site models that allow (M2a, M8) or disallow (M1a, M7) a class of sites to evolve with dN/dS (ω) >1 to the data using the F3x4 and the F61 codon frequency models. Input trees were generated by maximum-likelihood using the program PhyML (GTR+G as nucleotide substitution model) (Guindon *et al.* 2009).

Sites were called as positively selected if they were detected using both BEB (from model M8

with a cutoff of 0.90) (Anisimova *et al.* 2002) and MEME (with a cutoff of 0.1) (Murrell *et al.* 2012). We note that, whereas the PAML approach implicitly assumes that the strength and direction of natural selection is uniform across all lineages, MEME allows the distribution of dN/dS to vary from site to site and from branch to branch, allowing the identification of sites targeted by episodic positive selection. Thus, as expected, MEME generally detects more sites than BEB (Table S3, Supporting Information). Because BEB and MEME have a low but non-negligible false positive rate (Anisimova *et al.* 2002; Murrell *et al.* 2012), we applied the most conservative approach and declared selection at a site only if it was detected by both methods.

GARD (Kosakovsky Pond *et al.* 2006), MEME (Murrell *et al.* 2012), and SLAC (Kosakovsky Pond & Frost 2005) analyses were performed either through the DataMonkey server (Delport *et al.* 2010) (http://www.datamonkey.org) or run locally through HyPhy (Pond *et al.* 2005).

For the population genetics-phylogenetics analysis, genotype data from the Pilot 1 phase of the 1000 Genomes Project were retrieved from the dedicated website (1000 Genomes Project Consortium *et al.* 2010); in particular, SNP information were retrieved for individuals of three human populations: African (Yoruba), European, and East Asian (Chinese plus Japanese). Ancestral sequences were reconstructed by parsimony from the human, chimpanzee, orangutan and macaque sequences.

Analyses were performed with GammaMap (Wilson *et al.* 2011), that uses intra-specific variation and inter-specific diversity to estimate the distribution of population-scaled selection coefficients (γ) along coding regions. gammaMap classifies γ values into 12 categories, ranging from strongly beneficial (γ=100) to inviable (γ =-500), with γ equal to 0 indicating neutrality. In the analysis, we assumed θ (neutral mutation rate per site), k (transitions/transversions ratio), and T (branch length) to vary among genes following log-normal distributions. For p (the probability that adjacent codons share the same population-scaled selection coefficient) we assumed a uniform distribution. For each gene we set the neutral frequencies of non-STOP codons (pi) to 1/61. GammaMap analyses were also performed with pi estimated from the data for each gene and results were comparable with the ones obtained with pi = 1/61. For population-scaled selection coefficients we considered a uniform Dirichlet distribution with the same prior weight for each selection class. For each gene, two Markov Chain Monte Carlo runs of 10,000 iterations each were run with a thinning interval of 10 iterations. Runs were compared for convergence and merged for the analyses.

Uniform sampling was performed with 100,000 random samplings by assuming an equiprobable distribution of selected sites along protein sequences. Analyses were performed for domains that interact with pathogen-encoded molecules.

Evolutionary analysis of bacterial-encoded interactors

Sequences for *N. gonorrhoeae Por1A* and *Por1B* were obtained from a previous work (Posada *et al.* 2000).

PspC sequences were retrieved from Iannelli et al. (Iannelli *et al.* 2002); we included “typical” sequences belonging to allelic groups 1-3, 5, and 6 (Iannelli *et al.* 2002). *OspE* sequences were obtained from the NCBI database, with the exclusion of antigenic variants arising during infection (Table S4, Supporting Information). Finally, *CspZ* sequences derive from human clinical isolates as reported in Rogers et al. (Rogers *et al.* 2009).

Analysis of positive selection was performed using omegaMap (Wilson & McVean 2006). The program performs Bayesian inferences of ω and ρ (recombination parameter), allowing both parameters to vary along the sequence. An average block length of 10 and 30 codons was used to estimate ω and ρ, respectively. To determine the influence of the choice of priors on the posteriors, the analyses were repeated with two alternative sets of priors (Table S5, Supporting Information). For each alignment, three independent omegaMap runs, each with 500,000 iterations and a 50,000 iteration burn-in, were compared to assess convergence and merged to obtain the posterior probabilities.

Protein 3D structures, *in silico* mutagenesis and protein-protein docking

The structure of human CFH, C4BPA, CR2,

Table 1 Evolutionary analysis of primate complement genes

Gene*	Positively selected sites (human codons)†	% of sites (average dN/dS)‡	Tree length§
C1QB	A191	5.00 (3.402)	1.48
C4BPA	H2, R15, A19, R70, S88, H89, Y110¶, F145, R274, T317, S440, S442, Q463, H472¶	15.03 (3.060)	1.99
C5	S693, V705, I725, R1412, A1524	13.29 (1.840)	0.86
C5AR1	L187, E269	5.10 (3.881)	2.17
C6	S280, T765	8.79 (2.051)	1.02
C8A	V91, S523	6.56 (2.376)	1.30
C8B	S174, K251, P261	12.72 (2.116)	1.15
C9	A38, E222, A238, T248, C254, E393	20.28 (1.921)	1.35
CD55	L38, E57, P112, V124, P141¶, R170, P177, G178¶, S187, Q230, H263, T352	22.89 (2.775)	2.80
CD59	N33, V42, D47, D74, R78, R80, Y87, K91, L100	35,82 (3.829)	2.09
CFH	I49, N269, S354, H360, H402, A415, W787, H821, T1184, S1196, V1200, R1203, R1206	4.66 (4.799)	2.03
CFI	L4, S172, K405, R406, V408, Y411, Q485	6.74 (2.982)	1.61
CFP	T60, F62, R79, V189, P213, P265, N285¶, P377, E468	10.73 (2.119)	1.54
CR2	Y36, G53, T54, D72, N136, M137, G171, R840	9.75 (2.552)	1.19
CRP	L44, S92, G166¶, L194	17.65 (2.162)	1.60
ITGAX reg1	Q223, R238, H241, Y244	9.70 (3.259)	1.50
SERPING1	I50, V56, V62	10.16 (2.370)	1.57
VTN	T76, Y78, E91, S100, T113, A129, T325	15.51 (1.887)	1.52

*Only genes subject to positive selection (see text) are shown.
†Positively selected sites identified by both BEB and MEME.
‡Estimated percentage of sites evolving under positive selection by M8.
§Tree length is defined as the number of nucleotide substitutions per codon.
¶Positively selected sites also detected by gammaMap analysis; see also Tables S3 and S7 (Supporting information).

CD55, CRP , ITGAX, ITGB2, and *B. burgdorferi* CspZ were derived from the Protein Data Bank (PDB IDs: 2G7I, 2A55, 3OED, 1OJW, 1B09, 3K6S, 1DZI, and 4CBE). The structure of Por1B of *N. gonorrhoeae* was obtained by homology modelling using the Por1A (PDB ID: 4AUI_A) structure as a template; analysis was performed through the SWISS-MODEL server (Arnold *et al.* 2006). The accuracy of the model was assessed with VADAR (Volume, Area, Dihedral Angle Reporter), which uses several algorithms to calculate different parameters for individual residues and for the entire protein (Willard *et al.* 2003). We used these parameters to evaluate both general and residue-specific problems within the newly determined protein structure. The protein model produced had a low number of packaging defects and structural parameters (angles, dihedrals, buried charges) were in perfect agreement with expected values calculated on the sequence.

Structures of the CFH-OspE, CFH-FHbp, and CD59-ILY complexes were retrieved from the Protein Data Bank (PDB IDs: 4J38, 4AYD, and 4BIK, respectively). Protein-protein interaction analyses were performed using PIC (Protein Interaction Calculator) (Tina *et al.* 2007).

Protein-protein docking analysis was performed using ClusPro (Comeau *et al.* 2004). To validate the method, a first run of docking was carried out using the two interacting partners (as in the crystal) to verify that the output was comparable to the three-dimensional structure of the PDB file. After mutagenesis, a new run of docking was performed. We analyzed the 10 best cluster structures, to verify whether the 'native protein-protein conformation' can be found among them. Because the sampling on the conformation is very extensive (10^9 combinations), if the 'native conformation' is not found in the clusters, it is safe to assume that this type of binding is no longer stable and the mutation results in a perturbation of the native binding.

In silico mutations were generated with the FoldX tool run-muta (Schymkowitz *et al.* 2005). We performed the specific mutation 5 times to ensure convergence. Images were created using PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.2 Schrödinger, LLC).

Results

Widespread adaptive evolution at primate complement system genes

We analyzed the evolutionary history of 40 complement genes in primates (the complete list of analyzed genes is shown in Table S1, Supporting Information). DNA alignments of orthologous genes from at least 10 species (Table S2, Supporting Information) were generated and screened for the presence of breakpoints using GARD (genetic algorithm recombination detection) (Kosakovsky Pond *et al.* 2006). Only one recombination breakpoint was identified in *ITGAX*, whereas no evidence of recombination was detected in any other genes. The *ITGAX* alignment was therefore split into two sub-regions according to the breakpoint position.

Calculation of the average nonsynonymous/synonymous substitution rate ratio (dN/dS, also referred to as ω) yielded values lower than 1 in all cases, with the exclusion of CD59. This result is in line with the notion that purifying selection is a major force acting on primate coding regions (Lindblad-Toh *et al.* 2011). Nonetheless, positive selection may be localized and target few specific sites or domains. To address this possibility, we applied the likelihood ratio tests (LRT) implemented in the *codeml* program (Yang 1997; Yang 2007). For 18 genes (45% of the total number of genes we analyzed), two neutral models (site models M1a and M7) were rejected in favor of the positive selection models (M2a and M8); results were confirmed using different codon frequency models (Table S6, Supporting Information). These genes were therefore considered to be positively selected (Fig. 1, Table 1). We next applied the Bayes Empirical Bayes (BEB) analysis (Anisimova *et al.* 2002; Yang *et al.* 2005) and the Mixed Effects Model of Evolution (MEME) method (Murrell *et al.* 2012) to identify selected codons (Table S3, Supporting Information). To be conservative, only sites detected using both methods were considered (Table 1).

Positive selection and selective constraint in the human lineage

To perform an in-depth analysis of the more recent evolution of complement system genes in the human lineage, we applied a population genetics-phylogenetics approach. Specifically, we used gammaMap (Wilson *et al.* 2011) that integrates intra-specific variation and inter-specific diversity to estimate the distribution of population-scaled selection coefficients (γ) along coding regions. At the gene level we observed a good correspondence between γ (median over all codons) and dN/dS calculated in primates (Fig. 2), suggesting a relatively constant selective pressure. Notably, some complement regulators and receptors tended to be more selectively constrained in humans than expected based on dN/dS in primates; the most extreme example was accounted for by CD59, which showed strong negative selection in humans and the highest dN/dS values in primates (Fig. 2).

Despite the general preponderance of negative values of γ, gammaMap also identified several codons evolving under positive selection (cumulative probability ≥ 0.75 of $\gamma > 0$) in the human lineage (Table S7, Supporting Information).

Positive selection at interaction surfaces To investigate the role of sites targeted by positive selection - either in the primate phylogeny or in the human lineage - we integrated data from previous experimental analyses that dissected protein-protein interaction at the single amino acid resolution. In the following section we focus on host proteins that directly interact with pathogen-encoded molecules by detailing the location of selected sites in terms of interaction surfaces. In order to provide some quantitative estimation of the strength of pathogen-driven selection, we performed random uniform sampling of selected sites along protein regions.

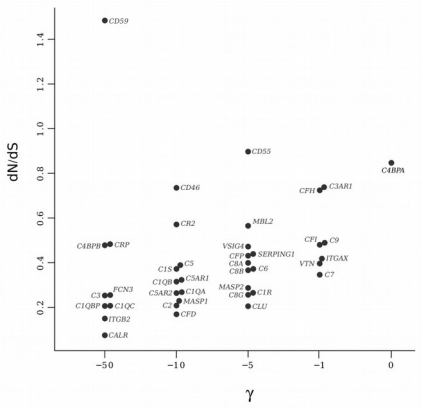


Figure 2. Selective pressure in primates and humans dN/dS (calculated in primates) is plotted against the median of population-scaled selection coefficients (γ) in the human lineage (from gammaMap) for all analyzed complement genes. γ are classified as

weakly deleterious (−1), moderately deleterious (−5, −10) and strongly deleterious (−50, −100). Complement receptors and regulators are shown in blue.

Specifically, we show that in most cases domains that are known to interact with pathogens represent preferential targets of selection.

CFH, C4BPA, CD55, and CR2 are constituted by a different number of complement control protein (CCP) domains. In CFH, CCP domains 6-7 and 19-20, where most positively selected sites are located, are targeted by an extremely wide array of pathogens (Meri *et al.* 2013). We determined that the CCP 19-20 domains represent preferential targets of positive selection: assuming an equiprobable distribution, the likelihood of having 8 selected sites in these regions amounts to $5x10^{-5}$; the observed/expected number of sites (enrichment) amounts to 5.0.

The same calculation yielded a suggestive but non significant p value of 0.07 for CCPs 6-7 (enrichment: 2.4). We exploited recent mutagenesis data (Meri *et al.* 2013) to analyze the overlap between positively selected residues and pathogen binding sites in the CCP 19-20 domain. Results showed that 3 (R1203, R1206, and R1210) of the 6 positions bound by at least 3 distinct pathogens (Meri *et al.* 2013) are positively selected (Fig. 3A). Interestingly, different amino acids at position 1203 are also observed in great apes (N1203 in chimpanzee and H1203 in gorilla) and R1203 represents a human-specific determinant for the binding of sialylated *N. gonorrhoeae* (Shaughnessy *et al.* 2011); indeed, a single N1203R substitution is sufficient to allow binding of sialylated gonococci to chimpanzee CFH.

The molecular details of interaction at CCPs 6-7 are less well characterized, with the exclusion of *N. meningitidis* CFH binding proteins (FHbp) (see below). Recent data, though, indicated that different residues at the positively selected 402 site (which is polymorphic in humans, rs1061170) modulate the binding of *Streptococcus pyogenes* M proteins in a strain-dependent fashion (Nilsson *et al.* 2013).

In C4BPA, pathogen-interacting sites are less localized than in CFH, with some preference for CCPs 1 and 2 (Fig. 3B) (random uniform sampling, p= 0.074; enrichment= 1.8) (Blom & Ram 2008). The CCP1 domain is bound by the porin 1A (Por1A) protein of *N. gonorrhoeae*. Por1A can bind C4BPA of human but not of chimpanzee origin and the two proteins only differ at 4 positions in CCP1 (Ngampasutadol *et al.* 2005). These 4 positions represent the species-specific determinants for binding (Fortin *et al.* 2002), and three of them are positively selected (M62, R70, L82). The M62 and R70 residues also flank the Por1B interaction surface (Fig. 3B) (Fortin *et al.* 2002). Another positively selected site (Y110) is located in the linking region between CCP1 and CCP2, which mediates interaction with complement component C4b. This region is also exploited by several strains of *Streptococcus pyogenes* to interact with C4BP with their M proteins (Blom *et al.* 2000; Jenkins *et al.* 2006).

In CD55, 11 of the 12 positively selected sites are located in the CCP 1-4 domain (random uniform sampling, p = 0.045; enrichment= 1.4). These CD55 domains are used as a receptor by some picornaviruses, including echoviruses 7 and 12, as well as coxsackievirus B3 (Plevka *et al.* 2010). These viruses, albeit related, have evolved different CD55 binding strategies, using distinct CD55 interacting sites located on the receptor surface. Consistent with the arms race hypothesis, we find seven positively selected sites on this CD55 capsid binding surface (Fig. 3D).

As for CR2, it displays 15 CCP domains and most positively selected sites (6 out of 11, random uniform sampling, p = $9x10^{-4}$; enrichment= 4.4) are located in CCP 1-2; these two domains form the binding surface for a C3 proteolytic fragment (C3d) and for the gp350 protein of Epstein-Barr virus (EBV).

Analysis of positively selected sites unveiled other interesting findings. Two positively selected sites are located on the CRP surface: one of these, S92, lies in the phosphocholine (Pch) binding pocket (Gang *et al.* 2012) (Fig. 3E). CRP binds *Streptococcus pneumoniae* through Pch residues present in pneumococcal C-polysaccharide, and similar Pch molecules are thought to be important for CRP binding to other bacterial species and non-bacterial parasites (Weiser *et al.* 1998; Gillespie *et al.* 1993; Culley *et al.* 1996).

Vitronectin (VTN) and CD59 act as complement regulators and interact with pathogens. In CD59 all positively selected sites

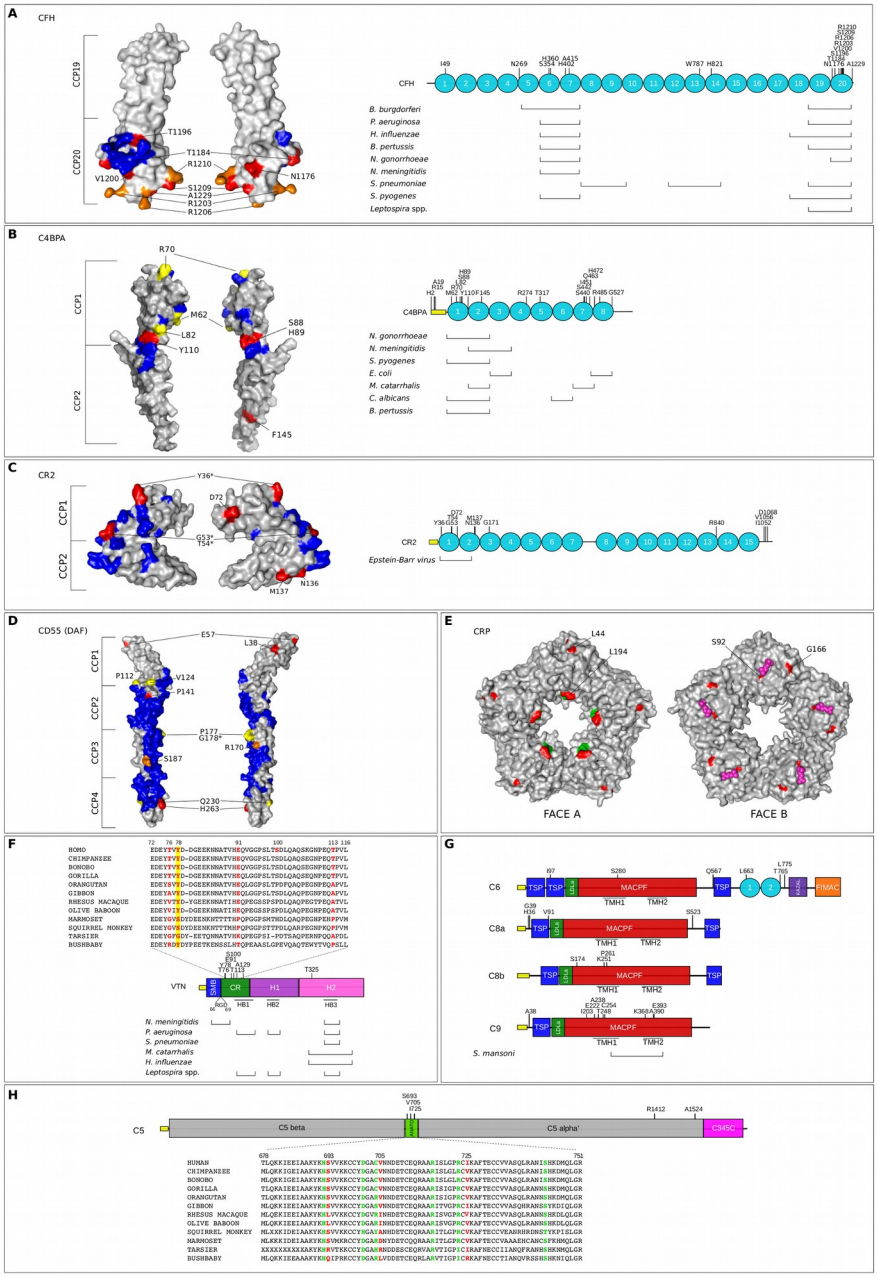


Figure 3. Positively selected sites in complement system proteins Domain representation of positively selected genes or/and 3D mapping of selected sites for CFH (A), C4BPA (B), CR2 (C), CD55 (D), CRP (E), VTN (F), MAC components (G), and C5 (H). Multiple alignments of the VTN N-terminal portion (F) and C5 anaphylatoxin (H) for representative primate species are reported. Positively selected sites detected by PAML and/or gammaMap analyses are shown. In structures and alignment, color codes are as follows: blue, residues involved in pathogen binding; green, residues involved in interactions among complement components; red, positively selected sites; yellow, positively selected sites interacting with a single pathogen; orange, positively selected sites involved in the binding of more than one pathogen; asterisks denote positively selected sites involved in the interaction among

complement components. Positively selected sites are reported in black on domain representation. Magenta spheres in the CRP structure (E) represent phosphocoline molecules. CCP, complement control protein domains (cyan circles); signal peptides (yellow rectangles); SMB, Somatomedin-B domain; RGD motif, integrin binding domain; CR, connecting region; H1-H2, Haemopexin-like domains; HB1-3, heparin-binding domains; TSP, thrombospondin type 1 repeats; LDLa, low-density lipoprotein receptor domain class A; MACPF, membrane-attack complex/perforin; KAZAL, kazal type serine protease inhibitors and follistatin-like domain; FIMAC, factor I membrane attack complex domain; TMH1-TMH2, transmembrane helices; ANATO, anaphylatoxin; C345C, netrin C-terminal domain.

are located in the short region that constitutes the mature protein (after removal of the signal peptide and propeptide) (random uniform sampling, p= 0.0083; enrichment= 1.7). A detailed analysis of selection at CD59 is reported below. As for VTN, most interactions with pathogens involve the heparin-binding region 3 (HB3), where no positively selected sites are located. In the same protein, the short "connecting region" (CR) represented a preferential selection target (random uniform sampling, p =$1x10^{-4}$; enrichment= 5.3) (Fig. 3). The CR is bound by the *N. meningitidis* Opc protein; in humans, Y78 is a sulphated tyrosine that directly interacts with Opc (Sa E Cunha *et al.* 2010) (Fig. 3F).

ITGAX encodes the α subunit of integrin αxβ2, also called complement receptor type 4. All the positively selected sites (Table 1) are located in the αI domain, a ~200 amino acid long structure that contains the metal ion-dependent adhesion site (MIDAS) (random uniform sampling, p = $7x10^{-4}$; enrichment= 6.1). The αI domain of αxβ2 and α2β1 integrins is bound by some rotaviruses (Fleming *et al.* 2011). By 3D-structure superimposition, three ITGAX selected sites (R238, H241 and Y244) were found to co-localize with α2β1 integrin residues that are involved in collagen and rotavirus binding (Fleming *et al.* 2011) (Fig. S1, Supporting Information).

Although not directly involved in interaction with pathogens, positively selected sites were also detected in the complement components responsible for MAC assembly. Most of them localize in the membrane attack complex perforin domain (MACPF) (Fig. 3G), and in particular in two functional regions that integrate into the plasma membrane (TMH elements) (Sonnen & Henneke 2014). In C9 one of these regions is targeted by the paramyosin of *Schistosoma mansoni* (Deng *et al.* 2003); a positively selected site (A390) in humans is also part of the primary CD59 recognition site (Huang *et al.* 2006).

Finally, we observed that 3 out of 5 positively selected residues in C5 are within the 73 amino acid region that constitutes anaphylatoxin C5a (Fig. 3H).

Adaptive evolution at bacterial-encoded complement regulators

Because host-pathogen conflicts are expected to shape genetic diversity both in the host's and in the pathogen's genomes, we investigated whether microbial genes that encode complement-evasion molecules were targeted by positive selection. In particular, we focused on bacterial proteins that have been characterized to such an extent that the molecular determinants of the interaction with CFH or C4BP, the two most common targets, are known. FHbp encoded by *N. meningitidis* were not analyzed because their evolutionary pattern has previously been described, and positively selected sites localize outside the CFH binding interface (Brehony *et al.* 2009).

Because recombination rates in bacterial genomes can be very high (Awadalla 2003), analyses were performed using omegaMap, a Bayesian method that simultaneously estimates recombination and selection (inferred through ω estimation) (Wilson & McVean 2006). The details of bacterial sequences used for these analyses are reported in the methods section.

The *porB* locus of *N. gonorrhoeae* was analyzed by separating *Por1A* and *Por1B* alleles, which form distinct monophyletic groups (Smith *et al.* 1995). A previous evolutionary analysis of *Por1A* and *Por1B* had detected several sites subject to positive selection, but recombination was not accounted for (Perez-Losada *et al.* 2005). omegaMap indicated high recombination at both loci and, by accounting for this effect, detected positive selection at several surface-exposed loop regions, some of which are implicated in the binding of CFH or C4BP (Figs 4A and 4B).

Likewise, multiple alleles have been described for PspC (also known as CbpA), a major surface-exposed protein of *Streptococcus pneumoniae*. We analyzed "typical" PspC sequences only; these have a cholin anchor and bind CFH (Iannelli *et al.* 2002). Positively selected sites were mainly found to be located at the N-terminus of PspC and include a short amino acid stretch which binds CFH (Fig. 4C).

Finally, we analyzed complement regulator-acquiring surface proteins (CRASPs) encoded by *Borrelia* species. In CspZ two signals of positive selection were detected (Fig. 4D); one of these contains a residue (N51) involved in the interaction with CFH or CFHR1 (Brangulis *et al.* 2014). As for OspE, a single signal of positive selection mapping to the CFH binding region was detected (Bhattacharjee *et al.* 2013) (Fig. 4E).

The effect of positively selected sites on protein-protein interactions

To gain further insight into the effect of positive selection on host-pathogen interactions, we applied an integrated approach based on previous *in vitro* experiments, *in silico* mutagenesis, and protein-protein docking. Using this strategy we analyzed the interaction between CD59 and intermedilysin (ILY), as well as the binding of CFH to OspE and FHbp. We stress that, although docking results were consistent with *in vitro* data, these analyses only provide qualitative information on the role of specific sites.

CD59 is bound by members of the cholesterol-dependent cytolysin family, including ILY, vaginolysin, and lectinolysin (produced by *Streptococcus intermedious* and *Gardnerella vaginalis*, respectively) (Wickham *et al.* 2011). These proteins are thought to share similar mechanisms of binding to CD59 and to use residues in this receptor that are also involved in the interaction with C8α/C9 (Wickham *et al.* 2011; Huang *et al.* 2005; Johnson *et al.* 2013). The ILY-CD59 interaction has been studied in detail and occurs at two major interfaces. The primary binding sites in CD59 are involved in the recruitment of ILY monomers to the host plasma membrane, while the secondary binding sites, located on the opposite side of CD59, facilitate intermolecular interactions that drive the formation of the ILY prepore ring (Johnson *et al.* 2013).

Four of the positively selected sites in CD59 are directly involved in the interaction with ILY. Y87 and D47 are primary and secondary binding sites, respectively; R80 forms a hydrogen bond with ILY D443, while K91 forms a salt-bridge with ILY D445, these two latter contacts are important to stabilize binding interfaces (Johnson *et al.* 2013). 3D structural mapping indicated that L100 is also located at the primary CD59-ILY interaction surface (Fig. 5A). In line with the notion that positive selection is pathogen-driven, the *in vitro* replacement of D47 with different residues modulates CD59 binding affinity for ILY and affects cytolytic activity (Wickham *et al.* 2011). Likewise, alanine mutagenesis of Y87 (Y62 in (Wickham *et al.* 2011)) was shown to increase the dissociation constant of the CD59-ILY interaction. We exploited this *in vitro* experiment to assess whether *in silico* mutagenesis and protein-protein docking can recapitulate this result. Docking analysis of wild-type human CD59 and ILY resulted in an interaction pose that is fully consistent with the three-dimensional structure of the solved complex; *in silico* mutagenesis of Y87 to alanine resulted in a displacement of the interaction partners (Fig. 5A), in agreement with the *in vitro* data (Wickham *et al.* 2011). We thus used this strategy to address the effect of amino acid replacements at positively selected sites. Mutagenesis was performed by replacing the human residue with the most common amino acid observed in primate sequences. Mutations Y87H, K91E, and L100P resulted in a complete misplacement of the interaction, strongly suggesting that selection-driven changes at these sites modulates CD59-ILY binding (Fig. 5A). It is worth mentioning that ILY was not analyzed for selection signatures because only 6 sequences were available. In the CFH-OspE interaction both partners are targeted by selection (Figs 3A and 4E). In OspE, two positively selected sites (V120 and I121) are located at the binding interface (Bhattacharjee *et al.* 2013) (Fig. 5B); both residues interact through hydrogen bonds with CFH, pointing to their central role in this process (Bhattacharjee *et al.* 2013). n CFH, positively selected sites are also located at the binding interface (e.g. S1196 interacts with S82 in OspE). As above, we first performed *in silico* alanine mutagenesis of a CFH residue (K1186) that, albeit being located at the interaction surface, does not modulate binding (Meri *et al.* 2013). In accordance with the *in*

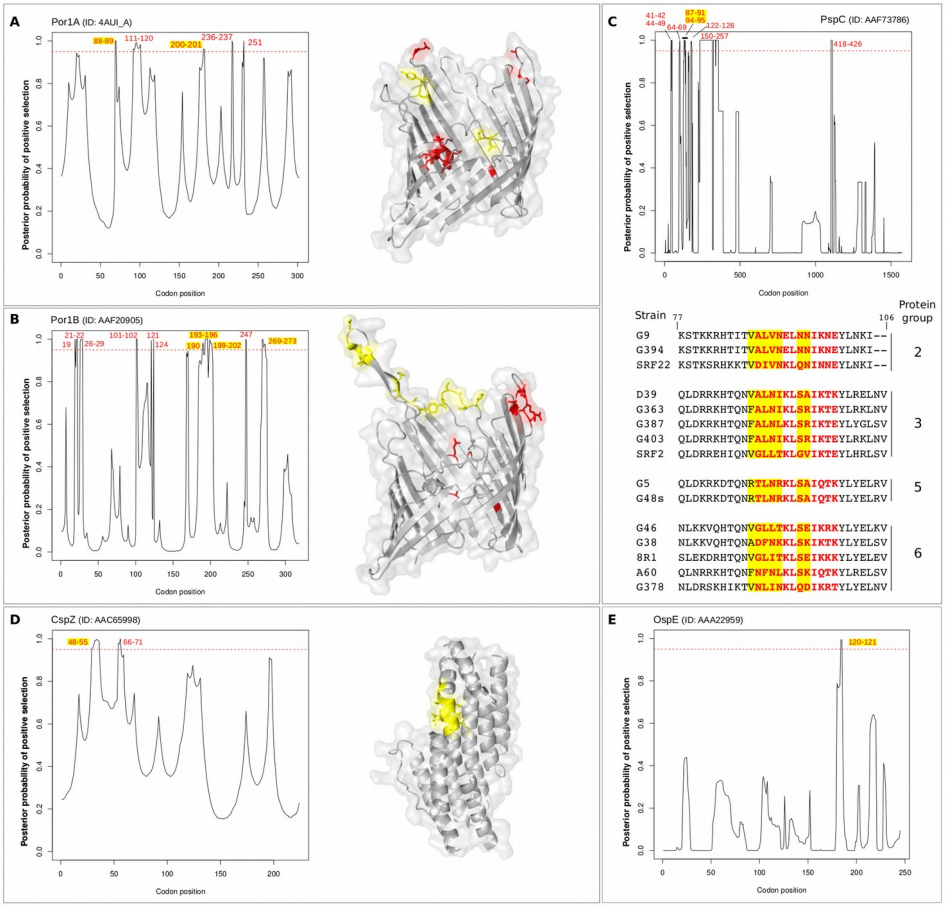


Figure 4. Positive selection at bacterial interactors Plots report the posterior probability of positive selection (dN/dS >1) along coding sequences. The hatched red lines correspond to a posterior probability of selection equal to 0.95. Positively selected codons or blocks are reported in red. Codons or blocks that participate in contact with complement proteins are highlighted in yellow. When available, sites were mapped onto 3D structures with the same color-code as in the plot. Data refer to *N. gonorrhoeae* Por1A (A) and Por1B (B), *S. Pneumoniae* PspC (C), *B. burgdorferi* CspZ (D) and OspE (E). For PspC a portion of the alignment is shown with color codes as in plots. Codon positions refer to reference strain indicated in each panel.

vitro data (Meri *et al.* 2013), docking results revealed no displacement (Fig. 5B). We thus analyzed the effect on binding of positively selected sites in OspE and CFH. Mutations V120L and I121A in OspE, as well as T1184K and S1196P in CFH, led to a misplacement of the interactors, strongly suggesting that they affect protein-protein interaction (Fig. 5B). Notably, variation at positions 120 and 121 in OspE has been proposed to contribute to complement resistance in different *Borrelia* species (Alitalo *et al.* 2005).

We add that T1184 was previously *in vitro* mutagenized and reported to decrease OspE binding, although not significantly (Meri *et al.* 2013). Nonetheless, the human threonine residue had been replaced with a different amino acid to the one we introduced and the effect on binding was much more evident for the T1184 mutant than for the K1186 replacement, that we also chain/side-chain interaction between H402 in CFH and H184 in FHbp. In line with previous reports (Schneider *et al.* 2009), mutagenesis of the CFH positively

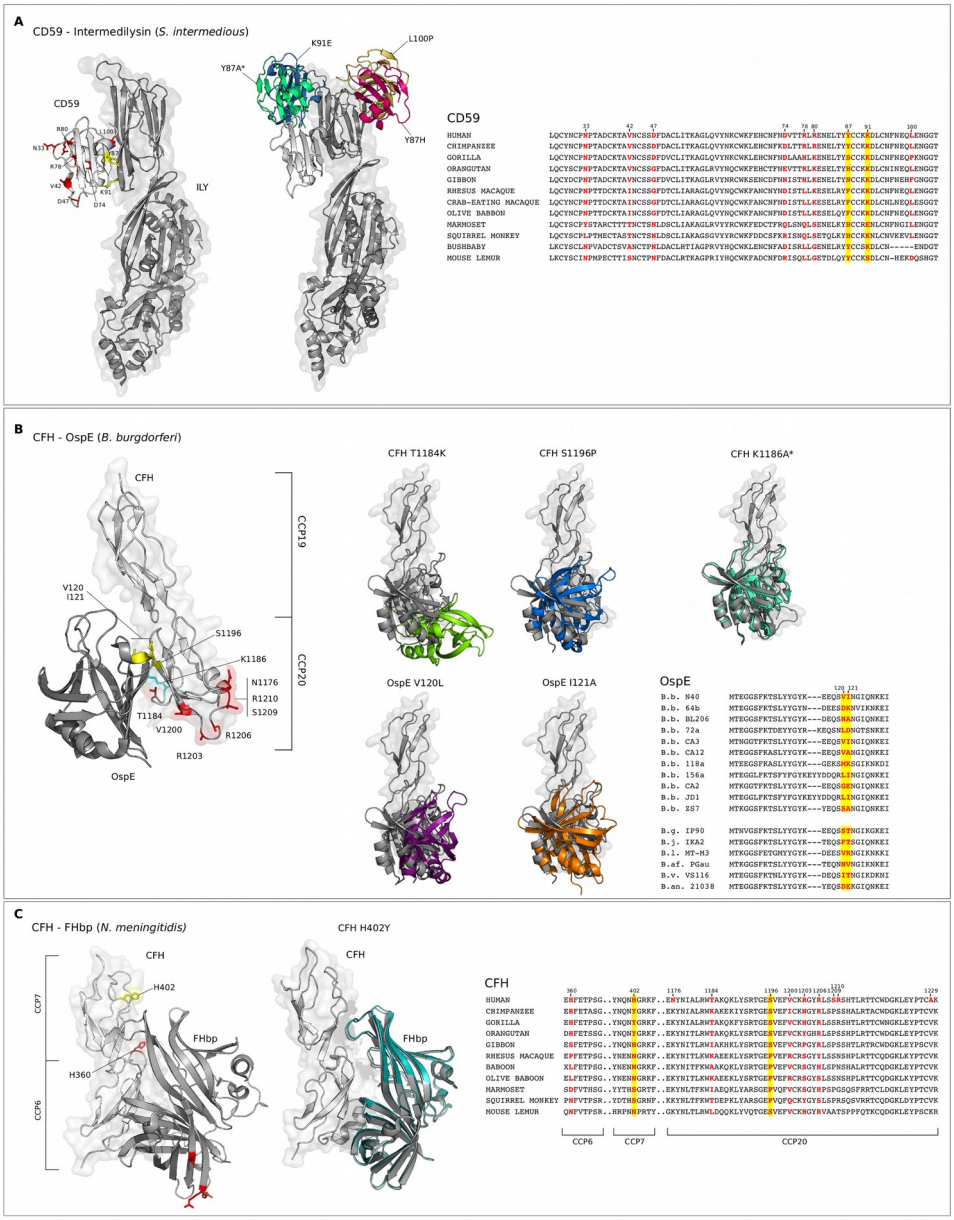


Figure 5. *In silico* mutagenesis and protein-protein docking (A) 3D structure of the CD59-ILY complex (left), docking analysis (middle) and portion of CD59 multiple alignment (right). Sites (structure and alignment) are color-coded with positively selected sites in red and positively selected sites involved in protein-protein interactions at primary binding sites in yellow. Results of docking analysis after mutagenesis of different residues in CD59 are colored in blue, green, magenta, and brown; docking with the un-mutated (human) protein is in grey. (B) 3D structure of the CFH-OspE complex, docking analysis (middle) and portion of OspE multiple alignment (right). Sites are color-coded as in A. Residue K1186, which was mutated as a negative control is in cyan. Protein-protein docking was performed after mutation of CFH (upper panels) or OspE (lower panels) and shown in different colors. Results with the un-mutated proteins are shown in grey. In the alignment, abbreviations are as follows: B.b., *Borrelia burgdorferi*; B.g., *Borrelia garinii*; B.j., *Borrelia japonica*, B.l. *Borrelia lusitaniae*; B.af., *Borrelia afzelii*; B.v., *Borrelia valaisiana*; B.an., *Borrelia andersonii*. In (A) and (B) asterisks denote mutants used as controls for docking experiments. (C) 3D structure of the CFH-FHbp complex, docking analysis (middle) and portion of CFH multiple alignment (right). Color codes are as in (A) and (B). Positively selected sites in FHbp derive from a previous work (Brehony *et al*. 2009).

selected site H402 and docking analysis resulted in no modification of the binding pose (Fig. 5C). This is probably due to H402 lying at the edge of the binding surface, with little contribution to the overall interaction energy. confirmed as inconsequential for binding (Meri *et al.* 2013).

Finally, interaction analysis of the CFH-FHbp complex highlighted the presence of a main-chain/side-chain interaction between H402 in CFH and H184 in FHbp. In line with previous reports (Schneider *et al.* 2009), mutagenesis of the CFH positively selected site H402 and docking analysis resulted in no modification of the binding pose (Fig. 5C). This is probably due to H402 lying at the edge of the binding surface, with little contribution to the overall interaction energy.

Discussion

This study was motivated by the observations that complement components are targeted by an extremely wide range of pathogens and that complement evasion underlies the virulence and/or outcome of several infections. Thus, the complement system is extremely important from a biomedical perspective, its microbial interactors are often used for vaccine design, and the system plus its interactors exemplify how host-pathogen conflicts play out.

We provide a comprehensive catalog of sites in complement system genes that were targeted by positive selection in primates and in the human lineage, and we report positive selection at bacterial interactors. We also integrated evolutionary analysis and docking studies with previous biochemical *in vitro* data to test the concept that arms race scenarios mainly involve host-pathogen interaction surfaces.

Results show that host components targeted by several microbial/viral proteins evolved under the strongest selective pressure in primates; these include CFH, CD59, C4BPA, and CD55. Selection acted at these genes and at others (e.g. *CR2* and *VTN*) on residues that are located at the interaction interface with microbial/viral components. In CFH, for instance, three sites we detected were previously shown to be bound by different pathogens belonging to distinct phyla (Meri *et al.* 2013), indicating very strong and long-standing selective

pressures. One of these sites also represents a human-specific determinant for the binding of *N. gonorrhoeae*, as is the case for three positively selected sites in C4BPA (Shaughnessy *et al.* 2011; Jarva *et al.* 2007). Partially because of the resistance to human complement-mediated killing, natural infectionwith *N. gonorrhoeae* is restricted to our species (Ngampasutadol *et al.* 2008). Examples of selection-driven species-specific susceptibility to infection have previously been reported for viral pathogens (Sawyer & Elde 2012; Daugherty & Malik 2012), but rarely for bacteria (Barber & Elde 2014).

Another member of the *Neisseria* genus, *N. meningitidis*, only infects humans and displays human-specific binding to CFH (Granoff *et al.* 2009). The only positively selected CFH site (H402) at the interaction surface with FHbp is not involved in modulating binding. Conversely, its evolution may be driven by staphylococcal M proteins, as different residues at the 402 site modulate the binding affinity of *Streptococcus pyogenes* M proteins (Nilsson *et al.* 2013). In line with this finding, positive selection at FHbp was previously shown to involve regions that do not contact CFH (Brehony *et al.* 2009), indicating that this latter and FHbp are not engaged in a mutual genetic conflict. One possible reason for this observation is that *N. meningitidis* has redundant mechanisms of complement evasion and the use of FHbp may be recent and dispensable; indeed, invasive meningococcal strains that lack FHbp have been isolated and shown to specifically bind human CFH (CCPs 6-7) through PorB2 (Lewis *et al.* 2013).

Unlike FHbp, a portion of positively selected sites in the other bacterial proteins we analyzed were found to be involved in the binding of the host interactor. These results reflect the expectations under a genetic conflict scenario whereby the host's and the pathogen's genes evolve within binding avoidance-binding seeking dynamics (Sironi *et al.* 2015). We provide evidence in favor of this possibility through protein-protein docking analyses. In particular, we exploited previous data to validate an *in silico* mutagenesis and docking strategy that was subsequently used to qualitatively evaluate the effect of selected sites. To this aim, a subset of host-pathogen protein-protein interactions was selected based on the

availability of the 3D structure of the complex. These analyses revealed that most positively selected sites we analyzed modulate binding between interactors. We note that the protein-protein docking program we used returns the best putative complex between the two protein structures given as input. Thus, the observed repositioning of the ligand after *in silico* mutagenesis can reliably be interpreted as an effect on the binding efficiency; the magnitude of this effect, though, cannot be estimated.

The host-pathogen interactions we analyzed herein are necessarily limited to those that have been reported and characterized in detail. Most likely, a large number of molecular interactions that involve complement system components and pathogen-encoded molecules remain to be detected or characterized; these interactions cannot therefore be analyzed within the framework we applied herein. This represents a limitation of our study, but the selected sites we identified may be prioritized in future functional analyses. For instance, we detected several selected sites in complement factor I (CFI) and properdin (CFP). This latter is recruited at the surface of *Chlamydia pneumoniae*, binds fungal glycans via C3, and interacts with LPS from certain *Escherichia coli* strains (Spitzer *et al.* 2007; Agarwal *et al.* 2011; Cortes *et al.* 2011). Likewise, CFI is targeted by important human pathogens such as *Staphylococcus aureus* and *Prevotella intermedia* (Malm *et al.* 2012; Hair *et al.* 2010). The protein regions mediating these interactions are unknown, but are expected to evolve under pathogen-exerted selective pressure and possibly involve some of the sites we describe herein. We note, however, that we failed to detect positive selection at complement components that do interact with numerous pathogens. One example is C3; *S. aureus* alone expresses at least three distinct molecules that bind C3 to prevent complement activation (Lambris *et al.* 2008). The lack of selection observed at this gene, as well as at others, may result either from the conservative approach we applied or from a lack of power due to the relatively small number of primate sequences we included. Still, C3 displays both a low dN/dS (0.25) and a signal of purifying selection in humans (median γ =-1). Thus, its central role in complement activation may impose constraint on C3 evolution and limit the sequence space accessible for adaptive evolution. In this respect, it is worth noting that several complement regulators (e.g CD59, CD55, CD46, CFH) and receptors (CR2 and C3AR1) tend to display stronger selective constraint in the human lineage than expected on the basis of their selective pattern in primates. The underlying selective forces responsible for this observation remain to be clarified. Infections are believed to have represented the major cause of mortality in both early human populations and in great apes (Finch 2010). Nonetheless, human-specific changes in immunological phenotypes have previously been described (O'Bleness *et al.* 2012). Unique among hominids, for instance, human T cells express little or no regulatory SIGLECs (sialic acid-recognizing Ig-superfamily lectins), resulting in more vigorous activation (Nguyen *et al.* 2006). One interesting possibility is that shifts in selective pressure have arisen as a cause of the increased pathogen load associated with human-specific features such as large settlements and extensive inter-group contacts (O'Bleness *et al.* 2012).

Finally, we note that several pathogens we included in this study are extremely important from a clinical perspective and are responsible for substantial morbidity and mortality. Bacterial proteins that interact with complement components have been regarded as attractive vaccine candidates, because vaccine efficacy is likely to be enhanced by virulence impairment (Meri *et al.* 2008). The N-terminal domain of PspC that interacts with CFH has been tested as a possible vaccine antigen in mice with promising results, although protection was influenced by the strain used for immunization and for challenge (Ricci *et al.* 2011). This effect was due to the high diversity of PspC molecules (Ricci *et al.* 2011) that, as shown herein, is driven by natural selection. Deeper understanding of the evolutionary dynamics of pathogen-encoded molecules targeted by vaccines or drugs may provide information on the most promising strategies for the development of novel interventions (Jefferies *et al.* 2011; Little *et al.* 2012).

In summary, in this study we use extensive data of inter- and intra-specific genetic diversity to show that complement proteins bound by several microbial components evolved under strong selective pressure at the host-pathogen

interaction interface. Bacterial proteins also evolved adaptively and positively selected sites are involved in the binding of the host complement interactor. These data provide insight into the human-specific susceptibility to gonorrhea and serve to validate more general hypotheses on host-pathogen conflicts.

Acknowledgments
CP is supported by a fellowship of the Doctorate School of Molecular Medicine, University of Milan**.**

Data Accessibility
DNA sequences accession numbers are reported in Tables S4, Supporting Information. Alignments and input files for all analyses have been submitted to Dryad (provisional DOI: doi:10.5061/dryad.5jt25).

Author contributions
MS and RC conceived the study; MS, RC, and MC supervised the project; CP, UP, and AM performed the evolutionary analysis in primates, with input from RC and DF; GF, AM, and DF performed the 3D mapping analyses; RC, DF, and NB performed the lineage-specific selection analyses; RC and DF performed the evolutionary analysis of bacterial genes; LDG and GF performed the *in silico* mutagenesis and docking studies; CP, AM and DF produced the figures, with input from all authors; UP provided support during the bioinformatic analyses; MS and RC wrote the manuscript, with critical input from MC and from the remaining authors.

References

1000 Genomes Project Consortium, Durbin RM, Abecasis GR, et al (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061-1073.

Agarwal S, Specht CA, Haibin H, et al (2011) Linkage specificity and role of properdin in activation of the alternative complement pathway by fungal glycans. *mBio*, **2**, 10.1128/mBio.00178-11. Print 2011.

Alitalo A, Meri T, Comstedt P, et al (2005) Expression of complement factor H binding immunoevasion proteins in Borrelia garinii isolated from patients with neuroborreliosis. *European journal of immunology*, **35**, 3043-3053.

Anisimova M, Bielawski JP, Yang Z (2002) Accuracy and power of bayes prediction of amino acid sites under positive selection. *Molecular biology and evolution*, **19**, 950-958.

Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics (Oxford, England)*, **22**, 195-201.

Awadalla P (2003) The evolutionary genomics of pathogen recombination. *Nature reviews.Genetics*, **4**, 50-60.

Ballanti E, Perricone C, Greco E, et al (2013) Complement and autoimmunity. *Immunologic research*, **56**, 477-491.

Barber MF, Elde NC (2014) Nutritional immunity. Escape from bacterial iron piracy through rapid evolution of transferrin. *Science (New York, N.Y.)*, **346**, 1362-1366.

Bhattacharjee A, Oeemig JS, Kolodziejczyk R, et al (2013) Structural basis for complement evasion by Lyme disease pathogen Borrelia burgdorferi. *The Journal of biological chemistry*, **288**, 18685-18695.

Blom AM, Berggard K, Webb JH, Lindahl G, Villoutreix BO, Dahlback B (2000) Human C4b-binding protein has overlapping, but not identical, binding sites for C4b and streptococcal M proteins. *Journal of immunology (Baltimore, Md.: 1950)*, **164**, 5328-5336.

Blom AM, Ram S (2008) Contribution of interactions between complement inhibitor C4b-binding protein and pathogens to their ability to establish infection with particular emphasis on Neisseria gonorrhoeae. *Vaccine*, **26 Suppl 8**, I49-55.

Brangulis K, Petrovskis I, Kazaks A, et al (2014) Structural characterization of CspZ, a

complement regulator factor H and FHL-1 binding protein from Borrelia burgdorferi. *The FEBS journal*, **281**, 2613-2622.

Brehony C, Wilson DJ, Maiden MC (2009) Variation of the factor H-binding protein of Neisseria meningitidis. *Microbiology (Reading, England)*, **155**, 4155-4169.

Comeau SR, Gatchell DW, Vajda S, Camacho CJ (2004) ClusPro: a fully automated algorithm for protein-protein docking. *Nucleic acids research*, **32**, W96-9.

Cortes C, Ferreira VP, Pangburn MK (2011) Native properdin binds to Chlamydia pneumoniae and promotes complement activation. *Infection and immunity*, **79**, 724-731.

Culley FJ, Harris RA, Kaye PM, McAdam KP, Raynes JG (1996) C-reactive protein binds to a novel ligand on Leishmania donovani and increases uptake into human macrophages. *Journal of immunology (Baltimore, Md.: 1950)*, **156**, 4691-4696.

Daugherty MD, Malik HS (2012) Rules of engagement: molecular insights from host-virus arms races. *Annual Review of Genetics*, **46**, 677-700.

Delport W, Poon AF, Frost SD, Kosakovsky Pond SL (2010) Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics (Oxford, England)*, **26**, 2455-2457.

Demogines A, Abraham J, Choe H, Farzan M, Sawyer SL (2013) Dual host-virus arms races shape an essential housekeeping protein. *PLoS biology*, **11**, e1001571.

Deng J, Gold D, LoVerde PT, Fishelson Z (2003) Inhibition of the complement membrane attack complex by Schistosoma mansoni paramyosin. *Infection and immunity*, **71**, 6402-6410.

Finch CE (2010) Evolution in health and medicine Sackler colloquium: Evolution of the human lifespan and diseases of aging: roles of infection, inflammation, and nutrition. *Proceedings of the National Academy of Sciences of the United States of America*, **107 Suppl 1**, 1718-1724.

Fleming FE, Graham KL, Takada Y, Coulson BS (2011) Determinants of the specificity of rotavirus interactions with the alpha2beta1 integrin. *The Journal of biological chemistry*, **286**, 6165-6174.

Fortin A, Stevenson MM, Gros P (2002) Susceptibility to malaria as a complex trait: big pressure from a tiny creature. *Human molecular genetics*, **11**, 2469-2478.

Gang TB, Hammond DJ,Jr, Singh SK, Ferguson DA,Jr, Mishra VK, Agrawal A (2012) The phosphocholine-binding pocket on C-reactive protein is necessary for initial protection of mice against pneumococcal infection. *The Journal of biological chemistry*, **287**, 43116-43125.

Gillespie SH, McWhinney PH, Patel S, et al (1993) Species of alpha-hemolytic streptococci possessing a C-polysaccharide phosphorylcholine-containing antigen. *Infection and immunity*, **61**, 3076-3077.

Granoff DM, Welsch JA, Ram S (2009) Binding of complement factor H (fH) to Neisseria meningitidis is specific for human fH and inhibits complement activation by rat and rabbit sera. *Infection and immunity*, **77**, 764-769.

Guindon S, Delsuc F, Dufayard JF, Gascuel O (2009) Estimating maximum likelihood phylogenies with PhyML. *Methods in molecular biology (Clifton, N.J.)*, **537**, 113-137.

Hair PS, Echague CG, Sholl AM, et al (2010) Clumping factor A interaction with complement factor I increases C3b cleavage on the bacterial surface of Staphylococcus aureus and decreases complement-mediated phagocytosis. *Infection and immunity*, **78**, 1717-1727.

Huang Y, Qiao F, Abagyan R, Hazard S, Tomlinson S (2006) Defining the CD59-C9 binding interaction. *The Journal of biological chemistry*, **281**, 27398-27404.

Huang Y, Smith CA, Song H, Morgan BP, Abagyan R, Tomlinson S (2005) Insights into the human CD59 complement binding interface

toward engineering new therapeutics. *The Journal of biological chemistry*, **280**, 34073-34079.

Iannelli F, Oggioni MR, Pozzi G (2002) Allelic variation in the highly polymorphic locus pspC of Streptococcus pneumoniae. *Gene*, **284**, 63-71.

Jarva H, Ngampasutadol J, Ram S, Rice PA, Villoutreix BO, Blom AM (2007) Molecular characterization of the interaction between porins of Neisseria gonorrhoeae and C4b-binding protein. *Journal of immunology (Baltimore, Md.: 1950)*, **179**, 540-547.

Jefferies JM, Clarke SC, Webb JS, Kraaijeveld AR (2011) Risk of red queen dynamics in pneumococcal vaccine strategy. *Trends in microbiology*, **19**, 377-381.

Jenkins HT, Mark L, Ball G, et al (2006) Human C4b-binding protein, structural basis for interaction with streptococcal M protein, a major bacterial virulence factor. *The Journal of biological chemistry*, **281**, 3690-3697.

Johnson S, Brooks NJ, Smith RA, Lea SM, Bubeck D (2013) Structural basis for recognition of the pore-forming toxin intermedilysin by human complement receptor CD59. *Cell reports*, **3**, 1369-1377.

Kosakovsky Pond SL, Frost SD (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular biology and evolution*, **22**, 1208-1222.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD (2006) Automated phylogenetic detection of recombination using a genetic algorithm. *Molecular biology and evolution*, **23**, 1891-1901.

Lambris JD, Ricklin D, Geisbrecht BV (2008) Complement evasion by human pathogens. *Nature reviews.Microbiology*, **6**, 132-142.

Lewis LA, Vu DM, Vasudhev S, Shaughnessy J, Granoff DM, Ram S (2013) Factor H-dependent alternative pathway inhibition mediated by porin B contributes to virulence of Neisseria meningitidis. *mBio*, **4**, e00339-13.

Lindblad-Toh K, Garber M, Zuk O, et al (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, **478**, 476-482.

Little TJ, Allen JE, Babayan SA, Matthews KR, Colegrave N (2012) Harnessing evolutionary biology to combat infectious disease. *Nature medicine*, **18**, 217-220.

Malm S, Jusko M, Eick S, Potempa J, Riesbeck K, Blom AM (2012) Acquisition of complement inhibitor serine protease factor I and its cofactors C4b-binding protein and factor H by Prevotella intermedia. *PloS one*, **7**, e34852.

Meri S, Jordens M, Jarva H (2008) Microbial complement inhibitors as vaccines. *Vaccine*, **26 Suppl 8**, I113-7.

Meri T, Amdahl H, Lehtinen MJ, et al (2013) Microbes bind complement inhibitor factor H via a common site. *PLoS pathogens*, **9**, e1003308.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL (2012) Detecting individual sites subject to episodic diversifying selection. *PLoS genetics*, **8**, e1002764.

Ngampasutadol J, Ram S, Blom AM, et al (2005) Human C4b-binding protein selectively interacts with Neisseria gonorrhoeae and results in species-specific infection. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 17142-17147.

Ngampasutadol J, Tran C, Gulati S, et al (2008) Species-specificity of Neisseria gonorrhoeae infection: do human complement regulators contribute? *Vaccine*, **26 Suppl 8**, I62-6.

Nguyen DH, Hurtado-Ziola N, Gagneux P, Varki A (2006) Loss of Siglec expression on T lymphocytes during human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 7765-7770.

Nilsson OR, Lannergard J, Morgan BP, Lindahl G, Gustafsson MC (2013) Affinity purification of human factor H on polypeptides derived from streptococcal m protein: enrichment of the Y402 variant. *PloS one*, **8**, e81303.

O'Bleness M, Searles VB, Varki A, Gagneux P, Sikela JM (2012) Evolution of genetic and genomic features unique to the human lineage. *Nature reviews.Genetics*, **13**, 853-866.

Pangburn MK, Ferreira VP, Cortes C (2008) Discrimination between host and pathogens by the complement system. *Vaccine*, **26 Suppl 8**, I15-21.

Penn O, Privman E, Ashkenazy H, Landan G, Graur D, Pupko T (2010) GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic acids research*, **38**, W23-8.

Perez-Losada M, Viscidi RP, Demma JC, Zenilman J, Crandall KA (2005) Population genetics of Neisseria gonorrhoeae in a high-prevalence community using a hypervariable outer membrane porB and 13 slowly evolving housekeeping genes. *Molecular biology and evolution*, **22**, 1887-1902.

Plevka P, Hafenstein S, Harris KG, et al (2010) Interaction of decay-accelerating factor with echovirus 7. *Journal of virology*, **84**, 12665-12674.

Pond SL, Frost SD, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics (Oxford, England)*, **21**, 676-679.

Posada D, Crandall KA, Nguyen M, Demma JC, Viscidi RP (2000) Population genetics of the porB gene of Neisseria gonorrhoeae: different dynamics in different homology groups. *Molecular biology and evolution*, **17**, 423-436.

Privman E, Penn O, Pupko T (2012) Improving the performance of positive selection inference by filtering unreliable alignment regions. *Molecular biology and evolution*, **29**, 1-5.

Ricci S, Janulczyk R, Gerlini A, et al (2011) The factor H-binding fragment of PspC as a vaccine antigen for the induction of protective humoral immunity against experimental pneumococcal sepsis. *Vaccine*, **29**, 8241-8249.

Ricklin D, Hajishengallis G, Yang K, Lambris JD (2010) Complement: a key system for immune surveillance and homeostasis. *Nature immunology*, **11**, 785-797.

Ricklin D, Lambris JD (2013) Complement in immune and inflammatory disorders: pathophysiological mechanisms. *Journal of immunology (Baltimore, Md.: 1950)*, **190**, 3831-3838.

Rogers EA, Abdunnur SV, McDowell JV, Marconi RT (2009) Comparative analysis of the properties and ligand binding characteristics of CspZ, a factor H binding protein, derived from Borrelia burgdorferi isolates of human origin. *Infection and immunity*, **77**, 4396-4405.

Sa E Cunha C, Griffiths NJ, Virji M (2010) Neisseria meningitidis Opc invasin binds to the sulphated tyrosines of activated vitronectin to attach to and invade human brain endothelial cells. *PLoS pathogens*, **6**, e1000911.

Sawyer SL, Elde NC (2012) A cross-species view on viruses. *Current opinion in virology*, **2**, 561-568.

Schneider MC, Prosser BE, Caesar JJ, et al (2009) Neisseria meningitidis recruits factor H using protein mimicry of host carbohydrates. *Nature*, **458**, 890-893.

Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L (2005) The FoldX web server: an online force field. *Nucleic acids research*, **33**, W382-8.

Serruto D, Rappuoli R, Scarselli M, Gros P, van Strijp JA (2010) Molecular mechanisms of complement evasion: learning from staphylococci and meningococci. *Nature reviews.Microbiology*, **8**, 393-399.

Shaughnessy J, Ram S, Bhattacharjee A, et al (2011) Molecular characterization of the interaction between sialylated Neisseria gonorrhoeae and factor H. *The Journal of biological chemistry*, **286**, 22235-22242.

Sironi M, Cagliani R, Forni D, Clerici M (2015) Evolutionary insights into host-pathogen interactions from mammalian sequence data. *Nature reviews.Genetics*, **16**, 224-236.

Smith NH, Maynard Smith J, Spratt BG (1995) Sequence evolution of the porB gene of Neisseria gonorrhoeae and Neisseria meningitidis: evidence of positive Darwinian selection. *Molecular biology and evolution*, **12**, 363-370.

Sonnen AF, Henneke P (2014) Structural biology of the membrane attack complex. *Sub-cellular biochemistry*, **80**, 83-116.

Spitzer D, Mitchell LM, Atkinson JP, Hourcade DE (2007) Properdin can initiate complement activation by binding specific target surfaces and providing a platform for de novo convertase assembly. *Journal of immunology (Baltimore, Md.: 1950)*, **179**, 2600-2608.

Tina KG, Bhadra R, Srinivasan N (2007) PIC: Protein Interactions Calculator. *Nucleic acids research*, **35**, W473-6.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research*, **19**, 327-335.

Weiser JN, Pan N, McGowan KL, Musher D, Martin A, Richards J (1998) Phosphorylcholine on the lipopolysaccharide of Haemophilus influenzae contributes to persistence in the respiratory tract and sensitivity to serum killing mediated by C-reactive protein. *The Journal of experimental medicine*, **187**, 631-640.

Wernersson R, Pedersen AG (2003) RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic acids research*, **31**, 3537-3539.

Wickham SE, Hotze EM, Farrand AJ, et al (2011) Mapping the intermedilysin-human CD59 receptor interface reveals a deep correspondence with the binding site on CD59 for complement binding proteins C8alpha and C9. *The Journal of biological chemistry*, **286**, 20952-20962.

Willard L, Ranjan A, Zhang H, et al (2003) VADAR: a web server for quantitative evaluation of protein structure quality. *Nucleic acids research*, **31**, 3316-3319.

Wilson DJ, Hernandez RD, Andolfatto P, Przeworski M (2011) A population genetics-phylogenetics approach to inferring natural selection in coding sequences. *PLoS genetics*, **7**, e1002395.

Wilson DJ, McVean G (2006) Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics*, **172**, 1411-1425.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, **24**, 1586-1591.

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences : CABIOS*, **13**, 555-556.

Yang Z, Wong WS, Nielsen R (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular biology and evolution*, **22**, 1107-1118.

3.2.2 Adaptation of genes not involved in immune response

3.2.2.1 Filovirus glycoproteins and their cellular receptor (NPC1) evolve under mutual selective pressure

The Filoviridae family (order Mononegavirales) includes the *Ebolavirus* and *Marburgvirus* genera, which cause highly fatal hemorrhagic fevers in humans. Recurrent outbreaks of filovirus are likely to have a zoonotic origin with subsequent human-to-human transmission. Uncertainty still exists about the reservoir host(s) for filoviruses, and it is presently unknown whether the spillover is initiated from a direct contact with the reservoir or rather from exposure to other wildlife that also contracted the infection from the reservoir. Moreover, whereas bats are generally asymptomatic carriers of Ebolaviruses and Marburgviruses, great apes, as well as humans, represent dead-end hosts due to severe pathology.

Filovirus infection is mediated by the surface-exposed trimeric glycoprotein (GP) and its engagement of a cellular receptor, the Niemann-Pick C1 (NPC1) protein, a housekeeping gene. NPC1 mainly resides in endosomes and lysosomes where it functions as a cholesterol transporter.

Interactions between viral molecules that mediate infection and host receptors are expected to develop into genetic arms-races, as the interacting partners are under mutual selective pressure within binding-seeking/binding-avoiding dynamics (rev. in [12]). This situation may favor amino acid replacements over silent substitutions, a situation referred to as positive (or diversifying) selection. Housekeeping genes have another characteristic that makes them appealing to viruses: they are highly conserved because most amino acid replacements cause a substantial reduction in fitness. From the viral perspective, this means that the sequence space available to the host for adaptive change is limited.

Before the molecular details of the interaction between filovirus glycoprotein

(GP) and its cellular receptor (NPC1) were known, Al-Daghri and co-workers [164] analyzed a limited set of mammalian sequences and they detected three positively selected codons in NPC1; it was suggested that these sites had evolved in response to filovirus-driven selective pressure. The recently solved structure of the GP-NPC1 complex [165] indicated that indeed these three sites are located at the interaction surface, demonstrating the predictive power of evolutionary inference in the field of host-pathogen interaction. Recently, Ng and co-workers investigated the evolution of *NPC1* in bats and detected positive selection in the second GP1-binding loop [166]. Changes at one of these sites (codon 502) indeed modulate binding to GP1 depending on the status of a single residue in the viral protein (position 141). Based on these findings, the authors hypothesized that, as expected under an arms-race scenario, codon 141 or other codons in *GP1* may be targeted by positive selection, as well [166].
In this work I investigated the evolutionary history of NPC1 in a large set of mammalian species and I apply a population genetics-phylogenetic approach to detect selection in humans and great apes. I then analyzed the adaptive events in *GP* during filovirus speciation. I also used state-of-the art molecular evolution approaches to analyze the interaction of filovirus GP with NPC1.

I integrated information on both positively and negatively selected sites to show that evolution of *NPC1* genes may be severely constrained by the necessity of maintaining the cellular function. Nevertheless, results indicate that a minority of residues in mammalian NPC1 proteins evolved adaptively, with most selected sites located within or at the anchor points of two protruding loops that directly interact with filovirus GP trimers. More specifically, the analysis of NPC1 indicated that positive selection drove the evolution of the receptor in all mammalian orders/superorders, from

Primates to Xenarthra, suggesting that most mammals are or were infected by filoviruses or related pathogens. In fact, the NPC1 domain that interacts with GP represented a preferential selection target (empirical *p* value= 0.0019). This is in line with the established long-lasting interaction between these pathogens and their mammalian hosts [164, 166-169]. On the other side of the race, filovirus *GP* was a target of positive selection, with one selected site directly contacting NPC1 and some other residues located within epitopes for neutralizing antibodies. I detect no evidence of selection at *GP* codon 141 but, rather, I observed such selection to involve the flanking 142 site. Site 142 establishes direct contacts with NPC1, indicating that it may represent a determinant of species tropism and virulence in specific hosts. Also, these data demonstrate that co-evolution, possibly in the form of epistasis, acted on GP in Ebolaviruses and in Marburgviruses. In both genera, one site in a co-evolving pair of residues is located within an antibody epitope, indicating that the underlying pressure is represented by the host immune system. Together with the identification of positively selected residues in Ebolavirus GP that localize to antibody epitopes, these results point to the host humoral immune response as a major selective pressure during filovirus speciation. Interestingly, residue 484, which we detected as positively selected, falls within a 9-amino acid long linear epitope recognized by the protective 14G7 antibody specific for EBOV [170]. Previous investigation of EBOV sequences from the recent West African outbreak identified 5 selected codons in the mucin domain, including sites 479 and 480 [171-174], within the 14G7 epitope. Evolutionary analysis of EBOV sequences from Sierra Leone also suggested that the T485A variant within the 14G7 epitope represents an escape mutant that originated during intra-host selection [173]. Thus, together with previous data, these findings point to the 14G7 epitope as a

major target of selection in Ebolaviruses.

In summary, although the recent analysis of NPC1 evolution in Chiroptera [166] was motivated by the well-known role of these mammals in MARV spill-overs and by their possible association with *Ebolavirus* transmission, these data indicate that bats are not the only mammals to be engaged in an arms-race with filoviruses. I suggest that the reservoir for these viruses may belong to any mammalian order. Also, because one of the GP selected sites (S142Q) establishes several atom-to-atom contacts with NPC1-C, I suggest that it modulates binding to NPC1 in different species and contributes to determine *Ebolavirus* host range in the wild. If this were the case, EBOV/BDBV (142S) and SUDV (142Q) may not share the same reservoir(s).

Moreover, the identification of selected sites and co-evolving sites that map to antigenic determinants for neutralizing antibodies will be extremely valuable for future efforts at developing effective treatments for filovirus hemorrhagic fever, as most therapies are currently based on antibody combinations [175].

Personal contribution to the work: I designed the study with my co-workers, I performed evolutionary analyses and I analyzed data. I also produced figures and tables and I contributed to write the manuscript.

Positive selection drives evolution at the host-filovirus interaction surface

Chiara Pontremoli[1], Diego Forni[1], Rachele Cagliani[1], Giulia Filippi[2], Luca De Gioia[2], Uberto Pozzoli[1], Mario Clerici[3,4], Manuela Sironi[1]

[1] Scientific Institute IRCCS E.MEDEA, Bioinformatics, 23842 Bosisio Parini, Italy;
[2] Department of Biotechnology and Biosciences, University of Milan-Bicocca, 20126 Milan, Italy;
[3]Department of Physiopathology and Transplantation, University of Milan, 20090 Milan, Italy;
[4]Don C. Gnocchi Foundation ONLUS, IRCCS, 20148 Milan, Italy.

Corresponding author: Manuela Sironi, Bioinformatics - Scientific Institute IRCCS E.MEDEA, 23842 Bosisio Parini, Italy. Tel: +39-031877826; Fax:+39-031877499; e-mail: manuela.sironi@bp.lnf.it

Abstract

Filovirus infection is mediated by engagement of the surface-exposed glycoprotein (GP) by its cellular receptor, NPC1 (Niemann-Pick C1). Two loops in the C domain of NPC1 (NPC1-C) bind filovirus GP. Herein we show that filovirus GP and NPC1-C evolve under mutual selective pressure. Analysis of a large mammalian phylogeny indicated that strong functional/structural constraints limit the NPC1 sequence space available for adaptive change and most sites at the contact interface with GP are under negative selection. These constraints notwithstanding, we detected positive selection at NPC1-C in all mammalian orders, from Primates to Xenarthra. Different codons evolved adaptively in distinct mammals, and most selected sites are located within the two NPC1-C loops that engage GP, or at their anchor points. In Homininae, NPC1-C was a preferential selection target, and the T419I variant possibly represents a human-specific adaptation to filovirus infection. On the other side of the arms-race, GP evolved adaptively during filovirus speciation. One of the selected sites (S142Q) establishes several atom-to-atom contacts with NPC1-C. Additional selected sites are located within epitopes recognized by neutralizing antibodies, including the 14G7 epitope, where sites selected during the recent EBOV epidemic also map. Finally, pairs of co-evolving sites in *Marburgviruses* and *Ebolaviruses* were found to involve antigenic determinants. These findings suggest that the host humoral immune response was a major selective pressure during filovirus speciation. The S142Q variant may contribute to determine *Ebolavirus* host range in the wild. If this were the case, EBOV/ BDBV (S142) and SUDV (Q142) may not share the same reservoir(s).

Introduction

Filoviruses are negative-sense, single-stranded RNA viruses; the *Filoviridae* family (order *Mononegavirales*) includes the *Ebolavirus* and *Marburgvirus* genera, which cause highly fatal hemorrhagic fevers in humans. The *Ebolavirus* genus consists of five recognized species: *Tai Forest ebolavirus* (TAFV), *Reston ebolavirus* (RESTV), *Sudan ebolavirus* (SUDV), *Zaire ebolavirus* (EBOV), and *Bundibugyo ebolavirus* (BDBV) (Kuhn et al. 2010). The fatality rate of at least 3 of these viruses (BDBV, EBOV, and SUDV) is extremely elevated, ranging from~38% for BDBV to ~79% for EBOV (de La Vega et al. 2015). TAFV has only been associated with a single, non-fatal human infection, whereas RESTV is thought to be non-pathogenic for humans (de La Vega et al. 2015). The *Marburg* genus only comprises one species (*Marburg marburgvirus*) which consists of two major lineages, Marburg virus (MARV, also referred to as Lake Victoria Marburg Complex) and Ravn virus (RAVV)

(Kuhn et al. 2010); both these viruses infect humans with variable rates of mortality (Brauburger et al. 2012). A third genus in the *Filoviridae* family, *Cuevavirus*, only includes one species (Lloviu virus, LLOV) isolated from bats in Northern Spain(Negredo et al. 2011).

Recurrent outbreaks of filovirus hemorrhagic fever have been described since 1976 (Messaoudi et al. 2015), the largest one being the recently contained epidemic in West Africa (http://apps.who.int/ebola/current-situation/ebola-situation-report-17-february-2016, last accessed February 2016). Outbreaks are likely to have a zoonotic origin with subsequent human-to-human transmission. Uncertainty still exists about the reservoir host(s) for filoviruses, and it is presently unknown whether the spillover is initiated from a direct contact with the reservoir or rather from exposure to other wildlife that also contracted the infection from the reservoir (Mari Saez et al. 2014; Olival and Hayman 2014). Field surveys have detected genome fragments or antibodies against EBOV and RESTV in different bat species from Africa and Asia (Olival and Hayman 2014; de La Vega et al. 2015). Nevertheless, infectious *Ebolaviruses* have never been isolated from bats (Olival and Hayman 2014; de La Vega et al. 2015). As for MARV/RAVV, live viruses were isolated only once from *Rousettus aegypticus* bats in Kitaka cave, Uganda (Towner et al. 2009). These observations suggest that whereas bats are generally asymptomatic carriers of *Ebolaviruses* and *Marburgviruses*, great apes, as well as humans, represent dead-end hosts due to severe pathology. Indeed, *Ebolaviruses* represent a continuing threat to humans as well as to the survival of gorilla and chimpanzee populations in Central Africa (Walsh et al. 2003).

Filovirus infection is mediated by the surface-exposed trimeric glycoprotein (GP), which is synthesized as a single peptide and subsequently cleaved by furin into a receptor-binding subunit (GP1) and a fusion subunit (GP2). The two subunits of each monomer remain linked through a disulfide bond and several non-covalent interactions (Jeffers et al. 2002; Lee et al. 2008). After cellular attachment (mediated by GP1) and endocytosis, the virus is trafficked to the late endosomes where the GP trimer is primed by cysteine proteases (Chandran et al. 2005; Schornberg et al. 2006; Kaletsky et al. 2007). Priming results in exposure of the receptor binding domain (RBD) and the consequent engagement of a cellular receptor, the Niemann-Pick C1 (NPC1) protein(Miller et al. 2012; Moller-Tank and Maury 2015). A multi-spanning membrane protein, NPC1 mainly resides in endosomes and lysosomes where it functions as a cholesterol transporter. Mutations in *NPC1* are responsible for a rare and fatal lipid storage disorder, Niemann-Pick disease type C (Garver et al. 2010; Peake and Vance 2010; Garver 2011). NPC1 is expressed ubiquitously and serves as an indispensable host entry factor for all known filoviruses (White and Schornberg 2012). Interaction with these pathogens involves the second luminal domain of NPC1 (the so called Domain C, NPC1-C), which is bound directly and specifically by GP; very recent crystallographic data indicated that two protruding loops in NPC1-C bind the RBD of *Ebolavirus* GP1 (Wang et al. 2016).

Interactions between viral molecules that mediate infection and host receptors are expected to develop into genetic arms races, as the interacting partners are under mutual selective pressure and cyclical adaptation and counter-adaptation occur (Sironi et al. 2015). This situation may favor amino acid replacements over silent substitutions, a situation referred to as positive (or diversifying) selection (Sironi et al. 2015). Positive selection is expected to be strong at codons corresponding to residues that form the physical viral-host interaction surfaces (Sironi et al. 2015). In line with this view, we have previously analyzed a limited number of mammalian species and observed that three codons in *NPC1-C* evolved under positive selection (Al-Daghri et al. 2012). The three codons are located in close proximity to or within one of the two loops (loop1) that engage *Ebolavirus* GP. More recently, Ng and coworkers investigated the evolution of *NPC1* in bats and detected positive selection in the second GP1-binding loop (loop2) (Ng et al. 2015). Changes at one of these sites (codon 502) indeed modulate binding to GP1 depending on the status of a single residue in the viral protein (position 141). Based on these findings, the authors hypothesized that,

as expected under an arms-race scenario, codon 141 or other codons in *GP1* may be targeted by positive selection, as well.

Herein we investigated the evolutionary history of *NPC1* in a large set of mammalian species and we show that, despite an overall strong constraint acting on the protein, positive selection drove the evolution of NPC1-C by targeting different residues in distinct mammalian orders. On the other side of the race, filovirus GP evolved adaptive changes at the RBD, not at site 141 but at the flanking 142 position. Finally, we show that the host humoral immune response exerted a major pressure during filovirus speciation.

Results

The little sequence space available for adaptive change in mammalian NPC1 is likely deployed to respond to filovirus-driven selection

Filoviruses have represented a selective pressure for several mammalian species, as witnessed by the detection of filovirus-derived endogenous viral elements in the genomes of bats, rodents, primates, and marsupials (Taylor et al. 2010; Taylor et al. 2011; Ng et al. 2015). Previous analyses identified different selected sites in NPC1: whereas we detected selection at NPC1-C loop 1 in a set of 41 mammalian sequences, Ng et al only analyzed bats and mainly detected selection at loop 2 (Al-Daghri et al. 2012; Ng et al. 2015). Although these differences may simply reflect variable power to detect selection, they may also suggest that different sites were target by selection in distinct mammals. To gain a thorough view of *NPC1* evolution and to assess the extent of filovirus-driven selection, we gathered a list of 80 *NPC1* mammalian sequences (fig. S1, Supplementary Material online). We first estimated the extent of functional constraint acting on NPC1 proteins by identification of sites under negative selection (total number = 897) using the Single-Likelihood Ancestor Counting (SLAC) method (Kosakovsky Pond and Frost 2005). SLAC estimates the probability of selection at each site in an alignment through the dN-dS metric (rate of nonsynonymous changes-rate of synonymous changes) (Kosakovsky Pond and Frost 2005). This is because the conventional dN/dS ratio is rendered to infinite for dS values equal to 0. The observation that a large proportion (71%) of sites shows evidence of negative selection indicates that most amino acid replacements in NPC1 are deleterious. In line with this view, we observed that codons that carry at least one missense mutation responsible for Niemann-Pick disease type C (n =172) display significantly lower dN-dS compared to sites where no missense mutation has been described (Wilcoxon rank sum test, $p= 4.2x10^{-7}$) (fig. 1A). These data suggest that adaptive evolution at NPC1 is restricted by functional/structural constraints to maintain its cellular role as a sterol transporter. This also applies to residues in NPC1-C loops 1 and 2: most sites at the contact interface with GP are under negative selection (7/8 sites in loop1 and 8/13 in loop2, fig. 1). Consistently, previous analyses of *NPC1* identified a very small proportion of positively selected sites (Al-Daghri et al. 2012; Ng et al. 2015).We next explored the presence and extent of positive selection in different mammalian orders, superorders or clades: Primates plus Scadentia, Laurasiatheria, Glires, and Xenarthra plus Afrotheria. Screening of the alignments for the presence of recombination identified only one breakpoint in the Laurasiatheria alignment at position 507; thus, the initial 169 codons were removed for evolutionary analysis.

Evidence of positive selection was searched for using the sites models (i.e. these models allow dN/dS to vary among sites in the alignment) implemented in the *codeml* program (Yang 1997; Yang 2007). Two neutral models (M8a and M7) were rejected in favor of the M8 positive selection model in all analyses, with the only exclusion of Primates/Scadentia (Table 1). Conversely, the M1a neutral model was never rejected in favor of the M2a positive selection model (all *p* values >0.05).Nonetheless, simulation experiments have shown that the M1a/M2a test has low power to detect selection even when positively selected sites truly exist in the data. This is particularly true when relatively few taxa are analyzed and/or if the proportion of positively selected sites is small (Wong et al. 2004).

We thus identified positively selected sites through Bayes Empirical Bayes (BEB)

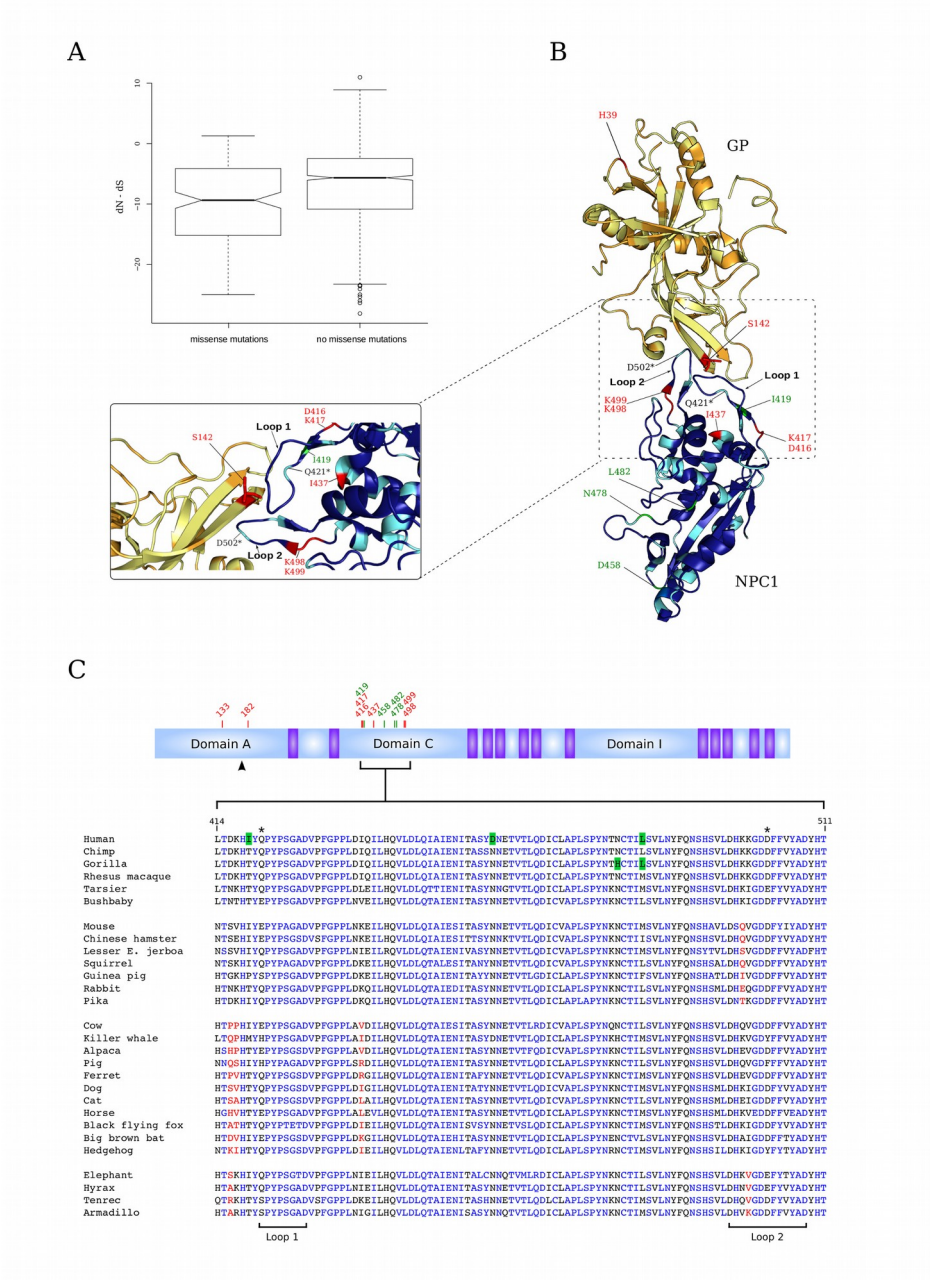


Figure 1. **Natural selection at mammalian *NPC1*. (A)** Boxplot representation of dN-dS (see methods for details). Mammalian *NPC1* codons carrying missense mutations responsible for Niemann-Pick disease type C are compared with sites where no missense mutation has been described. **(B)** 3D Structure of the GP-NPC1-C complex (PDB ID: 5F1B). *Ebolavirus* GP is colored in yellow, with sites under negative selection showed with a darker hue. Human NPC1 is showed in cyan with sites under negative selection in dark blue. Positively selected sites in the mammalian phylogeny and in *Ebolavirus* are shown in red, in Homininae are shown in green. Sites denoted with asterisks were previously detected as positively selected (see text). **(C)** Schematic representation of NPC1 domains. The three large luminal domains are indicated as A, C, and I; trans-membrane domains are in purple. Positively selected sites are reported on the NPC1 structure and in the multiple alignment; color and symbol codes are as in panel (B). Few representative mammalian species are shown in the alignment, and sites under negative selection as detected by SLAC are in blue. The black head arrow indicates the position (amino acid 169) of the recombination breakpoint identified in Laurasiatheria. Positions refer to the human sequence (NP_000262).

analysis (Anisimova et al. 2002; Yang et al. 2005) from model M8 (with a posterior probability cut-off of 0.95) and most of them were located in NPC1-C (fig. 1). More precisely, although distinct codons were targeted by selection in different mammalian orders/superorders or clades, most selected sites (5 out of 7) are within or at the base of either loop1 or loop2 (fig. 1, table 1). Because positive selection at NPC1 has previously been demonstrated in Chiroptera (Ng et al. 2015), we repeated the analysis after exclusion of bat sequences from the Laurasiatheria set: the likelihood ratio tests remained significant and two of the three selected sites in NPC1-C were still detected (table 1).

We identified a total of 7 positively selected sites in NPC1, five of them were located in domain C. To test whether this number is higher than expected, we performed random sampling across NPC1 sites that are not under negative selection (i.e. we assumed that all sites that are not under negative selection have the same probability of being called as positively selected). Results indicated that the likelihood of having 5 selected sites in NPC1-C amounts to 0.0019; thus, this domain represents a preferential selection target in mammals.

Adaptive evolution in humans and great apes acts on NPC1-C

The failure to detect positively selected sites in primate *NPC1* genes may derive from the relatively close relatedness of these species, resulting in reduced power. This limitation can nevertheless be bypassed by the availability of extensive genetic diversity data for humans and great apes (1000 Genomes Project Consortium et al. 2010; Prado-Martinez et al. 2013). This allows the application of population genetics-phylogenetics approaches to analyze *NPC1* evolution in the human, chimpanzee, and gorilla lineages. Thus, we used gammaMap (Wilson et al. 2011), a program that jointly uses intra-species variation and inter-specific diversity, to estimate population-scaled selection coefficients (γ) along *NPC1* and to identify positively selected sites. As previously suggested based on the distribution of γ values (Barreiro et al. 2009), we conservatively declared positively selected sites as those showing a posterior probability ≥ 0.75 of γ ≥ 1. Whereas no positively selected codons were identified in chimpanzees, three and two sites were detected in humans and gorillas, respectively (table S1, Supplementary Material online). All selected sites were located in NPC1-C (fig. 1) and the human selected site I419 is located at the base of loop1 (fig. 1). Site 482 was found to be selected both in the gorilla and in the human lineage (ancestral: methionine, human and gorilla: leucine); the same M482L change is observed in chimpanzee (although the posterior probability in this species did not reach the significance cut-off we set), suggesting that selection ensued in the common ancestor of humans and great apes.

Thus, a total of 4 independent selection events were detected and all sites were located in NPC1-C (fig. 1). Again, this finding is unlikely to be due to chance (random sampling, p= 0.0007), indicating that domain C represents a preferential target of positive selection in Homininae, as well.

***Ebolavirus* GP proteins evolve in response to host-driven selection**

We next investigated whether positive selection also drove the evolution of *Ebolavirus* GP proteins. Indeed, we expect viral GP to be under selective pressure both to optimize NPC1 binding and to elude the host immune system. In fact, GP represents a major target for antibody response as it is the only protein exposed on the virus surface (Murin et al. 2014).

The genus *Ebolavirus* includes 5 distinct species. To investigate the evolution of *GP* during speciation, we obtained sequence information for 20 EBOV, 6 BDV, 2 TAFV, 11 SUDV, and 9 RESTV (table S2, Supplementary Material online). Within each species, sequences were selected to represent viruses sampled during distinct outbreaks. In the case of the 2014 EBOV epidemic, isolates belonging to the 5 major lineages from Guinea, Mali, and Sierra Leone were included (Simon-Loriere et al. 2015). The sequence alignment was pruned of unreliably aligned codons (see Material and Methods), a procedure that resulted in the masking of a large portion of the mucin domain. Indeed, this region displays very little homology among *Ebolaviruses.* The pylogenetic tree of *GP* obtained with phyML

Table 1. Evolutionary Analysis of Mammalian *NPC1* Genes.

	N° of species	Tree length[a]	M8a vs. M8		M7 vs. M8		Positively selected
			$-2\Delta lnL^b$	*P* value	$-2\Delta lnL^b$	*P* value	Sites [c]
Primates plus Scadentia	**23**	**1.479**	**0.273**	**0.602**	**3.098**	**0.213**	**–**
Glires	**16**	**3.665**	**4.504**	**0.0338**	**10.240**	**0.00598**	**V133, K498**
Laurasiatheria (amino acids 170-1277)	**30**	**4.409**	**14.185**	$\mathbf{1.657 \times 10^{-4}}$	**49.139**	$\mathbf{2.14 \times 10^{-11}}$	**D182, D416[d], K417, I437[d]**
Xenarthra plus Afrotheria	**8**	**1.937**	**5.575**	**0.0182**	**16.925**	$\mathbf{2.11 \times 10^{-4}}$	**D416, K499**

[a]Tree length is defined as number of nucleotide substitutions per codon.
[b]2ΔlnL: twice the difference of the natural logs of the maximum likelihood of the models being compared.
[c]Aminoacid positions refer to the human sequence (NP_000262).
[d]Positively selected sites in Laurasiatheria with the exclusion of Chiroptera sequences.

was fully consistent with the previously reported Bayesian phylogeny (Gire et al. 2014).

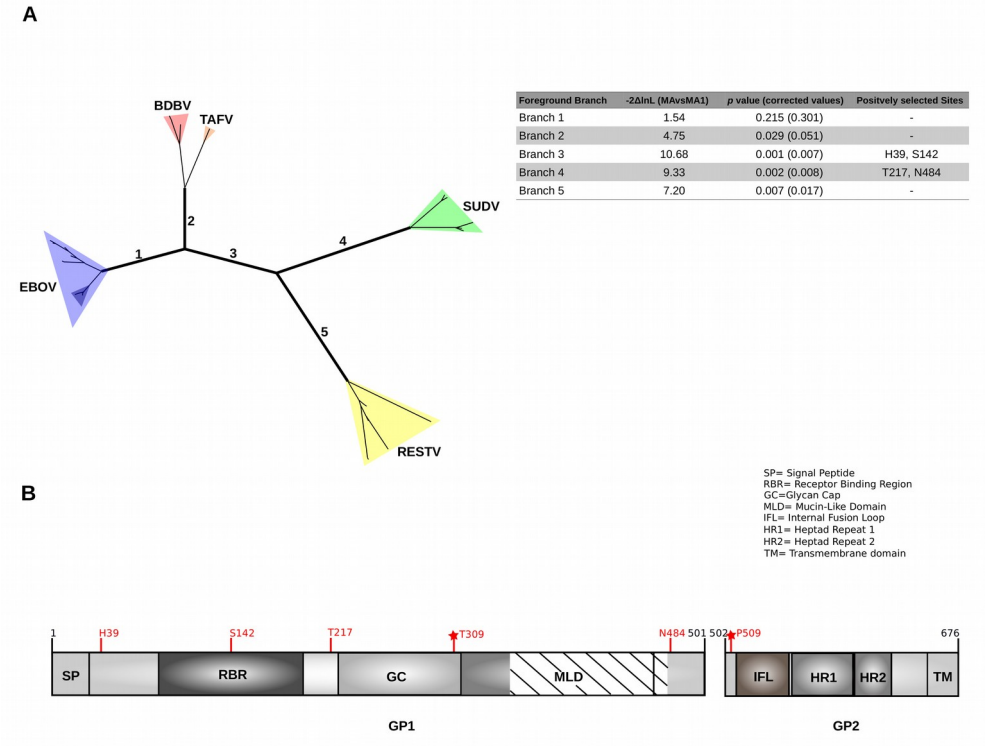
A search for codons under negative selection detected 175 sites, most of them located in the heptad repeats and in the fusion loop (fig. S2, Supplementary Material online).

The total tree length for *GP* amounted to 7.22, indicating that positive selection can be inferred with high confidence (but low power) using the *codeml* branch site tests, which we applied to test the internal branches of the phylogeny (fig. 2A). Statistically significant evidence of episodic positive selection was obtained for 3 branches (fig. 2A). Selected sites along these branches were identified using the BEB procedure from model MA (with a significant posterior probability cut-off of 0.95). A total of 4 selected sites were detected, all of them located in GP1 (fig. 2B). Among these, residue 142 is in the RBD, at the direct contact interface with NPC1-C (fig. 1B). Specifically, the S142 residue in EBOV GP establishes several atom-to-atom contacts with NPC1-C (Wang et al. 2016). Site 39 (threonine in SUDV, histidine in EBOV) is located at the base of the GP trimer, within an epitope bound by the 16F6 antibody that specifically neutralizes SUDV (fig. 3A) (Bale et al. 2012). Likewise, residue 484 falls within a 9-amino acid long linear epitope recognized by a protective antibody (14G7) specific for EBOV (not reactive against SUDV and TAFV) (fig. 3B) (Wilson et al. 2000). The fourth selected residue at position 217 is within the so-called “head” domain of GP1. A conservative isoleucine to alanine substitution at the flanking 218 site results in reduced infectivity possibly due to alteration of GP structure and decreased incorporation into virions (Manicassamy et al. 2005). Because epistasis is common in viruses and is thought to play an important role in the evolution of immune evasion and host shifts (Bedhomme et al. 2015), we searched for evidence of co-evolution between GP sites. To this aim, we applied BGM-Spidermonkey (Poon et al. 2007) and MISTIC (Simonetti et al. 2013) (see Material and Methods for details). One pair of co-evolving sites (309 and 509) was detected by both methods with high confidence (fig. 3C). Residue 509 (in GP2) is at the direct contact interface with the KZ52 antibody isolated from a human survivor of EBOV infection (fig. 3C) (Lee et al. 2008), whereas residue 309 is located in the glycan cap. This region is also bound by neutralizing antibodies, but the precise epitopes are not fully resolved (Murin et al. 2014). Interestingly, residue 509 (G509 in SUDV and P509 in EBOV) represents the anchor point of the N-terminal portion of GP2 to the GP core; the conformation of the GP2 N-terminus differs between SUDV and EBOV and is expected to result in different mobility of this portion (Bale et al. 2012). Finally, it is worth noting that mutations at the flanking 508 residue affect the binding of three different antibodies against EBOV GP, and Q508R escape mutants are associated with lethality in monkeys treated with a combination of three neutralizing mAbs (ZMAb) (Qiu et al. 2013). Site 508 showed evidence of positive selection on the branch separating SUDV from EBOV/TAFV/BDBV, although the posterior probability (0.94) did not reach the threshold we set for significance.

Evolution of *Marburgvirus GP*

Finally, we investigated the evolution of *GP* in *Marburg marburgvirus*. The analysis included representative sequences for the MARV and RAVV lineages, selected on the basis of their association with distinct outbreaks or isolation years (table S3,Supplementary Material

Figure 2



Foreground Branch	-2ΔlnL (MAvsMA1)	*p* value (corrected values)	Positvely selected Sites
Branch 1	1.54	0.215 (0.301)	-
Branch 2	4.75	0.029 (0.051)	-
Branch 3	10.68	0.001 (0.007)	H39, S142
Branch 4	9.33	0.002 (0.008)	T217, N484
Branch 5	7.20	0.007 (0.017)	-

Figure 2. **Branch-site analysis of positive selection for *Ebolavirus GP*. (A)** Analyses of episodic positive selection are shown in the box and selected sites are reported. -2ΔlnL is twice the difference of the natural logs of the maximum likelihood of the models being compared (see text for details). Numbers in the tree identify the corresponding tested branches. *Ebolavirus* species are highlighted by triangles. The darker blue triangle in the EBOV phylogeny denotes isolates from 2014 West Africa epidemic. **(B)** Mapping of co-evolving and positively selected sites onto the *Zaire ebolavirus* glycoprotein structure. The dashed portion of the structure corresponds to the alignment region masked by GUIDANCE. Positively s elected sites are in red, co-evolving sites are denoted with a star.

online). The *GP* gene tree was consistent with previously reported phylogenies based on whole genome sequences (Carroll et al. 2013).
MARV and RAVV belong to the same species and the *GP* phylogeny is much shallower (tree length=1.75) compared to that of *Ebolaviruses*. Therefore, we tested for positive selection using both the *codeml* site models (M7 vs M8 and M8a vs M8) and the branch-site models (MA1 vs MA). These latter models were used to test selection on the internal branch of the phylogeny. Evidence of positive selection was obtained using all tests (fig. 4A), although no positively selected sites were detected. Co-evolution analysis identified one pair of sites (fig. 4B): the 267 residue, co-evolving with codon 279, is located within the epitope for the 7G8 protective antibody (fig. 4C)(Hevey et al. 2003).

Discussion

In this study we show that mammalian NPC1 proteins evolved adaptively, with most selected sites located at the base of two protruding loops that directly interact with filovirus GP trimers. In turn, filovirus GP was a target of positive selection, with one selected site directly contacting NPC1 and some other residues located within epitopes for neutralizing antibodies. We detected positive selection at NPC1-C in all mammalian orders/superorders, from Primates to

Figure 3

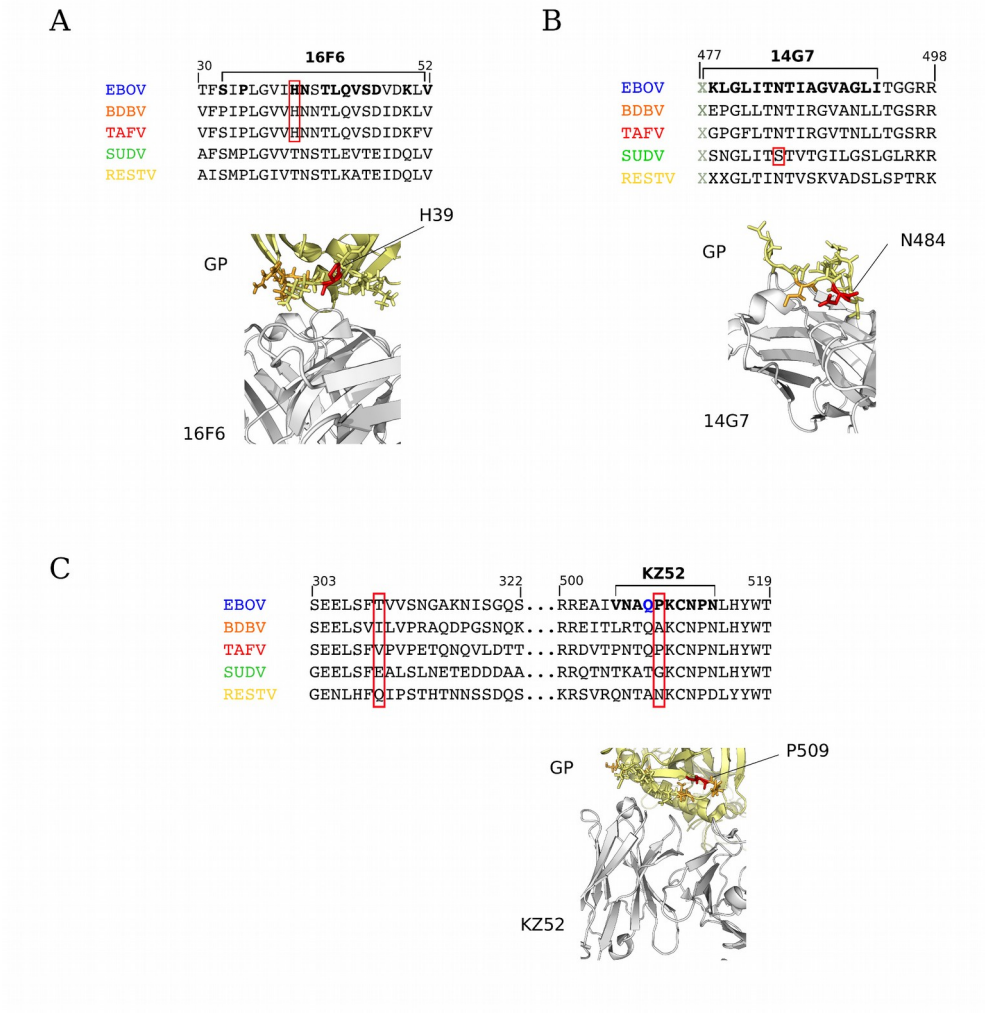


Figure 3. Selection at antibody binding interfaces. GP alignments for *Ebolavirus species* is shown along with the interaction surface with the protective 16F6 (PDB ID: 3VE0) **(A)** 14G7 (PDBID: 2Y6S) **(B)**, and KZ52 (PDB ID: 3CSY) **(C)** antibodies. In the alignments, positively selected sites are framed in red, GP residues involved in antibody binding are in bold. GP position 508 (see text) is in blue. In the structures, the epitope region is shown in stick representation with color codes as in Figure 1.

Xenarthra, suggesting that most mammals are or were infected by filoviruses or related pathogens. Together with previous results (Al-Daghri et al. 2012; Ng et al. 2015), these data strongly suggest that filoviruses and their mammalian hosts have been engaged in long-lasting genetic conflict. Indeed, based on the synteny of filovirus-derived endogenous viral elements, recent works indicated that these pathogens have infected bats and rodents for at least 25 MY

and 18 MY, respectively (Taylor et al. 2010; Taylor et al. 2011; Taylor et al. 2014).

The use of housekeeping proteins as cellular receptors is a common strategy employed by viruses to infect the host (Sironi et al. 2015). From a viral standpoint, housekeeping genes have the advantage of being expressed at high levels in multiple cell types. A notorious example of a housekeeping protein functioning as a viral entry gate is the transferrin receptor (TfR), which is also a target of positive selection in mammals (Kaelber et al. 2012; Demogines et al. 2013). TfR is exploited by some New World arenaviruses, parvoviruses, and retroviruses. These pathogens interact with the same domain of TfR and the binding surfaces are contiguous with small overlaps (Demogines et al. 2013). By analogy, we cannot exclude that the selected sites we detected in NPC1-C evolved in response to pathogens other than filoviruses, as several enveloped viruses are trafficked to endolysosomal compartments after endocytosis (Jae and Brummelkamp 2015). Nonetheless, current observations indicate that whereas all extant filoviruses are dependent on NPC1 for infection, there is no evidence that other viruses use this receptor. Housekeeping genes have another characteristic that makes them appealing to viruses: they are highly conserved because most amino acid replacements cause a substantial reduction in fitness. From the viral perspective, this means that the sequence space available to the host for adaptive change is limited. In NPC1, the majority of sites at the contact interface with GP are under negative selection and cannot therefore evolve to avoid viral binding. Indeed, replacements at highly conserved positions are likely to result in Niemann-Pick disease, as the analysis of reported mutations demonstrates. Only few sites within loop 1 and loop2 are not targeted by purifying selection. Among these, residue 502 was previously detected as a selection target in bats (Ng et al. 2015), whereas we had previously described codon 421 as positively selected in mammals (Al-Daghri et al. 2012). This latter site was also detected in this study, but it did not reach the threshold for statistical significance we set (Xenarthra/Afrotheria set; BEB, posterior probability = 0.902). Clearly, this does not exclude the possibility that it represents a selection target and that it modulates filovirus infection; indeed, residue 421 is located at the direct contact interface with GP (Wang et al. 2016). As for site 502, it did not show evidence of dN/dS >1 in our analyses, suggesting that it represents a selection target in bats only. This is consistent with the fact that distinct sites are selected in different mammalian orders.

Most of the selected sites we detected are located within the two NPC1-C loops that engage GP or at their anchor points. In the complex structure, none of them was reported to establish atom-to-atom contacts with GP. Nonetheless, loop structures are known to be highly flexible and mobile, suggesting that the selected residues might modulate loop conformation and, consequently, GP binding affinity. The I419 site, which is positively selected in the human lineage, immediately flanks the contact interface with GP, possibly representing a human-specific adaptation to filovirus infection. Additional sites targeted by positive selection in primates were found to be located in NPC1-C regions distant from the GP binding interface. Whether these variants modulate binding via long-range interactions or alteration of protein structure/stability remains to be evaluated. However, positive selection preferentially targeted NPC1-C in all mammalian orders. Enrichment of positively selected sites in this domain is unlikely to be accounted for by lower functional constraints; in fact, we performed random sampling allowing positively selected sites to only occur at sites that are not under negative selection. We thus suggest that the selective pressure acting on *NPC1* is mainly exerted by filoviruses or other pathogens that infect the host via endolysosomal trafficking.

In line with the arms race scenario, we also detected positively selected sites in *Ebolavirus* GP proteins. Notably, the branch-site test we used to detect positive selection during *Ebolavirus* speciation is robust to saturation issues and has a minimal false positive rate. Also, the test is largely insensitive to violations of the assumption of neutral evolution on the background branches (Zhang et al. 2005; Anisimova and Yang 2007). Nonetheless, for sequences as divergent as those from the five *Ebolavirus* species, the test lacks power (Gharib and Robinson-Rechavi 2013), indicating that several selected sites may have remained undetected.

One of the selected sites we identified (S/Q142) is located at the direct contact interface with NPC1-C and establishes several atom-to-atom contacts with the receptor (Wang et al. 2016). Ng and coworkers showed that a conservative change at the flanking 141 site (V141A) modulates the binding of GP to NPC1-C from Africa straw-colored fruit bats. The NPC1 molecule from these bats is a poor ligand for EBOV GP due to the positively selected D502F change (Ng et al. 2015). Because V141A is a naturally occurring variant in *Ebolaviruses*, the authors speculated that it may represent an adaptive change to broaden host tropism. Our results are consistent with this possibility, although we detect no evidence of selection at codon 141 but, rather, we observed such selection to involve the flanking 142 site. Interestingly, analysis of EBOV GP in complex with NPC1-C (which was not available when Ng and co-workers published their work) indicated that residue 141 in EBOV GP establishes no direct contact with the 502 residue in NPC1. Nevertheless, the V141A substitution restores GP binding to NPC1 molecules carrying the D502F substitution. This observation substantiates the view that changes within loop positions or in their flanks may affect loop structure and, consequently, binding properties. This is also in line with the view that loops are highly mobile and can exists as an ensemble of conformations (Shehu and Kavraki 2012), complicating the prediction of the functional effects of amino acid replacements. Experimental studies will be required to address the role of the S142Q variant in modulating host tropism or virulence in specific hosts. Clearly, this variant (as well as the V141A substitution) is unlikely to represent a major determinant of virulence in humans, as viruses causing a high fatality rate share the same amino acid with non-pathogenic species (i.e. RESTV). Nevertheless, the S142Q variant may contribute to determine *Ebolavirus* host range in wild animals (see also (Ng et al. 2015)). If this were the case, EBOV/BDBV (S142) and SUDV (Q142) may not share the same reservoir(s). Previous studies of positive selection in *Ebolaviruses* were based on intra-species analyses and consequently detected sites that are different from those reported herein (Li and Chen 2014; Azarian et al. 2015; Ladner et al. 2015; Park et al. 2015). Besides the fact that different evolutionary time frames are investigated compared to inter-species studies, intra-species analyses allow the assessment of protein regions that are poorly conserved among species, such as the mucin domain. Conversely, due to unreliable alignments, which in turn are known to originate false positive results, we filtered a large portion of codons in the mucin domain. This very domain was previously found to represent a preferential target of selection in EBOV (Walsh et al. 2005; Ladner et al. 2015; Park et al. 2015). For instance, an early study with relatively few EBOV sequences detected one single positively selected site (position 370) in the mucin domain (Walsh et al. 2005). More recently, investigation of EBOV sequences from the latest West African outbreak identified 5 selected codons in the mucin domain, including sites 479 and 480 (Ladner et al. 2015). As the 484 site we detected herein, these residues are within the epitope for the 14G7 antibody. Also, evolutionary analysis of EBOV sequences from Sierra Leone suggested that the T485A variant within the 14G7 epitope represents an escape mutant that originated during intra-host selection (Park et al. 2015). These results point to the 14G7 epitope as a major target of selection in *Ebolaviruses*. More generally, analysis of EBOV from Sierra Leone indicated an enrichment of nonsynonymous substitutions within B cell epitopes, and two out of four positively selected sites we detected in *Ebolavirus* GP are located within antibody epitopes (sites 39 and 484). Recent evidence indicates that epistasis in viral proteins may play an important role in the development of immune escape and drug resistance (Kryazhimskiy et al. 2011). This is particularly true for viral surface-exposed molecules. The co-evolving sites we detected in *Ebolavirus* GP are not located in spatial proximity in the folded protein structure.

This is not unexpected, though, as previous studies indicated that epistatic sites in the surface proteins of human influenza A virus (HAV) are not in physical proximity (Kryazhimskiy et al. 2011). This in confirmed also by experimental data showing that a HAV oseltamivir-resistant mutant carrying the H274Y substitution has decreased viral fitness, which can be restored by second-site mutations in residues that are in no close contact with

Figure 4

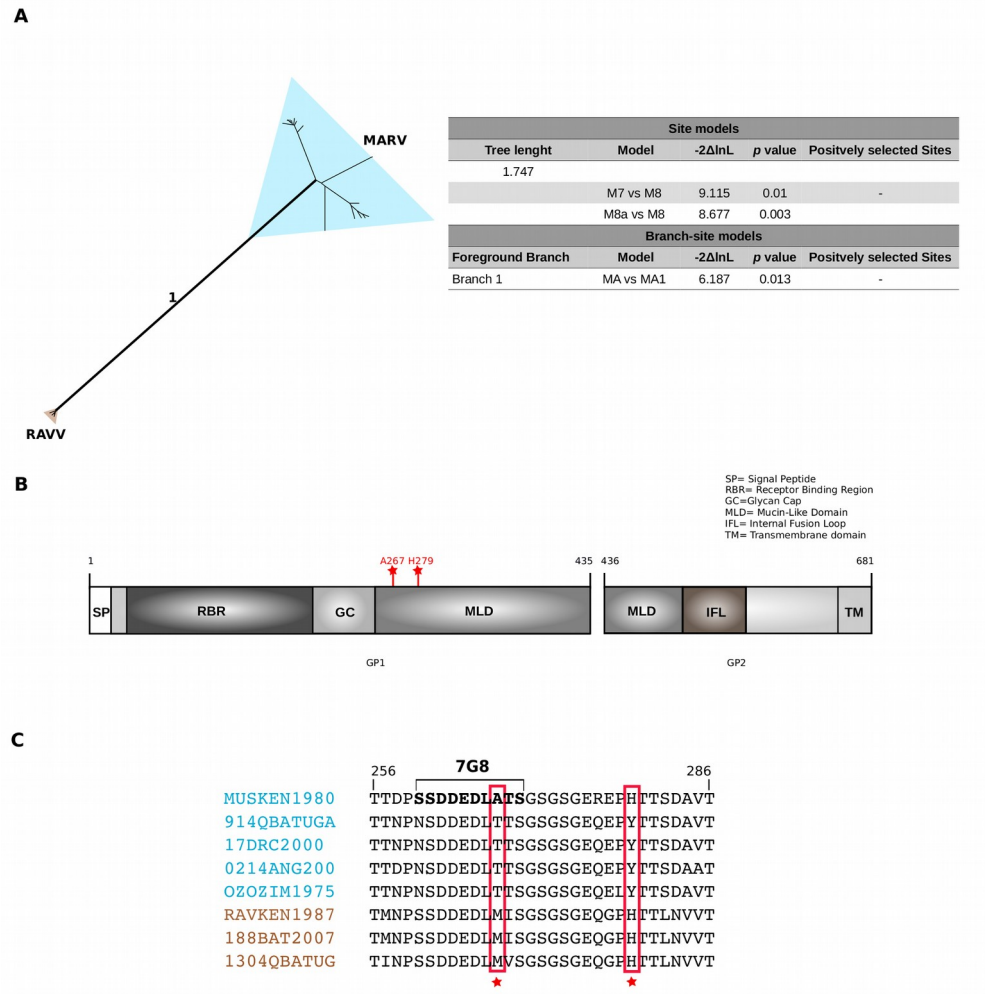


Figure 4. *Marburgvirus GP* adaptive evolution. (A) Phylogenetic tree of *GP* sequences from *Marburgviruses*. The tested branch is indicated on the phylogenetic tree by the corresponding number. - 2ΔlnL is twice the difference of the natural logs of the maximum likelihood of the models being compared (see methods for details). **(B)** Schematic representation of the domain structure of MARV glycoprotein. Co-evolving sites are shown in red with a star. **(C)** Alignment of the GP positions surrounding co-evolving sites for a few representative *Marburgvirus* strains; color codes are as in panel A; co-evolving sites are boxed in red and denoted with a star.

H274 in the neuraminidase structure (Bloom et al. 2010). In the context of HAV, epistatic sites were found to be common in hemagglutinin sites responsible for antigenic shifts (Neverov et al. 2015). Likewise, the fixation of destabilizing mutations in antigenic regions of the HAV nucleoprotein was found to be dependent on the acquisition of enabling substitutions. Again, nucleoprotein epistatic mutations involve residues in no physical contact, and their effect is mainly explained by a change in protein stability, with escape mutations determining a decrease in stability that is tolerated only in the presence of second-site stabilizing substitutions (Gong et al. 2013). For both *Ebolavirus* and *Marburgvirus* GP,

one site in a co-evolving pair of residues is located within an antibody epitope, suggesting that the underlying pressure driving their evolution is represented by the host immune system. Together with the identification of positive selected residues in *Ebolavirus* GP that localize to antibody epitopes, these results point to the host humoral immune response as a major selective pressure during filovirus speciation. Because the most promising treatment strategies for filovirus hemorrhagic fever are based on antibody combinations (Murin et al. 2014; Casadevall and Pirofski 2015), antigenic variability should pose a serious concern to their effectiveness in the long-term.

In summary, data herein indicate that positive selection has been driving the molecular evolution at the host-filovirus interaction surface, acting on the the cellular receptor NPC1 and on the viral glycoprotein. If the hypothesis of filovirus-driven positive selection at mammalian NPC1 is correct, our findings suggest that most mammals are or were infected by these pathogens; the filovirus reservoir(s) may thus belong to any mammalian order.

Materials and Methods

Sequences, alignments, and gene trees

Viral sequences were retrieved from the NCBI database and a list of accession numbers is provided as tables S2 and S3, Supplementary Material online. Available mammalian sequences for *NPC1* were also retrieved from the NCBI database (http://www.ncbi.nlm.nih.gov/, last accessed February 2016). A list of species is available as fig. S1, Supplementary Material online.

Alignment errors are common when divergent sequences are analyzed and can affect evolutionary inference. Thus, we used PRANK (Loytynoja and Goldman 2005) to generate multiple sequence alignments and GUIDANCE (Penn et al. 2010) for filtering unreliably aligned codons (we masked codons with a score <0.90), as suggested (Privman et al. 2012). Several codons in *Ebolavirus* GP mucin domain were filtered by GUIDANCE, whereas no codons were filtered in *Marburgvirus* and mammalian alignments.

All alignments were screened for the presence of recombination using GARD (Genetic Algorithm Recombination Detection) (Kosakovsky Pond et al. 2006), a Genetic Algorithm implemented in the HYPHY suite (Pond et al. 2005).

After running a codon model selection analysis in HYPHY, gene trees were generated using the program phyML with a maximum-likelihood approach, gamma-distributed rates, 4 substitution rate categories, and estimation of transition/transversion ratio and proportion of invariable sites (Guindon et al. 2010).

NPC1 mutations

The list of Niemann-Pick type C mutations was obtained from the Human Gene Mutation Database (HGMD, http://www.hgmd.cf.ac.uk/ac/, last accessed February 2016). From a total of 225 unique missense and nonsense mutations, we retained missense substitutions only (n = 200). These occurred in 172 unique codons.

Evolutionary analyses

The SLAC tool (Kosakovsky Pond and Frost 2005) from the HYPHY package was used to identify sites under negative selection and for calculating dN-dS (rate of nonsynonymous changes-rate of synonymous changes) at each site (Kosakovsky Pond and Frost 2005). Because SLAC is very conservative (Kosakovsky Pond and Frost 2005) a significance cut-off= 0.1 was used. An *NPC1* alignment with 80 species (including Eutheria, Metatheria and Monothremata, fig. S1, Supplementary Material online).

Evidence of positive selection was searched for using the codon-based *codeml* program implemented in the PAML (Phylogenetic Analysis by Maximum Likelihood) suite (Yang 2007). This tool analyzes gene alignments to evaluate the nonsynonymous/synonymous rate ratio (dN/dS, also referred as ω); positive selection can be defined when ω is higher than 1.

We applied different random site (NSsite) models with a F3x4 codon frequency model. M1a is a nearly neutral model that assumes one dN/dS (ω) class between 0 and 1, and one class with ω=1; M2a (positive selection model) is the same as M1a plus an extra class of ω >1. M7 is a null model that assumes that 0<ω<1

and is beta distributed among sites; M8 is a positive selection model: it is the same as M7 but also includes an additional category of sites in the alignment with ω>1; M8a is the same as M8, except that does not allow positive selection, but only neutral evolution. To assess statistical significance twice the difference of the likelihood (ΔlnL) for the models (M1a vs M2a, M7 vs M8 and M8a vs M8) is compared to a χ^2 distribution (2 degrees of freedom for the M1a vs M2a and M7 vs M8 comparisons, 1 degree of freedom for M8a vs M8). Positively selected sites were identified using the Bayes Empirical Bayes (BEB) analysis, which calculates the posterior probability that each codon is from the site class of positive selection (under model M8) (Anisimova et al. 2002). To be conservative, we considered a posterior probability ≥ 0.95.

To analyze the presence of episodic positive selection we applied the branch-site test (Zhang et al. 2005) from the PAML suite (Yang 2007). The test is based on the comparison between two nested models: a model (MA) that allows positive selection on one or more lineages (called foreground lineages), and a model (MA1) that does not allow such positive selection. The ΔlnL for the two models is then compared to a χ^2 distribution with one degree of freedom (Zhang et al. 2005). A false discovery rate correction was applied to take into account a multiple hypothesis issue generated by analyzing different branches on the same phylogeny (Anisimova and Yang 2007). When the likelihood ratio test suggested the action of positive selection, the Bayes Empirical Bayes (BEB) analysis was used to evaluate the posterior probability (with a cut-off of 0.95) that each codon belongs to the site class of positive selection on the foreground branch.

Detection of co-evolving sites

In order to analyze the presence of co-evolving sites in Filovirus GP, we applied two different methods: BGM (Bayesian Graphical Model)-Spidermonkey (Poon et al. 2007) and the Mutual Information Server To Infer Coevolution (MISTIC) (Simonetti et al. 2013). BGM-Spidermonkey identifies co-evolving sites from coding sequences; a Bayesian Graphical Model is used to evaluate the connection among codons in the alignment (represented by the nodes of the network). Significant statistical associations between nodes are indicated by the edges of the network, suggesting functional or structural interactions between codons. BGM-Spidermonkey is implemented in the HYPHY package.

MISTIC estimates the relationship between two or more alignment positions. The co-evolutionary association is evaluated by Mutual Information (MI), estimating whether the information from the amino acid at the first position can help to predict the amino acid information at the second position.

For BGM-Spidermonkey sites were filtered based on a minimum count of 4 substitutions across the phylogeny and each site was conditionally dependent on one other site. To be conservative, we considered a pair of residues as co-evolving if they showed a posterior probability >0.95. Likewise, for MISTIC site pairs were required to display a MI rank higher that the 95th percentile calculated using all MI scores from the alignment. Pairs of sites exceeding the threshold for both methods were declared to be co-evolving.

Positive selection in Homininae

For gammaMap (Wilson et al. 2011) analysis, genotype data from the phase 1 of the 1000 Genomes Project were retrieved from the dedicated website (1000 Genomes Project Consortium et al. 2012); we retrieved SNP information for the three human populations: African (Yoruba), European, and Chinese. For the chimpanzee and gorilla analyses, genotype information were retrieved from (Prado-Martinez et al. 2013) for 25 and 27 individuals, respectively.

The ancestral sequence was reconstructed from the human, chimpanzee, orangutan and macaque sequences. Ancestral sequence reconstruction (ASR) was performed through the DataMonkey sever (Delport et al. 2010) using the ASR utility, which implements three different methods based on maximum-likelihood or Bayesian inference (Kosakovsky Pond and Frost 2005). The three methods yieleded consistent results at all positions.

GammaMap applies a population genetics-phylogenetics approach using intra-specific

variation and inter-specific diversity to estimate the distribution of population-scaled selection coefficients (γ) along coding regions. γ values are classified into 12 categories, from inviable (γ =-500) to strongly beneficial (γ=100), with γ equal to 0 indicating neutrality. In the analysis, we assumed θ (neutral mutation rate per site) and k (transitions/transversions ratio) to vary following log-normal distributions; for T (branch length) a normal distribution was adopted. For p (the probability that adjacent codons share the same population-scaled selection coefficient) we assumed a value of 0.02. We set the neutral frequencies of non-STOP codons (pi) to 1/61. For population-scaled selection coefficients we considered a uniform Dirichlet distribution with the same prior weight for each selection class. Two Markov Chain Monte Carlo runs of 100,000 iterations each were run with a thinning interval of 10 iterations. Runs were compared for convergence and merged for the analyses.

3D structure mapping and rendering

Protein 3D structure of NPC1-GP complex (PDB ID: 5F1B) (Wang et al. 2016), *Sudan ebolavirus* glycoprotein bound to 16F6 (PDB ID: 3VE0) (Bale et al. 2012), *Zaire ebolavirus* glycoprotein bound to 14G7 (PDBID: 2Y6S) (Olal et al. 2012) and to KZ52 (PDB ID: 3CSY) (Lee et al. 2008) were derived from the Protein Data Bank (PDB). Sites were mapped onto structures using PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.2 Schrödinger, LLC).

Acknowledgments

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

References

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature. 491:56-65.

1000 Genomes Project Consortium, Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. Nature. 467:1061-1073.

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

Anisimova M, Yang Z. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol. 24:1219-1228.

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

Azarian T, Lo Presti A, Giovanetti M, Cella E, Rife B, Lai A, Zehender G, Ciccozzi M, Salemi M. 2015. Impact of spatial dispersion, evolution, and selection on Ebola Zaire Virus epidemic waves. Sci Rep. 5:10170.

Bale S, Dias JM, Fusco ML, Hashiguchi T, Wong AC, Liu T, Keuhne AI, Li S, Woods VL,Jr, Chandran K et al. . 2012. Structural basis for differential neutralization of ebolaviruses. Viruses. 4:447-470.

Barreiro LB, Ben-Ali M, Quach H, Laval G, Patin E, Pickrell JK, Bouchier C, Tichit M, Neyrolles O, Gicquel B et al. . 2009. Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. PLoS Genet. 5:e1000562.

Bedhomme S, Hillung J, Elena SF. 2015. Emerging viruses: why they are not jacks of all trades? Curr Opin Virol. 10:1-6.

Bloom JD, Gong LI, Baltimore D. 2010. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. Science. 328:1272-1275.

Brauburger K, Hume AJ, Muhlberger E, Olejnik J. 2012. Forty-five years of Marburg virus research. Viruses. 4:1878-1927.

Carroll SA, Towner JS, Sealy TK, McMullan LK, Khristova ML, Burt FJ, Swanepoel R, Rollin PE, Nichol ST. 2013. Molecular evolution of viruses of the family Filoviridae based on 97 whole-genome sequences. J Virol. 87:2608-2616.

Casadevall A, Pirofski LA. 2015. The Ebola epidemic crystallizes the potential of passive antibody therapy for infectious diseases. PLoS Pathog. 11:e1004717.

Chandran K, Sullivan NJ, Felbor U, Whelan SP, Cunningham JM. 2005. Endosomal proteolysis of the Ebola virus glycoprotein is necessary for infection. Science. 308:1643-1645.

de La Vega MA, Stein D, Kobinger GP. 2015. Ebolavirus Evolution: Past and Present. PLoS Pathog. 11:e1005221.

Delport W., A. F. Poon, S. D. Frost, and S. L. Kosakovsky Pond. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 26:2455-2457.

Demogines A, Abraham J, Choe H, Farzan M, Sawyer SL. 2013. Dual host-virus arms races shape an essential housekeeping protein. PLoS Biol. 11:e1001571.

Garver WS. 2011. Gene-diet interactions in childhood obesity. Curr Genomics. 12:180-189.

Garver WS, Jelinek D, Meaney FJ, Flynn J, Pettit KM, Shepherd G, Heidenreich RA, Vockley CM, Castro G, Francis GA. 2010. The National Niemann-Pick Type C1 Disease Database: correlation of lipid profiles, mutations, and biochemical phenotypes. J Lipid Res. 51:406-415.

Gharib WH, Robinson-Rechavi M. 2013. The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. Mol Biol Evol. 30:1675-1686

Gire SK, Goba A, Andersen KG, Sealfon RS, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G et al. . 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 345:1369-1372.

Gong LI, Suchard MA, Bloom JD. 2013. Stability-mediated epistasis constrains the evolution of an influenza protein. Elife. 2:e00631.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59:307-321.

Hevey M, Negley D, Schmaljohn A. 2003. Characterization of monoclonal antibodies to Marburg virus (strain Musoke) glycoprotein and identification of two protective epitopes. Virology. 314:350-357.

Jae LT, Brummelkamp TR. 2015. Emerging intracellular receptors for hemorrhagic fever viruses. Trends Microbiol. 23:392-400.

Jeffers SA, Sanders DA, Sanchez A. 2002. Covalent modifications of the ebola virus glycoprotein. J Virol. 76:12463-12472.

Kaelber JT, Demogines A, Harbison CE, Allison AB, Goodman LB, Ortega AN, Sawyer SL, Parrish CR. 2012. Evolutionary reconstructions of the transferrin receptor of Caniforms supports canine parvovirus being a re-emerged and not a novel pathogen in dogs. PLoS Pathog. 8:e1002666.

Kaletsky RL, Simmons G, Bates P. 2007. Proteolysis of the Ebola virus glycoproteins enhances virus binding and infectivity. J Virol. 81:13378-13384.

Kosakovsky Pond SL, Frost SD. 2005. Not so

different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 22:1208-1222.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. 2006. Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol. 23:1891-1901.

Kryazhimskiy S, Dushoff J, Bazykin GA, Plotkin JB. 2011. Prevalence of epistasis in the evolution of influenza A surface proteins. PLoS Genet. 7:e1001301.

Kuhn JH, Becker S, Ebihara H, Geisbert TW, Johnson KM, Kawaoka Y, Lipkin WI, Negredo AI, Netesov SV, Nichol ST et al. . 2010. Proposal for a revised taxonomy of the family Filoviridae: classification, names of taxa and viruses, and virus abbreviations. Arch Virol. 155:2083-2103.

Ladner JT, Wiley MR, Mate S, Dudas G, Prieto K, Lovett S, Nagle ER, Beitzel B, Gilbert ML, Fakoli L et al. . 2015. Evolution and Spread of Ebola Virus in Liberia, 2014-2015. Cell Host Microbe. 18:659-669.

Lee JE, Fusco ML, Hessell AJ, Oswald WB, Burton DR, Saphire EO. 2008. Structure of the Ebola virus glycoprotein bound to an antibody from a human survivor. Nature. 454:177-182.

Li YH, Chen SP. 2014. Evolutionary history of Ebola virus. Epidemiol Infect. 142:1138-1145.

Loytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci U S A. 102:10557-10562.

Manicassamy B, Wang J, Jiang H, Rong L. 2005. Comprehensive analysis of ebola virus GP1 in viral entry. J Virol. 79:4793-4805.

Mari Saez A, Weiss S, Nowak K, Lapeyre V, Zimmermann F, Dux A, Kuhl HS, Kaba M, Regnaut S, Merkel K et al. . 2014. Investigating the zoonotic origin of the West African Ebola epidemic. EMBO Mol Med. 7:17-23.

Messaoudi I, Amarasinghe GK, Basler CF. 2015. Filovirus pathogenesis and immune evasion: insights from Ebola virus and Marburg virus. Nat Rev Microbiol. 13:663-676.

Miller EH, Obernosterer G, Raaben M, Herbert AS, Deffieu MS, Krishnan A, Ndungo E, Sandesara RG, Carette JE, Kuehne AI et al. . 2012. Ebola virus entry requires the host-programmed recognition of an intracellular receptor. Embo j. 31:1947-1960.

Moller-Tank S, Maury W. 2015. Ebola virus entry: a curious and complex series of events. PLoS Pathog. 11:e1004731.

Murin CD, Fusco ML, Bornholdt ZA, Qiu X, Olinger GG, Zeitlin L, Kobinger GP, Ward AB, Saphire EO. 2014. Structures of protective antibodies reveal sites of vulnerability on Ebola virus. Proc Natl Acad Sci U S A. 111:17182-17187.

Negredo A, Palacios G, Vazquez-Moron S, Gonzalez F, Dopazo H, Molero F, Juste J, Quetglas J, Savji N, de la Cruz Martinez M et al. . 2011. Discovery of an ebolavirus-like filovirus in europe. PLoS Pathog. 7:e1002304.

Neverov AD, Kryazhimskiy S, Plotkin JB, Bazykin GA. 2015. Coordinated Evolution of Influenza A Surface Proteins. PLoS Genet. 11:e1005404.

Ng M, Ndungo E, Kaczmarek ME, Herbert AS, Binger T, Kuehne AI, Jangra RK, Hawkins JA, Gifford RJ, Biswas R et al. . 2015. Filovirus receptor NPC1 contributes to species-specific patterns of ebolavirus susceptibility in bats. Elife. 4:10.7554/eLife.11785.

Olal D, Kuehne AI, Bale S, Halfmann P, Hashiguchi T, Fusco ML, Lee JE, King LB, Kawaoka Y, Dye JM,Jr et al. . 2012. Structure of an antibody in complex with its mucin domain linear epitope that is protective against Ebola virus. J Virol. 86:2809-2816.

Olival KJ, Hayman DT. 2014. Filoviruses in bats: current knowledge and future directions. Viruses. 6:1759-1788.

Park DJ, Dudas G, Wohl S, Goba A, Whitmer SL, Andersen KG, Sealfon RS, Ladner JT, Kugelman JR, Matranga CB et al. . 2015. Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. Cell. 161:1516-1526.

Peake KB, Vance JE. 2010. Defective cholesterol trafficking in Niemann-Pick C-deficient cells. FEBS Lett. 584:2731-2739.

Penn O, Privman E, Ashkenazy H, Landan G, Graur D, Pupko T. 2010. GUIDANCE: a web server for assessing alignment confidence scores. Nucleic Acids Res. 38:W23-8.

Pond SL, Frost SD, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 21:676-679.

Poon AF, Lewis FI, Pond SL, Frost SD. 2007. An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope. PLoS Comput Biol. 3:e231.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G et al. . 2013. Great ape genetic diversity and population history. Nature. 499:471-475.

Privman E, Penn O, Pupko T. 2012. Improving the performance of positive selection inference by filtering unreliable alignment regions. Mol Biol Evol. 29:1-5.

Qiu X, Wong G, Fernando L, Audet J, Bello A, Strong J, Alimonti JB, Kobinger GP. 2013. mAbs and Ad-vectored IFN-alpha therapy rescue Ebola-infected nonhuman primates when administered after the detection of viremia and symptoms. Sci Transl Med. 5:207ra143.

Schornberg K, Matsuyama S, Kabsch K, Delos S, Bouton A, White J. 2006. Role of endosomal cathepsins in entry mediated by the Ebola virus glycoprotein. J Virol. 80:4174-4178.

Shehu A, Kavraki LE. 2012. Modeling Structures and Motions of Loops in Protein Molecules. Entropy. 14:252.

Simonetti FL, Teppa E, Chernomoretz A, Nielsen M, Marino Buslje C. 2013. MISTIC: Mutual information server to infer coevolution. Nucleic Acids Res. 41:W8-14.

Simon-Loriere E, Faye O, Faye O, Koivogui L, Magassouba N, Keita S, Thiberge JM, Diancourt L, Bouchier C, Vandenbogaert M et al. . 2015. Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic. Nature. 524:102-104.

Sironi M, Cagliani R, Forni D, Clerici M. 2015. Evolutionary insights into host-pathogen interactions from mammalian sequence data. Nat Rev Genet. 16:224-236.

Taylor DJ, Leach RW, Bruenn J. 2010. Filoviruses are ancient and integrated into mammalian genomes. BMC Evol Biol. 10:193-2148-10-193.

Taylor DJ, Dittmar K, Ballinger MJ, Bruenn JA. 2011. Evolutionary maintenance of filovirus-like genes in bat genomes. BMC Evol Biol. 11:336-2148-11-336.

Taylor DJ, Ballinger MJ, Zhan JJ, Hanzly LE, Bruenn JA. 2014. Evidence that ebolaviruses and cuevaviruses have been diverging from marburgviruses since the Miocene. Peerj. 2:e556.

Towner JS, Amman BR, Sealy TK, Carroll SA, Comer JA, Kemp A, Swanepoel R, Paddock CD, Balinandi S, Khristova ML et al. . 2009. Isolation of genetically diverse Marburg

viruses from Egyptian fruit bats. PLoS Pathog. 5:e1000536.

Walsh PD, Biek R, Real LA. 2005. Wave-like spread of Ebola Zaire. PLoS Biol. 3:e371.

Walsh PD, Abernethy KA, Bermejo M, Beyers R, De Wachter P, Akou ME, Huijbregts B, Mambounga DI, Toham AK, Kilbourn AM et al. . 2003. Catastrophic ape decline in western equatorial Africa. Nature. 422:611-614.

Wang H, Shi Y, Song J, Qi J, Lu G, Yan J, Gao GF. 2016. Ebola Viral Glycoprotein Bound to Its Endosomal Receptor Niemann-Pick C1. Cell. 164:258-268.

White JM, Schornberg KL. 2012. A new player in the puzzle of filovirus entry. Nat Rev Microbiol. 10:317-322.

Wilson DJ, Hernandez RD, Andolfatto P, Przeworski M. 2011. A population genetics-phylogenetics approach to inferring natural selection in coding sequences. PLoS Genet. 7:e1002395.

Wilson JA, Hevey M, Bakken R, Guest S, Bray M, Schmaljohn AL, Hart MK. 2000. Epitopes involved in antibody-mediated protection from Ebola virus. Science. 287:1664-1666.

Wong WS, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics. 168:1041-1051.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586-1591.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555-556.

Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol. 22:1107-1118.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 22:2472-2479.

3.2.2.2 Adaptive evolution underlies the species-specific binding of *P. falciparum* RH5 to human basigin

Plasmodium falciparum, the causative agent of most malaria-related deaths in humans, is a member of the Laverania subgenus, which includes ape-infecting *Plasmodium* protozoa. Although *P. falciparum* infection is now restricted to humans, phylogenetic reconstruction of *Laverania* sequences suggested that all extant *P. falciparum* strains resulted from a single gorilla to human cross-species transmission event [176].

Host shifts are often mediated by adaptive evolution at genes that encode host-pathogen interacting partners, and evolutionary analysis can provide valuable insight into the molecular determinants of zoonotic transmission events and spillovers [12]. I thus analyzed the evolutionary history of two parasite ligand-host receptor pairs that were suggested to determine the species tropism of *P. falciparum*. In particular, recent evidence indicated that binding of Reticulocyte-binding protein Homolog 5 (RH5) to human basigin (BSG) is a major determinant of species-specificity; a less important role was ascribed to the interaction between Erythrocyte Binding Antigen-175 (EBA-175) and human glycophorin A (GYPA). Basigin is a multifunctional protein with a role in trophoblast function and in spermatogenesis. This is therefore another example showing that host-pathogen interactions are not limited to the immune system components.

By sequencing of *BSG* genes from Old and New World monkeys, I show that BSG evolved adaptively in primates with selection targeting two sites (F27 and H102) that were previously shown to modulated PfRH5 binding [65]. A population-genetic phylogentics approach detected the strongest selection for the gorilla lineage; in both gorilla and chimpanzee, selection targeted residues that affect the binding specificity of PfRH5, as previously

demonstrated in binding assays [65].

On the parasite side, natural selection also operated. Using different methods to detect episodic positive selection we show that the RH5 gene evolved adaptively on the *P. falciparum* branch; one of the two positively selected sites we detected (W447) is known to stabilize the interaction with human basigin.

As for the *EBA175* gene, we detect no selection in the *P. falciparum* lineage. Its host receptor, *GYPA*, shows evidence of positive selection in all hominid lineages; selected codons include glycosylation sites that modulate PfEBA175 binding affinity.

Data herein suggest that BSG-binding *Plasmodium* parasites have been circulating for a long time among wild primate populations and support the notion that adaptive changes in the BSG-RH5 interacting partners contribute to host shifts. With respect to the more recent events that led to the emergence of *P. falciparum* as a human pathogen, these analyses implicate that its gorilla-infecting ancestor encoded a functional *RH5* gene and exerted a strong selective pressure on *BSG*. Adaptive changes in the gorilla *BSG* and in the parasite *RH5* genes most likely contributed to the shift to human hosts.

Conversely, these results suggest that variation at *PfEBA175* had no major role in the emergence of *P. falciparum* as a human pathogen. Clearly, this observation does not imply that interaction between GYPA and EBA-175 is unimportant for erythrocyte invasion, but supports the notion that it is not a determinant of host tropism [65].

In summary, these data provide an evolutionary explanation for species-specific binding of the PfRH5-BSG ligand-receptor pair and supports the role of adaptive evolution at these genes as a determinant of the host shift

that gave rise to *P. falciparum* as a major human pathogen. This work exemplifies how evolutionary information can provide insight into infectious disease emergence and zoonotic transmission.

Personal contribution to the work: I performed the analyses and I analyzed data. I also produced figures and tables for the manuscript.

Positive selection underlies the species-specific binding of *P. falciparum* RH5 to human basigin

Diego Forni[1]*, Chiara Pontremoli[1]*, Rachele Cagliani[1], Uberto Pozzoli[1], Mario Clerici[2,3], Manuela Sironi[1]

[1] Scientific Institute IRCCS E.MEDEA, Bioinformatics, 23842 Bosisio Parini, Italy;
[2] Department of Physiopathology and Transplantation, University of Milan, 20090 Milan, Italy;
[3] Don C. Gnocchi Foundation ONLUS, IRCCS, 20148 Milan, Italy.

* these authors equally contributed to this work

Corresponding author: Diego Forni - Bioinformatics - Scientific Institute IRCCS E.MEDEA, 23842 Bosisio Parini, Italy. Tel: +39-031877826; Fax:+39-031877499; e-mail: diego.forni@bp.lnf.it

Abstract

Plasmodium falciparum, the causative agent of the deadliest form of malaria, is a member of the *Laverania* subgenus, which includes ape-infecting parasites. *P. falciparum* is thought to have originated in gorillas, although infection is now restricted to humans. *Laverania* parasites display remarkable host-specificity, which is partially mediated by the interaction between parasite ligands and host receptors. We analyze the evolution of *BSG* (basigin) and *GYPA* (glycophorin A) in primates/hominins, as well as of their *Plasmodium*-encoded ligands, PfRH5 and PfEBA175. We show that, in primates, positive selection targeted two sites in BSG (F27 and H102), both involved in PfRH5 binding. A population-genetic phylogentics approach detected the strongest selection for the gorilla lineage: one of the positively selected sites (K191) is a major determinant of PfRH5 binding affinity. Analysis of *RH5* genes indicated episodic selection on the *P. falciparum* branch; the positively selected W447 site is known to stabilize the interaction with human basigin. Conversely, we detect no selection in the receptor binding region of EBA175 in the *P. falciparum* lineage. Its host receptor, GYPA, shows evidence of positive selection in all hominid lineages; selected codons include glycosylation sites that modulate PfEBA175 binding affinity. Data herein provide an evolutionary explanation for species-specific binding of the PfRH5-BSG ligand-receptor pair and support the hypothesis that positive selection at these genes drove the host shift leading to the emergence of *P. falciparum* as a human pathogen.

Introduction

Plasmodium falciparum, the causative agent of most malaria-related deaths in humans, is a member of the *Laverania* subgenus, which includes ape-infecting *Plasmodium* protozoa. Analysis of African wild-ape populations indicated that *Laverania* parasites have remarkable host-specificity and revealed a gorilla origin for *P. falciparum* (Liu *et al.* 2010). Indeed, phylogenetic reconstruction of *Laverania* sequences suggested that all extant *P. falciparum* strains resulted from a single gorilla to human cross-species transmission event (Liu *et al.* 2010). Despite its origin, field studies have found no evidence of *P. falciparum* infection in gorillas, suggesting that the parasite is unable to cross back the species barrier (Liu *et al.* 2010).

The molecular determinants underlying the species-specificity of *Laverania* parasites are only beginning to be elucidated. An essential step for *Plasmodium* infection is erythrocyte invasion. This requires a sequential cascade of events that involve multiple interactions between merozoite ligands and host receptors.

Erythrocyte Binding Antigen-175 (PfEBA175) was the first *P. falciparum* invasion ligand to be identified (Orlandi *et al.* 1990; Sim *et al.* 1994); it interacts with glycophorin-A (GYPA), a highly abundant erythrocyte surface sialoglycoprotein. Specifically, PfEBA175 binding requires sialic acid residues displayed on GYPA. This observation, together with the notion that humans lack a functional cytidine monophosphate-*N*-acetylneuraminic acid hydroxylase (*CMAH*) gene and therefore display different glycosylation patterns compared to other apes, led to the suggestion that the interaction between GYPA and PfEBA175 played a major role as a species-specific determinant (Martin *et al.* 2005). This view was recently challenged by *in vitro* binding assays showing that PfEBA175 orthologs from chimpanzee-infecting parasites (*P. reichenowi* and *P. billcollinsi*) can bind human GYPA in a sialic acid-dependent manner (Wanaguru *et al.* 2013b). Conversely, species-specific interactions were reported for another pair of ligand-receptor partners: the *P. falciparum* reticulocyte-binding protein homologue 5 (PfRH5) and basigin (BSG), its erythrocyte cell-surface receptor (Wanaguru *et al.* 2013b; Crosnier *et al.* 2011). In particular, Wanaguru and coworkers showed that PfRH5 can bind chimpanzee BSG, but with lower affinity compared to the human protein, and it is unable to bind BSG of gorilla origin (Wanaguru *et al.* 2013b). PfRH5 belongs to the reticulocyte binding-like (RBL) family of *Laverania* proteins that includes different numbers of active paralogs in distinct species (Otto *et al.* 2014). *PfRH5* polymorphisms in laboratory-adapted *P. falciparum* strains allow invasion of red blood cells from New World primates and rodents, substantiating the role of PfRH5 as an important determinant of host range (Hayton *et al.* 2013).

By providing information on the location and nature of functional genetic variation, evolutionary analysis may help explain changes in pathogen tropism (Sironi *et al.* 2015). Using different strategies, we investigated the evolutionary history of primate *BSG* and *GYPA*, and of the *Plasmodium* ligands *EBA175* and *RH5*. Results are consistent with selection acting on both primate genes, as well as on the *PfRH5* gene during *P. falciparum* speciation.

Materials and Methods

Sequences and samples

Available primate sequences for *BSG* were retrieved from the Ensembl website (http://www.ensembl.org/index.html) and NCBI database (http://www.ncbi.nlm.nih.gov/). *BSG* coding sequencing information for *Aotus trivirgatus*, *Saguinus oedipus* and *Chlorocebus aethiops* was obtained by direct sequencing of cDNA derived from OMK, EBV B95 UK and COS1 cells, respectively (obtained by the European Collection of Cell Cultures , ECACC). The genomic DNA of *Colobus guereza*, *Ateles fusciceps*, *Pithecia pithecia*, and *Lagothrix lagotricha* was kindly provided by the Gene Bank of Primates, Primate Genetics (Germany) and used as a template for *BSG* amplification and sequencing. The genomic DNA of one *Gorilla gorilla* and two *Pan troglodytes* were obtained from ECACC and used to amplify and sequence the *BSG* coding region. This procedure was meant to assess the presence of sequence errors or rare variants in the gorilla and chimpanzee reference genome assemblies (gorGor3.1/gorGor3 and CSAC 2.1.4/panTro4).

The cDNAs/DNAs were amplified by PCR (primer sequences are listed in Table S1, Supporting information), and then treated with ExoSAP-IT (USB Corporation Cleveland Ohio, USA). Purified PCR products were 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 (Life Technologies). Sequences were assembled using AutoAssembler version 1.4.0 (Applied Biosystems), and manually inspected. The obtained sequences have been submitted to NCBI database (accession numbers: KR425477-KR425483, KT358396-KT358397). A list of analyzed species and their accession numbers is reported in Table S2, Supporting information.

The chimpanzee and gorilla *GYPA* coding sequences were obtained from the respective reference genomes (gorGor3.1/gorGor3 and CSAC 2.1.4/panTro4) and checked against *GYPA* sequences obtained by Sanger sequencing experiments (Baum *et al.* 2002; Xie *et al.* 1997). The reference chimpanzee *GYPA*

sequence was identical to that previously obtained by Baum and co-workers (Baum *et al.* 2002). As for gorilla, the reference sequence was compared to those obtained through cDNA sequencing (Xie *et al.* 1997): minor differences were detected, which were in all cases consistent with the presence of polymorphisms (Xie *et al.* 1997; Prado-Martinez *et al.* 2013) with the exclusion of two codons (reference: 72S and 81G; all other sequences: 72P and 81V). These were thus replaced before performing ancestral sequence reconstruction (see below).

For the analysis of positive selection in *RH5* and *EBA175*, *Plasmodium* sequences were retrieved from the NCBI database (a list of accession numbers is provided as Table S3, Supporting information) or derived from previous works (Wanaguru *et al.* 2013b; Otto *et al.* 2014; Hayton *et al.* 2008).

Evolutionary analyses

DNA alignments were performed with the RevTrans 2.0 utility, (http://www.cbs.dtu.dk/services/RevTrans/, using MAFFT v6.240 as an aligner) (Wernersson & Pedersen 2003), which uses the protein sequence alignment as a scaffold to construct the corresponding DNA multiple alignment. This latter was checked and edited by TrimAl to remove alignment uncertainties (http://phylemon.bioinfo.cipf.es/utilities.html, with the *automated1* method) (Capella-Gutierrez *et al.* 2009). Primate gene trees were generated by maximum-likelihood using the program phyML with a GTR plus gamma-distributed rates model (Guindon *et al.* 2009). For analyses using *Plasmodium* sequences were used previously generated trees (Wanaguru *et al.* 2013b).

Positive selection in primate *BSG* gene was detected using the PAML (Phylogenetic Analysis by Maximum Likelihood) package (Yang 2007). The site models implemented in the *codeml* program were developed to detect positive selection affecting only a few amino acid residues in a protein: positive selection is characterized by a non-synonymous substitution/synonymous substitution rate ratio (dN/dS, also referred to as ω) >1. To detect selection, site models that allow (M2a, M8) or disallow (M1a, M7, M8a) a class of sites to evolve with ω >1 were fitted to the data using the F61 model (frequencies of each of the 61 non-stop codons estimated from the data). Positively selected sites were identified using two different methods: the Bayes Empirical Bayes (BEB) analysis (with a cut-off of 0.90), which calculates the posterior probability that each codon is from the site class of positive selection (under model M8) (Anisimova *et al.* 2002), and the Mixed Effects Model of Evolution (MEME, with the default cut-off of 0.1) (Murrell *et al.* 2012), which allows the distribution of ω to vary from site to site and from branch to branch at a site. Only sites detected using both methods were considered positively selected.

To detect positive selection in *RH5* and *EBA175* we applied the branch-site test from the PAML suite (Zhang *et al.* 2005) and BUSTED (branch-site unrestricted statistical test for episodic diversification, Murrell *et al.* 2015), which is implemented in the HYPHY package (Pond *et al.* 2005). The branch-site test compares a model (MA) that allows positive selection on one or more lineages (foreground lineages) with a model (MA1) that does not allow such positive selection. Twice the difference of likelihood for the two models (ΔlnL) is then compared to a χ^2 distribution with one degree of freedom (Zhang *et al.* 2005). Positively selected sites can be identified through the BEB analysis.

BUSTED is designed to detect the action of episodic positive selection that is acting on a subset of branches in the phylogeny at a proportion of sites within the alignment. To detect selection at individual sites, twice the difference of the likelihood for the alternative and the null model is compared to a χ^2 distribution (df=1); we considered a site as positively selected if it showed a p value < 0.05.

To identify the action of recombination along the gene alignment, we applied GARD (genetic algorithm recombination detection) (Kosakovsky Pond *et al.* 2006).

SLAC (Kosakovsky Pond & Frost 2005), GARD and MEME analyses were performed through the DataMonkey server (Delport *et al.* 2010) (http://www.datamonkey.org).

Population genetics-phylogenetics analysis

Genotype data from the Pilot 1 phase of the

1000 Genomes Project were retrieved from the dedicated website for human individuals (1000 Genomes Project Consortium *et al.* 2010). SNP genotype informations for 25 unrelated chimpanzees and 23 unrelated gorillas were retrieved from a previous work (Prado-Martinez *et al.* 2013). Coding sequence information was obtained for *GYPA* and *BSG*. For *BSG* the ancestral sequence was reconstructed by parsimony from the human, chimpanzee, orangutan and macaque sequences; for *GYPA*, due to the absence of an unambiguous orthologue in orangutan and macaque genomes, the ancestral sequence was reconstructed from the human, chimpanzee, and gorilla sequences.

Analyses were performed with gammaMap (Wilson *et al.* 2011): we assumed θ (neutral mutation rate per site), k (transitions/transversions ratio), and T (branch length) to vary among the gene following log-normal distributions. For each gene we set the neutral frequencies of non-STOP codons (1/61) and the probability that adjacent codons share the same selection coefficient (p=0.02). For selection coefficients we considered a uniform Dirichlet distribution with the same prior weight for each selection class. For each gene we run 100,000 iterations with thinning interval of 10 iterations.

To be conservative, we declared a codon to be targeted by positive selection when the cumulative posterior probability of $\gamma \geq 1$ was > 0.75, as suggested (Quach *et al.* 2013).

Results

Natural selection at the primate *BSG* and *Plasmodium RH5* genes contributes to species-specific binding

To analyze the evolutionary history of *BSG*, we obtained coding sequence information for 28 primate species either from public databases or by sequencing (see methods). These sequence data allow sufficient power to detect positive selection at primate genes (McBee *et al.* 2015). The tree shrew and colugo sequences were also included (Table S2, Supporting information). Notably, sequencing of genomic DNA from two chimpanzees and one gorilla revealed some differences compared to the reference *BSG* sequences (as predicted from the respective genome assemblies) (Fig. 1A). These differing positions were checked against genotype data from 25 *Pan troglodytes* and 23 *Gorilla gorilla* (Prado-Martinez *et al.* 2013), which revealed no polymorphism and were consistent with the sequence data we obtained. Thus, the following analyses were performed using these newly generated sequences. We conclude that the reference *BSG* sequences contain either errors or rare ploymorhisms/haplotypes. DNA alignments were generated using RevTrans (Wernersson & Pedersen 2003) and screened for the presence of recombination using GARD (genetic algorithm recombination detection) (Kosakovsky Pond *et al.* 2006). This is because recombination can be mistaken as positive selection. No breakpoint was detected.

The average nonsynonymous/synonymous substitution rate ratio (dN/dS, also referred to as ω) was calculated using the single-likelihood ancestor counting (SLAC) method (Kosakovsky Pond & Frost 2005). dN/dS was lower than 1 (ω =0.31, 95% CI = 0.29 - 0.34), indicating purifying selection as the major driving force acting on *BSG* gene in primates. This finding does not exclude that localized positive selection acts on specific sites or domains of the protein. To test this possibility we analyzed the *BSG* alignment using the *codeml* program (Yang 2007; Yang 1997). Using likelihood ratio tests (LRTs), c*odeml* compares models of gene evolution that allow (NSsite models M2a and M8, positive selection models) or disallow (NSsite models M1a, M7, and M8a null models) a class of codons to evolve with dN/dS > 1. In the case of *BSG* all the neutral models were rejected in favor of the positive selection models (Table 1). We next applied the Bayes Empirical Bayes (BEB) analysis (Anisimova *et al.* 2002; Yang *et al.* 2005) and the Mixed Effects Model of Evolution (MEME) (Murrell *et al.* 2012) to identify specific sites targeted by positive selection in *BSG*; to be conservative only sites detected using both methods were considered. Two positively selected sites were identified in the extracellular domain (Table 1), F27 and H102 (Schlegel *et al.* 2009). Both sites are located at the direct contact interface with PfRH5 (Wright *et al.* 2014) (Fig. 1A-B); consistently, the F27L mutation reduces the affinity of human BSG for PfRH5 (Wanaguru *et al.* 2013b). Remarkably, some of the

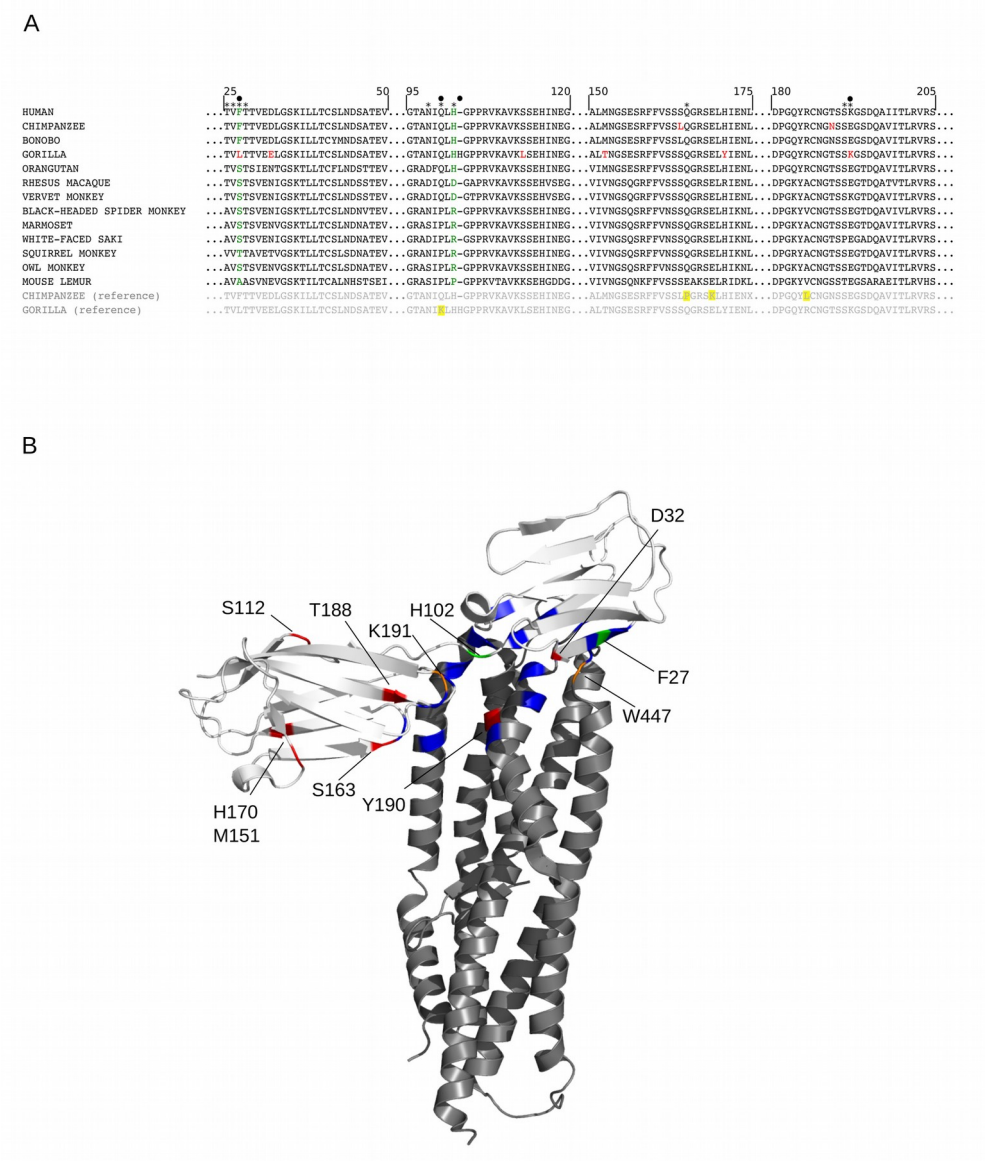


Figure 1. Positive selection at BSG PfRH5 interface. (A) Multiple sequence alignment of BSG. Alignment of a portion of BSG protein for a few representative primate species is shown. The amino acid translation corresponding to the *BSG* coding regions inferred from the chimpanzee and gorilla reference genomes are also shown in gray; amino acids that represent sequence errors or rare variants are shaded in yellow. Positively selected sites identified by gammaMap or BUSTED are shown in red; residues identified with BEB and MEME are shown in green. Sites that interact with PfRH5 are marked with a star. Black dots indicate basigin residues which, when mutated, affect PfRH5 affinity (Wanaguru *et al.* 2013b). (B) Ribbon representation (PDB ID: 4U0Q) of human basigin (white) in complex with PfRH5 (dark gray). Color codes are as in Figure 1A; residue W447 in PfRH5, as well as BSG positively selected sites that interact with PfRH5, are shown in orange; residues involved in the interaction are shown in blue.

sequence errors/rare variants in the reference genomes also correspond to codons that map to the binding interface with PfRH5. This is the case of Q164 in chimpanzee (P164 in the reference sequence) and Q100 in gorilla (K100 in the reference). Based on the inferred Q100K change, this position was previously indicated as a determinant of species-specific binding (Wanaguru et al. 2013b) (Fig. 1A).The only known ortholog of PfRH5 is the PrRH5 gene from P. reichenowi, a chimpanzee-infecting parasite. Evolutionary analysis of *RH5* was thus performed by including all available *PfRH5* genes from 11 different *P. falciparum* strains and the *P. reichenowi* sequence (Table S3, Supporting information); the *PfRH2b* paralog was included to anchor the phylogeny.

Screening with GARD, identified a recombination breakpoint at codon position 23. We thus omitted the first 23 codons from the analysis. Evidence of episodic selection along the branch leading to *P. falciparum* was searched for by using the branch-site models implemented in *codeml* (Zhang *et al.* 2005), which provided statistically significant evidence of positive selection, but detected no positively selected sites (Table 2). To validate this result, and to further investigate the selection pattern of *RH5*, we applied BUSTED (branch-site unrestricted statistical test for episodic diversification), a recently developed method to detect episodic positive selection (Murrell *et al.* 2015). This analysis confirmed selection on the *Plasmodium falciparum* branch (Table 2, Fig. 2A). The implemented maximum-likelihood test identified Y190 and W447 as positively selected (Table 2). W447 was previously shown to stabilize the interaction between PfRH5 and human basigin (Wright *et al.* 2014) (Fig. 1B).

The receptor-binding region of *EBA175* is not a selection target in the *P. falciparum* lineage

The evolutionary history of *PfEBA175* was analyzed by retrieval of 45 *Laverania* coding sequences of the receptor binding region (RII) (Wanaguru *et al.* 2013b) (Fig. 2B and Table S3, Supporting information). GARD detected no breakpoint and, in analogy to *RH5*, episodic positive selection was searched for using the *codeml* branch-site tests. We tested both the *P. falciparum* branch and the node which includes *P. falciparum* and *P. praefalciparum* (the closest species (Rayner *et al.* 2011)). The analysis didn't provide any statistically significant evidence of positive selection; similar results were obtained with BUSTED (Table 2).

Positive selection in primate lineages targets residues at the interaction surface with parasite proteins

To gain insight into the more recent selective events in specific lineages, we applied a population genetics-phylogenetics approach (gammaMap (Wilson et al. 2011)) to study the evolution of the BSG gene in Hominidae. This approach also allowed analysis of *GYPA*, which belongs to a multigene family that expanded through duplication events about 9-13 million years ago (Ko et al. 2011). Therefore, orthologs of human *GYPA* can only be found in the chimpanzee and gorilla genomes.

gammaMap jointly uses intra-species variation and inter-specific diversity to estimate the distribution of selection coefficients (γ) along coding regions. For humans we exploited data from the 1000 Genomes Pilot Project for Europeans (CEU), Chinese plus Japanese (CHBJPT) and Yoruba (YRI) individuals (1000 Genomes Project Consortium *et al.* 2010). For chimpanzees and gorillas, we used SNP information from 25 and 23 apes, respectively (Prado-Martinez *et al.* 2013).

Analysis of the overall distribution of selection coefficients indicated a similar evolutionary pattern for both genes in the human and chimpanzee lineages (Fig. 3A). In these species purifying selection drove the evolution of a major proportion of sites in *BSG*, with a median γ in the range of -10 to -5 (moderately deleterious, Fig. 3A). Conversely, *GYPA* displayed median γ values of 5 to 10 (moderately beneficial, Fig. 3A). The gorilla lineage showed opposite patterns at the two genes with average γ values of 5 (moderately beneficial) and -5 (moderately deleterious) for *BSG* and *GYPA*, respectively (Fig. 3A). We next used gammaMap to identify specific codons evolving under positive selection (defined as those having a cumulative probability >0.75 of $\gamma > 0$) in each lineage. These codons may occur even if the majority of sites evolve under selective constraint. In fact, in *BSG*, 2 and 6 sites were found to

represent positive selection targets in chimpanzees and gorillas, respectively; no positively selected sites were found in the human lineage (Table S4, Supporting information). One positively selected site in the chimpanzee (S163) and two (L27 and K191) in the gorilla lineage are located at the interaction surface of human BSG with PfRH5 (Wright *et al.* 2014) (Fig. 1A-B). In particular, residue 191 was shown to play an important role in the interaction with PfRH5, as its mutation to the gorilla counterpart is sufficient to confer chimpanzee BSG the ability to bind PfRH5 (Wanaguru *et al.* 2013b). The L27 residue is known to reduce the binding of PfRH5 with human basigin when mutated to the gorilla counterpart (Wanaguru *et al.* 2013b).

Liu *et al.* 2014; Krief *et al.* 2010; Pacheco *et al.* 2013; Boundenga *et al.* 2015). Although cases of cross-species infection have been described - for example, chimpanzees and bonobos can acquire *P. falciparum* from humans (Krief *et al.* 2010; Pacheco *et al.* 2013) and *P. billcollinsi* infects both gorillas and chimpanzees (Boundenga *et al.* 2015) - *Laverania* infections display remarkable host-specificity in the wild (Rayner *et al.* 2011). Overall, these observations suggest that, over millions of years, primate hosts have co-evolved with their parasites. Herein we show that natural selection has shaped genetic diversity at the *BSG* gene in primates and at its parasite-encoded ligand, *PfRH5*. The two BSG residues previously shown to modulate the affinity of human BSG for PfRH5 were found to represent targets of positive selection either in the entire primate phylogeny or in specific primate lineages. Likewise, one of the two positively selected sites in PfRH5, W447, is known to stabilize the interaction with human BSG by packing into hydrophobic pockets on the protein (Wright *et al.* 2014). In PrRH5 position 447 is occupied by a polar residue (N447), suggesting very different properties at the contact surface. The evolutionary analysis of PfRH5 was nevertheless conducted using a limited number of sequences and one single ortholog. Under these circumstances, the power to detect selection is limited, especially when inference is extended both to the branch and to the sites targeted by selection. This might explain why the branch-site test, which is robust but lacks power (Zhang *et al.* 2005), failed to detect positively selected sites on the *P. falciparum* branch of the *RH5* phylogeny. Therefore, results from this analysis will need to be validated by further analysis, and the sequencing of additional orthologs from other *Plasmodium* species will be necessary to gain full insight into the evolutionary history of *RH5*.

Basigin is a multifunctional protein with a role in trophoblast function and in spermatogenesis, two processes that may be targeted by natural selection in primates (Nielsen *et al.* 2005). Therefore, although several selected sites are located at the PfRH5 binding interface, we cannot exclude that the evolution of BSG was driven by selective pressures unrelated to *Plasmodium* infection.

Previous analyses indicated that positive selection in primates also targeted DARC (Duffy antigen/receptor for chemokines), which is used by *P. vivax* and *P. knowlesi* for erythrocyte invasion, with most selected sites located in the region interacting with the parasite-encoded ligand (Demogines *et al.* 2012). On the one hand, these data testify to the long-lasting interaction between *Plasmodium* parasites and primates; on the other hand, *Laverania* infection in great apes is generally thought to result in no or very weak pathology, although transitory malaria-like

Table 1 Likelihood ratio test (LRT) statistics for models of variable selective pressure among sites in *BSG*

Selection model	Degrees of freedom	$-2\Delta\ln L^{\dagger}$	*P* value	% of sites (average dN/dS)	Positively selected sites (BEB and MEME)
M1a vs. M2a[‡]	2	8.52	0.014	2.0% (3.2)	
M7 vs. M8[§]	2	10.89	0.004	4.0% (2.5)	F27, H102
M8a[¶] vs. M8	1	8.66	0.003		

[†]2ΔlnL: twice the difference of the natural logs of the maximum likelihood of the models being compared.
[‡]M1a is a nearly neutral model that assumes one ω class between 0 and 1, and one class with ω = 1; M2a (positive selection model) is the same as M1a plus an extra class of ω >1.
[§]M7 is a null model that assumes that 0 < ω < 1 is beta distributed among sites; M8 (positive selection model) is the same as M7 but also includes an extra category of sites with ω > 1.
[¶]M8a is the same as M8, except that the 11th category cannot allow positive selection, but only neutral evolution.

Table 2 Tests of episodic positive selection in the *Plasmodium falciparum* and *P. praefalciparum* lineages

Gene/foreground branch	Model	$-2\Delta\ln L^{\dagger}$	*P* value	Positively selected sites
RH5				
Plasmodium falciparum	MA vs. MA1[‡]	11.99	0.0005	–
	BUSTED[§]	7.51	0.0236	Y190, W447
EBA175				
Plasmodium falciparum	MA vs. MA1	2.6×10^{-5}	0.996	–
	BUSTED	0	1	
Plasmodium falciparum and *praefalciparum* node	MA vs. MA1	0.003	0.954	–
	BUSTED	0	1	

[†]2ΔlnL: twice the difference of the natural logs of the maximum likelihood of the models being compared.
[‡]MA and MA1 are branch-site models that assume four classes of sites: the MA model allows a proportion of codons to have $\omega \geq 1$ on the foreground branches, whereas the MA1 model does not.
[§]BUSTED analysis uses the designated branches as foreground.

symptoms were recently described in a young chimpanzee (Herbert *et al.* 2015). Also, recent evidence indicated that parasite detection increases during pregnancy in *Pan troglodytes*, either because of increased susceptibility or of higher parasitemia (De Nys *et al.* 2014). Nevertheless, it is presently unknown whether *Plasmodium* infection has an adverse effect on pregnancy outcome in chimpanzees and, more generally, if *Laverania* imposed a fitness cost in great apes. It is also worth noting that it is presently unknown which other *Laverania* species carry a functional *RH5* ortholog because the RBL family is highly dynamic (Otto *et al.* 2014). Moreover, whereas the interaction between human BSG and PfRH5 occurs through protein-protein contacts (Wright *et al.* 2014), PfRH5 variants from distinct strains display different affinity for erythrocytes of New World monkeys depending on the presence of surface sialic acids (Hayton *et al.* 2013). Hence, it is not clear if such PfRH5 mutants depend on BSG for invasion or whether they acquired the ability to exploit other cellular receptors. By analogy, other field strains or *Plasmodium* species may express RH5 molecules or, more generally, RBL proteins with affinity for host receptors distinct from BSG, as is the case for PfRH4, which binds human CR1 (complement receptor type 1). Whereas these points will require further investigation, data herein suggest that BSG-binding *Plasmodium* parasites have been circulating for a long time among wild primate populations and support the notion that positive selection at the BSG-RH5 interacting partners contribute to host shifts. With respect to the more recent events that led to the emergence of *P. falciparum* as a human pathogen, gammaMap analysis suggests that its gorilla-infecting ancestor encoded a functional *RH5* gene and exerted a strong selective pressure on *BSG*. Positive selection at the gorilla *BSG* and at the parasite *RH5* genes most likely contributed to the shift to human hosts. It is worth mentioning here that previous *in vitro* assays that analyzed the interaction between primate BSG and PfRH5 used expression vectors based on the chimpanzee and gorilla reference genomes, which contain either sequence errors or rare variants (Wanaguru *et al.* 2013b). Whereas the results of the single-site mutagenesis remain mostly valid (e.g. for sites F27L, E191K, and the H103 insertion), in some cases their biological significance should be revised; for instance the Q100K substitution, which decreases PfRH5 binding affinity (Wanaguru *et al.* 2013b), does not correspond to the replacement of a human residue with the gorilla counterpart, as most gorillas are likely to carry a glutamine, as well. Most, importantly, the binding affinity of PfRH5 for the wild-type gorilla and chimpanzee BSG proteins should be re-evaluated. We found no evidence of selection for *EBA175* in the *P. falciparum* and *praefalciparum* lineages. We note that, due to limited sequence information availability, our analysis only coveredthe RII domain. Recently, Wanaguru and coworkers (Wanaguru *et al.* 2013a) demonstrated that protein regions other than the RII domain are important for interaction with GYPA. In fact, binding of EBA175 to GYPA is thought to occur in two steps: after a sialic-dependent recognition, proteolytic cleavage generates a 65-kD EBA175 fragment that binds the glycophorin backbone (Kain *et al.* 1993). This fragment includes a dimorphic region

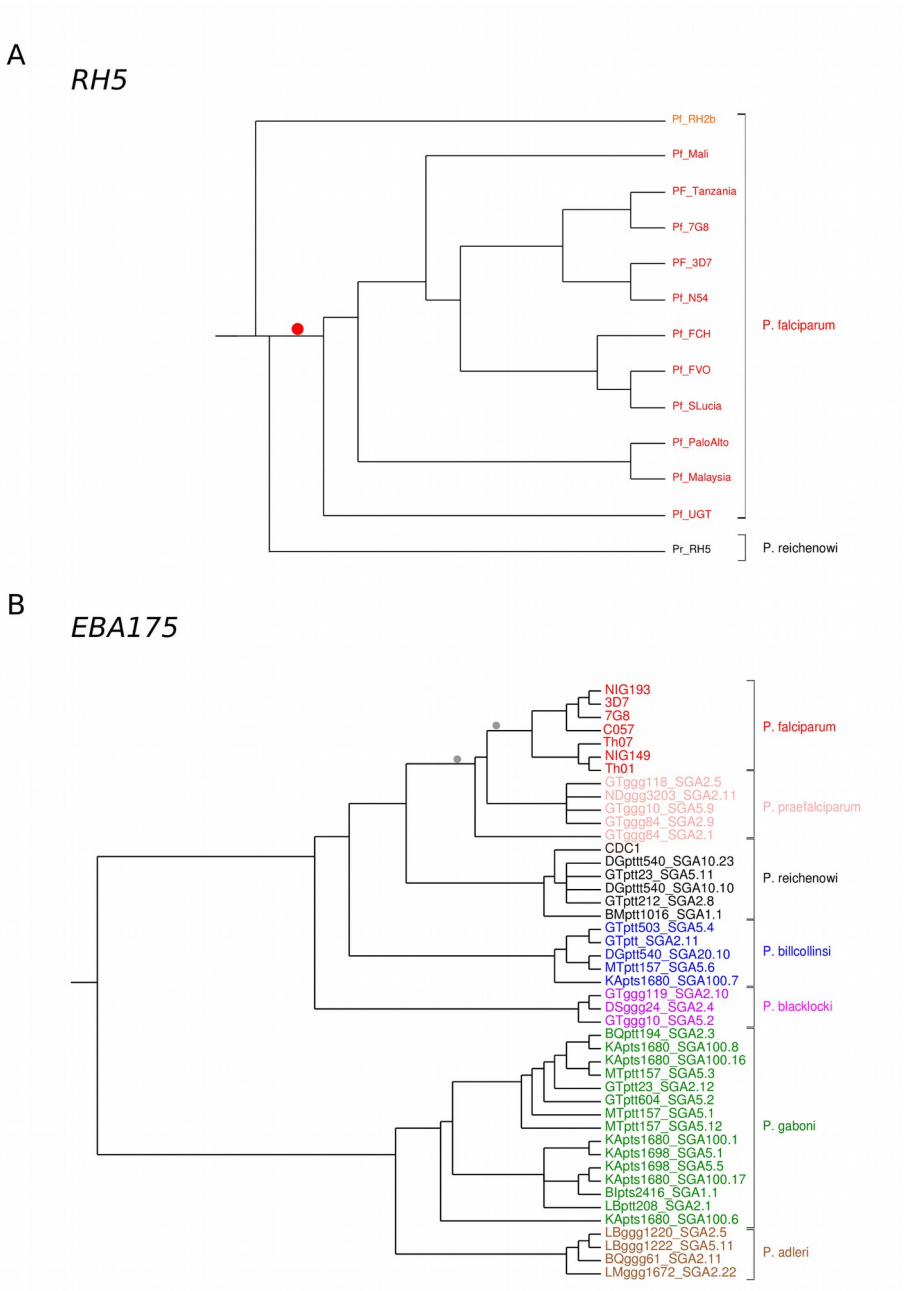


Figure 2. Phylogenetic trees of *Laverania* subgenus. Phylogenetic tree of (A) *RH5* sequences from *P. falciparum* and *P. reichenowi* and (B) the receptor binding region (RII) of *EBA175* strains (Wanaguru *et al.* 2013b). Branches set as foreground lineages are marked with dots: red indicates statistically significant evidence of positive selection.

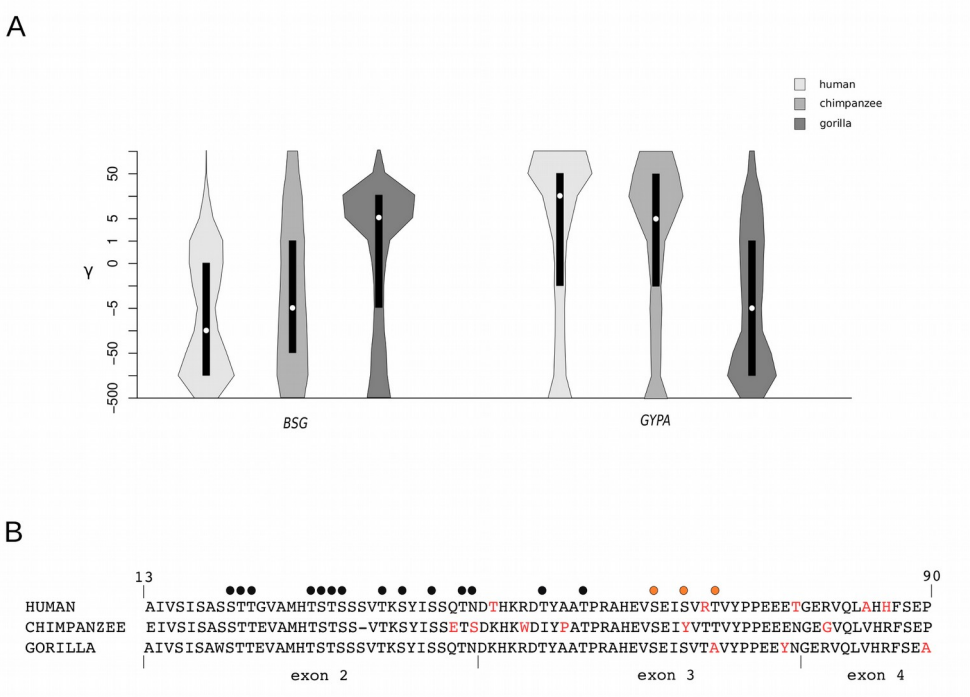


Figure 3. Analysis of selective pressure in the Homininae lineages. (A) Violin plot of selection coefficients (median, white dot; interquartile range, black bar) for *BSG* and *GYPA* genes. Selection coefficients (γ) are classified as strongly beneficial (100, 50), moderately beneficial (10, 5), weakly beneficial (1), neutral (0), weakly deleterious (−1), moderately deleterious (−5, −10), strongly deleterious (−50, −100), and inviable (−500). (B) Multiple sequence alignment of GYPA exons 2-4. Sites with a cumulative probability >0.75 of γ > 0 are shown in red, glycosylation sites are marked with a black dot. Orange dots indicated glycosylation sites which, when mutated, affect PfEBA175 binding (Salinas *et al.* 2014).

characterized by the presence of two alternative and divergent amino acid sequences, the so-called F- and C- segments (after the strains they were initially isolated from, FCR3 and CAMP, respectively), with different size and location in the *EBA175* coding region (Kain *et al.* 1993; Ware *et al.* 1993). *PfEBA175* genes carrying either the F- or C-segment alleles have been detected in all *P. falciparum* isolates analyzed to date (Ware *et al.* 1993; Cramer *et al.* 2004; Ahouidi *et al.* 2010), although it is presently unknown whether the encoded proteins display differential efficiency for GYPA binding. Irrespective of the possible role of this dimorphic region, these observations indicate that selection may target EBA175 residues outside the RII region and, if so, would go undetected in our study. Also, the analyses we performed were not devised to detect intra-species selection in *P. falciparum*.

Baum and coworkers (Baum *et al.* 2002) previously analyzed 30 P*fEBA175* RII sequences isolated in Nigeria from infected subjects and concluded that the region was a target of diversifying selection, highlighting its importance for erythrocyte invasion and, possibly, as a target of the host immune response. Herein we did not analyze the recent evolutionary events driving diversity in *P. falciparum* populations, but focused on the changes that contributed to the emergence of *P. falciparum* as a human pathogen. We conclude that changes in the RII region of EBA175 played no or minor role in determining host tropism, in line with the observation that *EBA175* orthologs from chimpanzee-restricted *Laverania* parasites encode proteins that bind human GYPA with the same affinity as PfEBA175 (Wanaguru *et al.* 2013b).

The binding of PfEBA175 to GYPA is sialic-acid dependent and the parasite ligand engages

several sialoglycans (Salinas *et al.* 2014). We found *GYPA* to display several positively selected sites in the three primate lineages we analyzed, in line with previous analyses (Baum *et al.* 2002; Ko *et al.* 2011). Selected codons also included glycosylation sites that modulate PfEBA175 binding affinity (Salinas *et al.* 2014), suggesting that the selective pressure imposed by *Laverania* parasites contributed to shaping diversity at GYPA. Although large uncertainty affects time estimates, the timing of the glycophorin gene family expansion and the time of origin for *Laverania* species broadly correspond (Krief *et al.* 2010; Pacheco *et al.* 2013), suggesting long-lasting interactions between parasite ligands and primate glycophorins. Nonetheless, GYPA is bound by several other pathogens including common viruses and bacteria; these also exploit complex sugars for binding (Gagneux & Varki 1999). This observation led to the suggestion that GYPA may act as a decoy receptor that attracts pathogens to erythrocytes where infection is unproductive due to the absence of a nucleus. Other authors (Baum *et al.* 2002; Ko *et al.* 2011) favored an "evasion" hypothesis as opposed to the decoy function, by placing the emphasis on the role of glycophorins as receptors for *Plasmodium* parasites. While these two views of glycophorin evolution are not mutually exclusive and remain difficult to test, we limit our conclusions to the observation that pathogens unrelated to *Laverania* may have acted as selective pressures on GYPA. Finally, we note that gene conversion has been shown to contribute significantly to the shaping of nucleotide diversity at glycophorin genes, due to the high sequence homology among family members (Ko *et al.* 2011). We cannot therefore exclude that a portion of the selected sites we detected may indeed result from gene conversion events rather than selection.

In summary, data herein provide an evolutionary explanation for species-specific binding of the PfRH5-BSG ligand-receptor pair and support the view that natural selection at these genes underlies the host shift that gave rise to *P. falciparum*, a major human pathogen. These observations are in line with recent *in vitro* data (Wanaguru *et al.* 2013b) and exemplify how evolutionary studies can provide insight into infectious diseases emergence and zoonotic transmission (Sironi *et al.* 2015).

Acknowledgments

CP is supported by a fellowship of the Doctorate School of Molecular Medicine, University of Milan**.**

Data Accessibility

DNA sequences accession numbers are reported in Tables S2-S3, Supporting information.

Alignments and trees for inter-species analyses and input files for gammaMap: Dryad provisional DOI:doi:10.5061/dryad.3g23m

Author Contributions

MS and MC conceived the study; DF and CP performed the analysis and analyzed the data; RC and UP provided support during the analyses; CP and DF produced the figures; MS and DF wrote the manuscript, with critical input from MC and from the remaining authors.

References

1000 Genomes Project Consortium, Durbin RM, Abecasis GR, et al (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061-1073.

Ahouidi AD, Bei AK, Neafsey DE, et al (2010) Population genetic analysis of large sequence polymorphisms in Plasmodium falciparum blood-stage antigens. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, **10**, 200-206.

Anisimova M, Bielawski JP, Yang Z (2002) Accuracy and power of bayes prediction of amino acid sites under positive selection. *Molecular biology and evolution*, **19**, 950-958.

Baum J, Ward RH, Conway DJ (2002) Natural selection on the erythrocyte surface. *Molecular biology and evolution*, **19**, 223-229.

Boundenga L, Ollomo B, Rougeron V, et al (2015) Diversity of malaria parasites in great apes in Gabon. *Malaria journal*, **14**, 111-015-0622-6.

Capella-Gutierrez S, Silla-Martinez JM,

Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)*, **25**, 1972-1973.

Cramer JP, Mockenhaupt FP, Mohl I, et al (2004) Allelic dimorphism of the erythrocyte binding antigen-175 (eba-175) gene of Plasmodium falciparum and severe malaria: Significant association of the C-segment with fatal outcome in Ghanaian children. *Malaria journal*, **3**, 11.

Crosnier C, Bustamante LY, Bartholdson SJ, et al (2011) Basigin is a receptor essential for erythrocyte invasion by Plasmodium falciparum. *Nature*, **480**, 534-537.

De Nys HM, Calvignac-Spencer S, Boesch C, et al (2014) Malaria parasite detection increases during pregnancy in wild chimpanzees. *Malaria journal*, **13**, 413-2875-13-413.

Delport W, Poon AF, Frost SD, Kosakovsky Pond SL (2010) Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics (Oxford, England)*, **26**, 2455-2457.

Demogines A, Truong KA, Sawyer SL (2012) Species-specific features of DARC, the primate receptor for Plasmodium vivax and Plasmodium knowlesi. *Molecular biology and evolution*, **29**, 445-449.

Gagneux P, Varki A (1999) Evolutionary considerations in relating oligosaccharide diversity to biological function. *Glycobiology*, **9**, 747-755.

Guindon S, Delsuc F, Dufayard JF, Gascuel O (2009) Estimating maximum likelihood phylogenies with PhyML. *Methods in molecular biology (Clifton, N.J.)*, **537**, 113-137.

Hayton K, Dumoulin P, Henschen B, Liu A, Papakrivos J, Wellems TE (2013) Various PfRH5 polymorphisms can support Plasmodium falciparum invasion into the erythrocytes of owl monkeys and rats. *Molecular and biochemical parasitology*, **187**, 103-110.

Hayton K, Gaur D, Liu A, et al (2008) Erythrocyte binding protein PfRH5 polymorphisms determine species-specific pathways of Plasmodium falciparum invasion. *Cell host & microbe*, **4**, 40-51.

Herbert A, Boundenga L, Meyer A, et al (2015) Malaria-like symptoms associated with a natural Plasmodium reichenowi infection in a chimpanzee. *Malaria journal*, **14**, 220-015-0743-y.

Kain KC, Orlandi PA, Haynes JD, Sim KL, Lanar DE (1993) Evidence for two-stage binding by the 175-kD erythrocyte binding antigen of Plasmodium falciparum. *The Journal of experimental medicine*, **178**, 1497-1505.

Ko WY, Kaercher KA, Giombini E, et al (2011) Effects of natural selection and gene conversion on the evolution of human glycophorins coding for MNS blood polymorphisms in malaria-endemic African populations. *American Journal of Human Genetics*, **88**, 741-754.

Kosakovsky Pond SL, Frost SD (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular biology and evolution*, **22**, 1208-1222.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD (2006) Automated phylogenetic detection of recombination using a genetic algorithm. *Molecular biology and evolution*, **23**, 1891-1901.

Krief S, Escalante AA, Pacheco MA, et al (2010) On the diversity of malaria parasites in African apes and the origin of Plasmodium falciparum from Bonobos. *PLoS pathogens*, **6**, e1000765.

Liu W, Li Y, Learn GH, et al (2010) Origin of the human malaria parasite Plasmodium falciparum in gorillas. *Nature*, **467**, 420-425.

Liu W, Li Y, Shaw KS, et al (2014) African origin of the malaria parasite Plasmodium vivax. *Nature communications*, **5**, 3346.

Martin MJ, Rayner JC, Gagneux P, Barnwell JW, Varki A (2005) Evolution of human-chimpanzee differences in malaria susceptibility: relationship to human genetic loss of N-glycolylneuraminic acid. *Proceedings of the National Academy of*

Sciences of the United States of America, **102**, 12819-12824.

McBee RM, Rozmiarek SA, Meyerson NR, Rowley PA, Sawyer SL (2015) The effect of species representation on the detection of positive selection in primate gene data sets. *Molecular biology and evolution*, **32**, 1091-1096.

Murrell B, Weaver S, Smith MD, et al (2015) Gene-Wide Identification of Episodic Selection. *Molecular biology and evolution*, **32,** 1365-71.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL (2012) Detecting individual sites subject to episodic diversifying selection. *PLoS genetics*, **8**, e1002764.

Nielsen R, Bustamante C, Clark AG, et al (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS biology*, **3**, e170.

Orlandi PA, Sim BK, Chulay JD, Haynes JD (1990) Characterization of the 175-kilodalton erythrocyte binding antigen of Plasmodium falciparum. *Molecular and biochemical parasitology*, **40**, 285-294.

Otto TD, Rayner JC, Bohme U, et al (2014) Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nature communications*, **5**, 4754.

Pacheco MA, Cranfield M, Cameron K, Escalante AA (2013) Malarial parasite diversity in chimpanzees: the value of comparative approaches to ascertain the evolution of Plasmodium falciparum antigens. *Malaria journal*, **12**, 328-2875-12-328.

Pond SL, Frost SD, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics (Oxford, England)*, **21**, 676-679.

Pozzoli U, Fumagalli M, Cagliani R, et al (2010) The role of protozoa-driven selection in shaping human genetic variability. *Trends in genetics*, **26,** 95-9.

Prado-Martinez J, Sudmant PH, Kidd JM, et al (2013) Great ape genetic diversity and population history. *Nature*, **499**, 471-475.

Quach H, Wilson D, Laval G, et al (2013) Different selective pressures shape the evolution of Toll-like receptors in human and African great ape populations. *Human molecular genetics*, **22**, 4829-4840.

Rayner JC, Liu W, Peeters M, Sharp PM, Hahn BH (2011) A plethora of Plasmodium species in wild apes: a source of human infection? *Trends in parasitology*, **27**, 222-229.

Salinas ND, Paing MM, Tolia NH (2014) Critical glycosylated residues in exon three of erythrocyte glycophorin a engage Plasmodium falciparum EBA-175 and define receptor specificity. *mBio*, **5**, e01606-14.

Schlegel J, Redzic JS, Porter CC, et al (2009) Solution characterization of the extracellular region of CD147 and its interaction with its enzyme ligand cyclophilin A. *Journal of Molecular Biology*, **391**, 518-535.

Sim BK, Chitnis CE, Wasniowska K, Hadley TJ, Miller LH (1994) Receptor and ligand domains for invasion of erythrocytes by Plasmodium falciparum. *Science (New York, N.Y.)*, **264**, 1941-1944.

Sironi M, Cagliani R, Forni D, Clerici M (2015) Evolutionary insights into host-pathogen interactions from mammalian sequence data. *Nature reviews.Genetics*, **16**, 224-236.

Wanaguru M, Crosnier C, Johnson S, Rayner JC, Wright GJ (2013a) Biochemical analysis of the Plasmodium falciparum erythrocyte-binding antigen-175 (EBA175)-glycophorin-A interaction: implications for vaccine design. *The Journal of biological chemistry*, **288**, 32106-32117.

Wanaguru M, Liu W, Hahn BH, Rayner JC, Wright GJ (2013b) RH5-Basigin interaction plays a major role in the host tropism of Plasmodium falciparum. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 20735-20740.

Ware LA, Kain KC, Lee Sim BK, Haynes JD, Baird JK, Lanar DE (1993) Two alleles of the 175-kilodalton Plasmodium falciparum erythrocyte binding antigen. *Molecular and biochemical parasitology*, **60**, 105-109.

Wernersson R, Pedersen AG (2003) RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic acids research*, **31**, 3537-3539.

Wilson DJ, Hernandez RD, Andolfatto P, Przeworski M (2011) A population genetics-phylogenetics approach to inferring natural selection in coding sequences. *PLoS genetics*, **7**, e1002395.

Wright KE, Hjerrild KA, Bartlett J, et al (2014) Structure of malaria invasion protein RH5 with erythrocyte basigin and blocking antibodies. *Nature*, **515**, 427-430.

Xie SS, Huang CH, Reid ME, Blancher A, Blumenfeld OO (1997) The glycophorin A gene family in gorillas: structure, expression, and comparison with the human and chimpanzee homologues. *Biochemical genetics*, **35**, 59-76.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, **24**, 1586-1591.

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences : CABIOS*, **13**, 555-556.

Yang Z, Wong WS, Nielsen R (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular biology and evolution*, **22**, 1107-1118.

Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular biology and evolution*, 22, 2472-2479.

4. CONCLUSIONS

Evolutionary analyses and the identification of natural selection patterns in different species, have manifold purposes.

In the first study, for example, the availability of genetic information of other primates and of preagricultural human population, allowed the opportunity to address the tempo and mode of evolution for genes involved in carbohydrate digestion and absorption. In this case I found that some non-coding positively selected variants identified in human populations are “modern alleles” when Neanderthals and Denisovans are taken as a comparison, but not as modern as to unequivocally support their agriculture-driven spread. The introduction of agriculture might have spurred the frequency increase of variants that were already weakly adaptive in hunter-gatherers, resulting in a continuum rather than an abrupt onset of selective events. These data have a relevance in explaining the high prevalence of diabetes and obesity seen in modern populations. When applied to the study of host-pathogen conflict, evolutionary analyses may help to gain insight into the genetic determinants modulating the susceptibility to infectious diseases and the virulence of pathogens. In turn, this knowledge may be translated into novel treatment strategies in terms of molecular targets for drug development.

Inter-specific and population genetic-phylogenetic analyses allowed the identification of specific regions and sites evolving under positive selection: in the case of *ADAR* genes (genes involved in A-to-I RNA editing), for instance, I found that most positively selected sites are located in the additional amino-terminal portion of the interferon-induced isoform of ADAR (p150). One of these, in particular, is located within the nuclear exporting sequence (NES). Cytoplasmic localization of ADARp150 may be relevant to viral detection and binding, as well to stress responses. This confirms the

idea that regions under positive selection are the ones directly related with immune responses.

As for *MX2* (myxovirus resistance 2), five of the positively selected sites were located in an unstructured loop (loop 4) where several positively selected sites were also detected in primate *MX1* genes. In this letter case, the positively selected sites explain species-specific difference in antiviral activity against orthomyxoviruses. Interestingly, one of the positive selected sites detected in MX2, S518, corresponds to a polymorphic position in the pig coding sequence: amino acid variation at this residue was shown to modulate the ability of the porcine MX2 protein to inhibit vescicular stomatitis virus replication, supporting the hypothesis that positively selected sites determine antiviral activity.

This work also clearly shows how evolutionary analysis could be exploited to identify novel host determinant of HIV-1 infection susceptibility. Indeed, population genetic analyses on *MX2* demonstrated that distinct selective events drove the frequency increase of two haplotypes in populations of Asian and European ancestry: the Asian haplotype carries a susceptibility allele for melanoma; the European haplotype is tagged by rs2074560, an intronic variant. Because natural selection targets variants are expected to entail a phenotypic effect, and given the role of MX2 as a HIV-1 restriction factor, I investigated whether the allelic status at rs2074560 modulates susceptibility to HIV-1 infection. Genotyping analyses performed on three independent European cohorts of HIV-1 exposed seronegative individuals with different geographic origin and distinct exposure route showed that the ancestral (G) allele protects from HIV-1 infection with a recessive effect. The same allele is associated with lower *in vitro* HIV-1 replication and increases MX2 expression levels in response to IFN-α. Results herein establish, therefore, a possible target for therapeutic intervention in HIV-1

treatment and prevention. They also show that evolutionary analysis can identify susceptibility alleles for modern human mutations.

Still in line with the tenet that natural selection targets functionally relevant residues, position 230 in STING (stimulator of interferon genes) was found to be positively selected in both the primate and bat phylogenies. This site lies in the flexible loop that acts as a lid above the cyclic di-nucleotide binding pocket of the receptor. Substitutions at this site greatly affect the response to natural ligands and to mimetic drugs, such as DMXAA. Different aminoacid residues at position 230 were also shown to be responsible for the species-specific differences in the induction of the type I interferon pathway in response to DMXAA in human and mouse. DMXAA is a mimetic drug showed promising antitumor effects in mice, but failed in human clinical trials because the human protein does not bind to or signal in response to DMXAA. Therefore, evolutionary analyses of immune response genes may also provide valuable information on the molecular determinants underlying species-specific infection susceptibility and may clarify the differential response to natural or synthetic molecules.

Under a genetic conflict scenario whereby the host and the pathogen genes evolve within binding avoidance-binding seeking dynamics, evolutionary analyses may predict host-pathogen interaction surfaces at the single amino acid resolution. That was clearly shown in the analyses of the evolutionary history of the complement system and of bacterial-encoded complement-interacting proteins. In this case, evolutionary analyses predicted host-pathogen interaction surfaces at the single amino acid resolution. In the CFH-OspE interaction (Complement factor H and *Borrelia burgdorferi* surface protein, respectively), for example, both partners are targeted by selection. In OspE, two positively selected sites (V120 and I121) are located at the binding interface and this is the same for

the positively selected sites in CFH (e.g. S1196 interacts with S82 in OspE). Mutations V120L and I121A in OspE, as well as T1184K and S1196P in CFH, led to a misplacement of the interactors in a docking analysis, strongly suggesting that they affect protein-protein interaction.

Studying the pattern of inter-species evolution may provide valuable information on the differential susceptibility to infection within and among species. Primates for example, show marked differences in the susceptibility and severity of several viral infections including those caused by HIV/SIV. Moreover, several emerging and re-emerging viral diseases affecting humans originate through the zoonotic transmission from a reservoir animal host: evolutionary analysis help identify the most likely reservoirs of zoonotic pathogens. Examples of pathogen spillover events include *Ebola virus* and *Plasmodium falciparum*.

Uncertainty still exists about the reservoir host(s) for filoviruses, and it is presently unknown whether the spillover is initiated from a direct contact with the reservoir or rather from exposure to other wild-life that also contracted the infection from the reservoir. The exploration of the presence and extent of positively selection in *NPC1* (Niemann-Pick disease, type C1 membran protein), led to identify distinct codons targeted by selection in different mammalian orders/superorders or clades, most of them within or at the base of loops previously shown to be involved in the interaction with filovirus glycoprotein (GP). This data indicate that positive selection has been driving the molecular evolution at the host-filovirus interaction surface and that most mammals are or were infected by these pathogens; the filovirus reservoir(s) may thus belong to any mammalian order.

Plasmodium falciparum, the causative agent of most malaria-related deaths in humans, instead, originated in gorillas, although infection is now restricted to humans. A population genetics-phylogenetics approach applied

on basigin (BSG) gene, the erytrocyte cell-surface receptor of *Plasmodium falciparum* reticulocyte-binding protein homologue 5 (PfRH5), detected the strongest selection for the gorilla lineage, suggesting that gorilla-infecting *Plasmodium* ancestor encoded a functional *RH5* gene and exerted a strong selective pressure on *BSG*. Positive selection at the gorilla *BSG* and at the parasite *RH5* genes most likely contributed to the shift to human hosts. In particular, one of the positively selected sites (K191) is a major determinant of PfRH5 binding affinity: its mutation to the gorilla counterpart is sufficient to confer chimpanzee BSG the ability to bind PfRH5. This data demonstrate that adaptive changes in BSG-RH5 interacting partners contributed to the host shift that gave rise to *P. falciparum* as a major human pathogen. Therefore, by providing information on the location and nature of functional genetic variation, evolutionary analysis may help explain changes in pathogen tropism.

Finally, evolutionary analysis can provide information on possible therapeutic targets. Investigation of Ebola GP sequences identified positively selected sites that resides within epitopes for antibodies: residue 484 is within the epitope for 14G7 antibody and site 39 is located within an epitope bound by the 16F6 antibody that specifically neutralizes Sudan strains. Furthermore, one pair of co-evolving sites (309 and 509) was also detected. Residue 509 is at the direct contact interface with the KZ52 antibody isolated from a human survivor of Ebola infection, whereas residue 309 is located in the glycan cap, a region bound by neutralizing antibodies. It is worth noting that mutations at the flanking 508 residue affect the binding of three different antibodies against EBOV GP, and Q508R escape mutants are associated with lethality in monkeys treated with a combination of three neutralizing mAbs (ZMAb). These results point to the host humoral immune response as a major selective pressure on

filovirus GP.

Because promising treatment strategies for filovirus infection are based on antibody combinations, evolutionary studies suggest from one hand caution in their use, since antigenic variability should pose a serious concern to their effectiveness in the long-term. On the other, they identify in the less variable regions the most likely targets for these treatments.

In conclusion, these works highlight the importance of applying evolutionary approaches. Indeed, evolutionary analyses can provide information not only on the past adaptive events that shaped human-specific traits, but also on the presence and location of functional genetic variants with particular relevance to phenotypic diversity and, ultimately, to human health.

Moreover, the experimental design (based on the integration between inter and intra-species analyses, and between host and pathogen's interacting partners) used in my works may serve as a model for the analysis of other host-pathogen interactors.

My future perspective is to focus on host-pathogen interactions, the strongest selective pressure. In particular I'm going to integrate evolutionary analysis with the availability of an ever increasing number of mammalian and pathogen genomes and with an increasing number of crystal structures of interacting host and pathogen partners to search for selective pressures acting on host and emergence pathogen interactions.

I also envisage experimental validation of the results from evolutionary analysis using the two hybrid system, a method used to assay for protein-protein interactions and protein-DNA interactions. The final aim is to clarify the mechanisms underling the susceptibility to and the severity of the diseases caused by emergence pathogens, as well as to provide a list of proteins/regions that may be targets for therapeutic preventive strategies.

5. REFERENCES

1. Vitti J. J., Grossman S. R., Sabeti P. C., "Detecting natural selection in genomic data", *Annu Rev Genet,* Vol. 47, 2013, pp. 97-120.

2. Fumagalli M., Sironi M., "Human genome variability, natural selection and infectious diseases", *Curr Opin Immunol,* Vol. 30, 2014, pp. 9-16.

3. Waldman Y. Y., Tuller T., Keinan A., Ruppin E., "Selection for translation efficiency on synonymous polymorphisms in recent human evolution", *Genome Biol Evol,* Vol. 3, 2011, pp. 749-761.

4. Fischer K., Bot A. N., Zwaan B. J., Brakefield P. M., "Genetic and environmental sources of egg size variation in the butterfly bicyclus anynana", *Heredity (Edinb),* Vol. 92, no. 3, 2004, pp. 163-169.

5. Saastamoinen M., "Heritability of dispersal rate and other life history traits in the glanville fritillary butterfly", *Heredity (Edinb),* Vol. 100, no. 1, 2008, pp. 39-46.

6. Loewe L., Hill W. G., "The population genetics of mutations: Good, bad and indifferent", *Philos Trans R Soc Lond B Biol Sci,* Vol. 365, no. 1544, 2010, pp. 1153-1167.

7. Masel J., "Genetic drift", *Curr Biol,* Vol. 21, no. 20, 2011, pp. R837-8.

8. 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.

9. Karlsson E. K., Kwiatkowski D. P., Sabeti P. C., "Natural selection and infectious disease in human populations", *Nat Rev Genet,* Vol. 15, no. 6, 2014, pp. 379-393.

10. Barreiro L. B., Quintana-Murci L., "From evolutionary genetics to human immunology: How selection shapes host defence genes", *Nat Rev Genet,* Vol. 11, no. 1, 2010, pp. 17-30.

11. Donaldson P, Daly A, Ermini L, Bevitt D. (2016) Genetics of complex disease. New York, United States: New York, NY : Garland Science/Taylor & Francis Group, LLC. 406 pages p.

12. Sironi M., Cagliani R., Forni D., Clerici M., "Evolutionary insights into host-pathogen interactions from mammalian sequence data", *Nat Rev Genet,* Vol. 16, no. 4, 2015, pp. 224-236.

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

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

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

16. Murrell B., Wertheim J. O., Moola S., Weighill T., Scheffler K., Kosakovsky Pond S. L., "Detecting individual sites subject to episodic diversifying selection", *PLoS Genet,* Vol. 8, no. 7, 2012, pp. e1002764.

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

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

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

20. International HapMap Consortium, "A haplotype map of the human genome", *Nature,* Vol. 437, no. 7063, 2005, pp. 1299-1320.

21. Li B., Leal S. M., "Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data", *Am J Hum Genet,* Vol. 83, no. 3, 2008, pp. 311-321.

22. 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.

23. Nei M. (1987) Molecular evolutionary genetics. New York, NY, USA: Columbia University Press.

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

25. Bowcock A. M., Kidd J. R., Mountain J. L., Hebert J. M., Carotenuto L., Kidd K. K., Cavalli-Sforza L. L., "Drift, admixture, and selection in human evolution: A study with DNA polymorphisms", *Proc Natl Acad Sci U S A,* Vol. 88, no. 3, 1991, pp. 839-843.

26. Barreiro L. B., Laval G., Quach H., Patin E., Quintana-Murci L., "Natural selection has driven population differentiation in modern humans", *Nat Genet,* Vol. 40, no. 3, 2008, pp. 340-345.

27. Vatsiou A. I., Bazin E., Gaggiotti O. E., "Detection of selective sweeps in structured populations: A comparison of recent methods", *Mol Ecol,* Vol. 25, no. 1, 2016, pp. 89-103.

28. Quintana-Murci L., "Genetic and epigenetic variation of human populations: An adaptive tale", *C R Biol,* Vol. 339, no. 7-8, 2016, pp. 278-283.

29. Karasov W. H., Martinez del Rio C., Caviedes-Vidal E., "Ecological physiology of diet and digestive systems", *Annu Rev Physiol,* Vol. 73, 2011, pp. 69-93.

30. Luca F., Perry G. H., Di Rienzo A., "Evolutionary adaptations to dietary changes", *Annu Rev Nutr,* Vol. 30, 2010, pp. 291-314.

31. Clark A. G., Glanowski S., Nielsen R., Thomas P. D., Kejariwal A., Todd M. A., Tanenbaum D. M., Civello D., Lu F., Murphy B., et al, "Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios", *Science,* Vol. 302, no. 5652, 2003, pp. 1960-1963.

32. Nielsen R., "Molecular signatures of natural selection", *Annu Rev Genet,* Vol. 39, 2005, pp. 197-218.

33. 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.

34. Helgason A., Palsson S., Thorleifsson G., Grant S. F., Emilsson V., Gunnarsdottir S., Adeyemo A., Chen Y., Chen G., Reynisdottir I., et al, "Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution", *Nat Genet,* Vol. 39, no. 2, 2007, pp. 218-225.

35. Hu J. C., Yamakoshi Y., Yamakoshi F., Krebsbach P. H., Simmer J. P., "Proteomics and genetics of dental enamel", *Cells Tissues Organs,* Vol. 181, no. 3-4, 2005, pp. 219-231.

36. Hu C. C., Hart T. C., Dupont B. R., Chen J. J., Sun X., Qian Q., Zhang C. H., Jiang H., Mattern V. L., Wright J. T., et al, "Cloning human enamelin cDNA, chromosomal localization, and analysis of expression during tooth development", *J Dent Res,* Vol. 79, no. 4, 2000, pp. 912-919.

37. Mardh C. K., Backman B., Holmgren G., Hu J. C., Simmer J. P., Forsman-Semb K., "A nonsense mutation in the enamelin gene causes local hypoplastic autosomal dominant amelogenesis imperfecta (AIH2)", *Hum Mol Genet,* Vol. 11, no. 9, 2002, pp. 1069-1074.

38. Gurven M., Kaplan H., "Longevity among hunter- gatherers: A cross-cultural examination", *Population and Development Review,* Vol. 33, no. 2, 2007, pp. 321-365.

39. Chivers D. J., "GoodallJ.1986. the chimpanzees of gombe: Patterns of behavior. harvard university press, cambridge (massachusetts). 673 pages. ISBN 0-674-11649-6. price: \pounds19.95 (hardback).", *J Trop Ecol,* Vol. 3, no. 2, 1987, pp. 190-191.

40. Finch C. E., "Evolution in health and medicine sackler colloquium: Evolution of the human lifespan and diseases of aging: Roles of infection, inflammation, and nutrition", *Proc Natl Acad Sci U S A,* Vol. 107 Suppl 1, 2010, pp. 1718-1724.

41. Taylor L. H., Latham S. M., Woolhouse M. E., "Risk factors for human disease emergence", *Philos Trans R Soc Lond B Biol Sci,* Vol. 356, no.

1411, 2001, pp. 983-989.

42. Allocati N., Petrucci A. G., Di Giovanni P., Masulli M., Di Ilio C., De Laurenzi V., "Bat-man disease transmission: Zoonotic pathogens from wildlife reservoirs to human populations", *Cell Death Discov,* Vol. 2, 2016, pp. 16048.

43. Ludwig B., Kraus F. B., Allwinn R., Doerr H. W., Preiser W., "Viral zoonoses - a threat under control?", *Intervirology,* Vol. 46, no. 2, 2003, pp. 71-78.

44. Bliven K. A., Maurelli A. T., "Evolution of bacterial pathogens within the human host", *Microbiol Spectr,* Vol. 4, no. 1, 2016, pp. 10.1128/microbiolspec.VMBF-0017-2015.

45. Cagliani R., Riva S., Fumagalli M., Biasin M., Caputo S. L., Mazzotta F., Piacentini L., Pozzoli U., Bresolin N., Clerici M., et al, "A positively selected APOBEC3H haplotype is associated with natural resistance to HIV-1 infection", *Evolution,* Vol. 65, no. 11, 2011, pp. 3311-3322.

46. Cagliani R., Guerini F. R., Fumagalli M., Riva S., Agliardi C., Galimberti D., Pozzoli U., Goris A., Dubois B., Fenoglio C., et al, "A trans-specific polymorphism in ZC3HAV1 is maintained by long-standing balancing selection and may confer susceptibility to multiple sclerosis", *Mol Biol Evol,* Vol. 29, no. 6, 2012, pp. 1599-1613.

47. 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,.

48. Sironi M., Biasin M., Cagliani R., Forni D., De Luca M., Saulle I., Lo Caputo S., Mazzotta F., Macias J., Pineda J. A., et al, "A common polymorphism in TLR3 confers natural resistance to HIV-1 infection", *J Immunol,* Vol. 188, no. 2, 2012, pp. 818-823.

49. Elde N. C., Child S. J., Geballe A. P., Malik H. S., "Protein kinase R reveals an evolutionary model for defeating viral mimicry", *Nature,* Vol. 457, no. 7228, 2009, pp. 485-489.

50. Sawyer S. L., Emerman M., Malik H. S., "Discordant evolution of the adjacent antiretroviral genes TRIM22 and TRIM5 in mammals", *PLoS Pathog,* Vol. 3, no. 12, 2007, pp. E197.

51. Sawyer S. L., Wu L. I., Akey J. M., Emerman M., Malik H. S., "High-frequency persistence of an impaired allele of the retroviral defense gene TRIM5alpha in humans", *Curr Biol,* Vol. 16, no. 1, 2006, pp. 95-100.

52. Khan N., Gowthaman U., Pahari S., Agrewala J. N., "Manipulation of costimulatory molecules by intracellular pathogens: Veni, vidi, vici!!", *PLoS Pathog,* Vol. 8, no. 6, 2012, pp. E1002676.

53. Hansen T. H., Bouvier M., "MHC class I antigen presentation: Learning from viral evasion strategies", *Nat Rev Immunol,* Vol. 9, no. 7, 2009, pp. 503-513.

54. Forni D., Cagliani R., Tresoldi C., Pozzoli U., De Gioia L., Filippi G., Riva S., Menozzi G., Colleoni M., Biasin M., et al, "An evolutionary analysis of antigen processing and presentation across different timescales reveals pervasive selection", *PLoS Genet,* Vol. 10, no. 3, 2014, pp. E1004189.

55. Forni D., Cagliani R., Pozzoli U., Colleoni M., Riva S., Biasin M., Filippi G., De Gioia L., Gnudi F., Comi G. P., et al, "A 175 million year history of T cell regulatory molecules reveals widespread selection, with adaptive evolution of disease alleles", *Immunity,* Vol. 38, no. 6, 2013, pp. 1129-1141.

56. Prufer K., Racimo F., Patterson N., Jay F., Sankararaman S., Sawyer S., Heinze A., Renaud G., Sudmant P. H., de Filippo C., et al, "The complete genome sequence of a neanderthal from the altai mountains", *Nature,* Vol. 505, no. 7481, 2014, pp. 43-49.

57. Meyer M., Kircher M., Gansauge M. T., Li H., Racimo F., Mallick S., Schraiber J. G., Jay F., Prufer K., de Filippo C., et al, "A high-coverage genome sequence from an archaic denisovan individual", *Science,* Vol. 338, no. 6104, 2012, pp. 222-226.

58. Olalde I., Allentoft M. E., Sanchez-Quinto F., Santpere G., Chiang C. W., DeGiorgio M., Prado-Martinez J., Rodriguez J. A., Rasmussen S., Quilez J., et al, "Derived immune and ancestral pigmentation alleles in a 7,000-year-old mesolithic european", *Nature,* Vol. 507, no. 7491, 2014, pp. 225-228.

59. Raghavan M., Skoglund P., Graf K. E., Metspalu M., Albrechtsen A., Moltke I., Rasmussen S., Stafford T. W.,Jr, Orlando L., Metspalu E., et al, "Upper palaeolithic siberian genome reveals dual ancestry of native americans", *Nature,* Vol. 505, no. 7481, 2014, pp. 87-91.

60. Vilella A. J., Severin J., Ureta-Vidal A., Heng L., Durbin R., Birney E., "EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates", *Genome Res,* Vol. 19, no. 2, 2009, pp. 327-335.

61. Posada D., Crandall K. A., Nguyen M., Demma J. C., Viscidi R. P., "Population genetics of the porB gene of neisseria gonorrhoeae: Different dynamics in different homology groups", *Mol Biol Evol,* Vol. 17, no. 3, 2000, pp. 423-436.

62. Iannelli F., Oggioni M. R., Pozzi G., "Allelic variation in the highly polymorphic locus pspC of streptococcus pneumoniae", *Gene,* Vol. 284, no. 1-2, 2002, pp. 63-71.

63. Rogers E. A., Abdunnur S. V., McDowell J. V., Marconi R. T., "Comparative analysis of the properties and ligand binding characteristics of CspZ, a factor H binding protein, derived from borrelia burgdorferi isolates of human origin", *Infect Immun,* Vol. 77, no. 10, 2009, pp. 4396-4405.

64. Otto T. D., Rayner J. C., Bohme U., Pain A., Spottiswoode N., Sanders M., Quail M., Ollomo B., Renaud F., Thomas A. W., et al, "Genome

sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts", *Nat Commun,* Vol. 5, 2014, pp. 4754.

65. Wanaguru M., Liu W., Hahn B. H., Rayner J. C., Wright G. J., "RH5-basigin interaction plays a major role in the host tropism of plasmodium falciparum", *Proc Natl Acad Sci U S A,* Vol. 110, no. 51, 2013, pp. 20735-20740.

66. Hayton K., Gaur D., Liu A., Takahashi J., Henschen B., Singh S., Lambert L., Furuya T., Bouttenot R., Doll M., et al, "Erythrocyte binding protein PfRH5 polymorphisms determine species-specific pathways of plasmodium falciparum invasion", *Cell Host Microbe,* Vol. 4, no. 1, 2008, pp. 40-51.

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

68. Capella-Gutierrez S., Silla-Martinez J. M., Gabaldon T., "trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses", *Bioinformatics,* Vol. 25, no. 15, 2009, pp. 1972-1973.

69. Penn O., Privman E., Ashkenazy H., Landan G., Graur D., Pupko T., "GUIDANCE: A web server for assessing alignment confidence scores", *Nucleic Acids Res,* Vol. 38, no. Web Server issue, 2010, pp. W23-8.

70. Privman E., Penn O., Pupko T., "Improving the performance of positive

selection inference by filtering unreliable alignment regions", *Mol Biol Evol,* Vol. 29, no. 1, 2012, pp. 1-5.

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

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

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

74. Guindon S., Dufayard J. F., Lefort V., Anisimova M., Hordijk W., Gascuel O., "New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0", *Syst Biol,* Vol. 59, no. 3, 2010, pp. 307-321.

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

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

77. Delport W., Poon A. F., Frost S. D., Kosakovsky Pond S. L., "Datamonkey 2010: A suite of phylogenetic analysis tools for evolutionary biology", *Bioinformatics,* Vol. 26, no. 19, 2010, pp. 2455-2457.

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

79. Awadalla P., "The evolutionary genomics of pathogen recombination", *Nat Rev Genet,* Vol. 4, no. 1, 2003, pp. 50-60.

80. Wilson D. J., McVean G., "Estimating diversifying selection and functional constraint in the presence of recombination", *Genetics,* Vol. 172, no. 3, 2006, pp. 1411-1425.

81. Poon A. F., Lewis F. I., Pond S. L., Frost S. D., "An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope", *PLoS Comput Biol,* Vol. 3, no. 11, 2007, pp. E231.

82. Simonetti F. L., Teppa E., Chernomoretz A., Nielsen M., Marino Buslje C., "MISTIC: Mutual information server to infer coevolution", *Nucleic Acids Res,* Vol. 41, no. Web Server issue, 2013, pp. W8-14.

83. 1000 Genomes Project Consortium, Abecasis G. R., Auton A., Brooks L. D., DePristo M. A., Durbin R. M., Handsaker R. E., Kang H. M., Marth G. T., McVean G. A., "An integrated map of genetic variation from 1,092 human genomes", *Nature,* Vol. 491, no. 7422, 2012, pp. 56-65.

84. Prado-Martinez J., Sudmant P. H., Kidd J. M., Li H., Kelley J. L., Lorente-Galdos B., Veeramah K. R., Woerner A. E., O'Connor T. D., Santpere G., et al, "Great ape genetic diversity and population history", *Nature,* Vol. 499, no. 7459, 2013, pp. 471-475.

85. Quach H., Wilson D., Laval G., Patin E., Manry J., Guibert J., Barreiro L. B., Nerrienet E., Verschoor E., Gessain A., et al, "Different selective pressures shape the evolution of toll-like receptors in human and african great ape populations", *Hum Mol Genet,* Vol. 22, no. 23, 2013, pp. 4829-4840.

86. 1000 Genomes Project Consortium, Durbin R. M., Abecasis G. R., Altshuler D. L., Auton A., Brooks L. D., Durbin R. M., Gibbs R. A., Hurles M. E., McVean G. A., "A map of human genome variation from population-scale sequencing", *Nature,* Vol. 467, no. 7319, 2010, pp. 1061-1073.

87. Cereda M., Sironi M., Cavalleri M., Pozzoli U., "GeCo++: A C++ library for genomic features computation and annotation in the presence of variants", *Bioinformatics,* Vol. 27, no. 9, 2011, pp. 1313-1315.

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

89. 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.

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

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

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

93. Barreiro L. B., Ben-Ali M., Quach H., Laval G., Patin E., Pickrell J. K., Bouchier C., Tichit M., Neyrolles O., Gicquel B., et al, "Evolutionary dynamics of human toll-like receptors and their different contributions to host defense", *PLoS Genet,* Vol. 5, no. 7, 2009, pp. E1000562.

94. Guindon S., Delsuc F., Dufayard J. F., Gascuel O., "Estimating maximum likelihood phylogenies with PhyML", *Methods Mol Biol,* Vol. 537, 2009, pp. 113-137.

95. Karolchik D., Hinrichs A. S., Furey T. S., Roskin K. M., Sugnet C. W., Haussler D., Kent W. J., "The UCSC table browser data retrieval tool", *Nucleic Acids Res,* Vol. 32, no. Database issue, 2004, pp. D493-6.

96. Kuhn R. M., Haussler D., Kent W. J., "The UCSC genome browser and associated tools", *Brief Bioinform,* Vol. 14, no. 2, 2013, pp. 144-161.

97. Eswar N., Webb B., Marti-Renom M. A., Madhusudhan M. S., Eramian D., Shen M. Y., Pieper U., Sali A., "Comparative protein structure modeling using modeller", *Curr Protoc Bioinformatics,* Vol. Chapter 5, 2006, pp. Unit 5.6.

98. Zhang Y., "I-TASSER server for protein 3D structure prediction", *BMC Bioinformatics,* Vol. 9, 2008, pp. 40-2105-9-40.

99. Roy A., Kucukural A., Zhang Y., "I-TASSER: A unified platform for automated protein structure and function prediction", *Nat Protoc,* Vol. 5, no. 4, 2010, pp. 725-738.

100. Willard L., Ranjan A., Zhang H., Monzavi H., Boyko R. F., Sykes B. D., Wishart D. S., "VADAR: A web server for quantitative evaluation of protein structure quality", *Nucleic Acids Res,* Vol. 31, no. 13, 2003, pp. 3316-3319.

101. McGuffin L. J., Bryson K., Jones D. T., "The PSIPRED protein structure prediction server", *Bioinformatics,* Vol. 16, no. 4, 2000, pp. 404-405.

102. Tina K. G., Bhadra R., Srinivasan N., "PIC: Protein interactions calculator", *Nucleic Acids Res,* Vol. 35, no. Web Server issue, 2007, pp. W473-6.

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

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

105. Miyazawa M., Lopalco L., Mazzotta F., Lo Caputo S., Veas F., Clerici M., ESN Study Group, "The 'immunologic advantage' of HIV-exposed seronegative individuals", *Aids,* Vol. 23, no. 2, 2009, pp. 161-175.

106. 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, "PLINK: A tool set for whole-genome association and population-based linkage analyses", *Am J Hum Genet,* Vol. 81, no. 3, 2007, pp. 559-575.

107. 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.

108. Axelsson E., Ratnakumar A., Arendt M. L., Maqbool K., Webster M. T., Perloski M., Liberg O., Arnemo J. M., Hedhammar A., Lindblad-Toh K., "The genomic signature of dog domestication reveals adaptation to a starch-rich diet", *Nature,* Vol. 495, no. 7441, 2013, pp. 360-364.

109. Freedman A. H., Gronau I., Schweizer R. M., Ortega-Del Vecchyo D., Han E., Silva P. M., Galaverni M., Fan Z., Marx P., Lorente-Galdos B., et al, "Genome sequencing highlights the dynamic early history of dogs", *PLoS Genet,* Vol. 10, no. 1, 2014, pp. E1004016.

110. Caviedes-Vidal E., McWhorter T. J., Lavin S. R., Chediack J. G., Tracy C. R., Karasov W. H., "The digestive adaptation of flying vertebrates: High intestinal paracellular absorption compensates for smaller guts", *Proc Natl Acad Sci U S A,* Vol. 104, no. 48, 2007, pp. 19132-19137.

111. Laden G., Wrangham R., "The rise of the hominids as an adaptive shift in fallback foods: Plant underground storage organs (USOs) and australopith origins", *J Hum Evol,* Vol. 49, no. 4, 2005, pp. 482-498.

112. Humphrey L. T., De Groote I., Morales J., Barton N., Collcutt S., Bronk Ramsey C., Bouzouggar A., "Earliest evidence for caries and exploitation of starchy plant foods in pleistocene hunter-gatherers from morocco", *Proc Natl Acad Sci U S A,* Vol. 111, no. 3, 2014, pp. 954-959.

113. Choi Y. K., Johlin F. C.,Jr, Summers R. W., Jackson M., Rao S. S., "Fructose intolerance: An under-recognized problem", *Am J Gastroenterol,* Vol. 98, no. 6, 2003, pp. 1348-1353.

114. Skoog S. M., Bharucha A. E., "Dietary fructose and gastrointestinal symptoms: A review", *Am J Gastroenterol,* Vol. 99, no. 10, 2004, pp. 2046-2050.

115. Werling D., Jann O. C., Offord V., Glass E. J., Coffey T. J., "Variation matters: TLR structure and species-specific pathogen recognition", *Trends Immunol,* Vol. 30, no. 3, 2009, pp. 124-130.

116. Varki A., "A chimpanzee genome project is a biomedical imperative", *Genome Res,* Vol. 10, no. 8, 2000, pp. 1065-1070.

117. Willer D. O., Ambagala A. P., Pilon R., Chan J. K., Fournier J., Brooks J., Sandstrom P., Macdonald K. S., "Experimental infection of cynomolgus macaques (macaca fascicularis) with human varicella-zoster virus", *J Virol,* Vol. 86, no. 7, 2012, pp. 3626-3634.

118. Jones K. E., Patel N. G., Levy M. A., Storeygard A., Balk D., Gittleman J. L., Daszak P., "Global trends in emerging infectious diseases", *Nature,* Vol. 451, no. 7181, 2008, pp. 990-993.

119. Yi G., Brendel V. P., Shu C., Li P., Palanathan S., Cheng Kao C., "Single nucleotide polymorphisms of human STING can affect innate immune response to cyclic dinucleotides", *PLoS One,* Vol. 8, no. 10, 2013, pp. E77846.

120. Gao P., Zillinger T., Wang W., Ascano M., Dai P., Hartmann G., Tuschl T., Deng L., Barchet W., Patel D. J., "Binding-pocket and lid-region substitutions render human STING sensitive to the species-specific drug DMXAA", *Cell Rep,* Vol. 8, no. 6, 2014, pp. 1668-1676.

121. Gao P., Ascano M., Zillinger T., Wang W., Dai P., Serganov A. A., Gaffney B. L., Shuman S., Jones R. A., Deng L., et al, "Structure-function analysis of STING activation by c[G(2',5')pA(3',5')p] and targeting by antiviral DMXAA", *Cell,* Vol. 154, no. 4, 2013, pp. 748-762.

122. Ablasser A., Goldeck M., Cavlar T., Deimling T., Witte G., Rohl I., Hopfner K. P., Ludwig J., Hornung V., "cGAS produces a 2'-5'-linked cyclic dinucleotide second messenger that activates STING", *Nature,* Vol. 498, no. 7454, 2013, pp. 380-384.

123. Diner E. J., Burdette D. L., Wilson S. C., Monroe K. M., Kellenberger C. A., Hyodo M., Hayakawa Y., Hammond M. C., Vance R. E., "The innate immune DNA sensor cGAS produces a noncanonical cyclic dinucleotide that activates human STING", *Cell Rep,* Vol. 3, no. 5, 2013, pp. 1355-1361.

124. Gao P., Ascano M., Wu Y., Barchet W., Gaffney B. L., Zillinger T., Serganov A. A., Liu Y., Jones R. A., Hartmann G., et al, "Cyclic [G(2',5')pA(3',5')p] is the metazoan second messenger produced by DNA-activated cyclic GMP-AMP synthase", *Cell,* Vol. 153, no. 5, 2013, pp. 1094-1107.

125. Zhang G., Cowled C., Shi Z., Huang Z., Bishop-Lilly K. A., Fang X., Wynne J. W., Xiong Z., Baker M. L., Zhao W., et al, "Comparative analysis of bat genomes provides insight into the evolution of flight and immunity", *Science,* Vol. 339, no. 6118, 2013, pp. 456-460.

126. Jin W., Wu D. D., Zhang X., Irwin D. M., Zhang Y. P., "Positive selection on the gene RNASEL: Correlation between patterns of evolution and function", *Mol Biol Evol,* Vol. 29, no. 10, 2012, pp. 3161-3168.

127. Han Y., Donovan J., Rath S., Whitney G., Chitrakar A., Korennykh A., "Structure of human RNase L reveals the basis for regulated RNA decay in the IFN response", *Science,* Vol. 343, no. 6176, 2014, pp. 1244-1248.

128. Ferguson W., Dvora S., Fikes R. W., Stone A. C., Boissinot S., "Long-term balancing selection at the antiviral gene OAS1 in central african chimpanzees", *Mol Biol Evol,* Vol. 29, no. 4, 2012, pp. 1093-1103.

129. Dubensky T. W.,Jr, Kanne D. B., Leong M. L., "Rationale, progress and development of vaccines utilizing STING-activating cyclic dinucleotide adjuvants", *Ther Adv Vaccines,* Vol. 1, no. 4, 2013, pp. 131-143.

130. Nishikura K., "Functions and regulation of RNA editing by ADAR deaminases", *Annu Rev Biochem,* Vol. 79, 2010, pp. 321-349.

131. Slotkin W., Nishikura K., "Adenosine-to-inosine RNA editing and human disease", *Genome Med,* Vol. 5, no. 11, 2013, pp. 105.

132. Savva Y. A., Rieder L. E., Reenan R. A., "The ADAR protein family", *Genome Biol,* Vol. 13, no. 12, 2012, pp. 252.

133. Ramaswami G., Li J. B., "RADAR: A rigorously annotated database of A-to-I RNA editing", *Nucleic Acids Res,* Vol. 42, no. Database issue, 2014, pp. D109-13.

134. Tomaselli S., Locatelli F., Gallo A., "The RNA editing enzymes ADARs: Mechanism of action and human disease", *Cell Tissue Res,* Vol. 356, no. 3, 2014, pp. 527-532.

135. Chen J. Y., Peng Z., Zhang R., Yang X. Z., Tan B. C., Fang H., Liu C. J., Shi M., Ye Z. Q., Zhang Y. E., et al, "RNA editome in rhesus macaque shaped by purifying selection", *PLoS Genet,* Vol. 10, no. 4, 2014, pp. E1004274.

136. Bahn J. H., Lee J. H., Li G., Greer C., Peng G., Xiao X., "Accurate identification of A-to-I RNA editing in human by transcriptome sequencing", *Genome Res,* Vol. 22, no. 1, 2012, pp. 142-150.

137. Doria M., Neri F., Gallo A., Farace M. G., Michienzi A., "Editing of HIV-1 RNA by the double-stranded RNA deaminase ADAR1 stimulates viral infection", *Nucleic Acids Res,* Vol. 37, no. 17, 2009, pp. 5848-5858.

138. Doria M., Tomaselli S., Neri F., Ciafre S. A., Farace M. G., Michienzi A., Gallo A., "ADAR2 editing enzyme is a novel human immunodeficiency virus-1 proviral factor", *J Gen Virol,* Vol. 92, no. Pt 5, 2011, pp. 1228-1232.

139. Phuphuakrat A., Kraiwong R., Boonarkart C., Lauhakirti D., Lee T. H., Auewarakul P., "Double-stranded RNA adenosine deaminases enhance expression of human immunodeficiency virus type 1 proteins", *J Virol,* Vol. 82, no. 21, 2008, pp. 10864-10872.

140. Biswas N., Wang T., Ding M., Tumne A., Chen Y., Wang Q., Gupta P., "ADAR1 is a novel multi targeted anti-HIV-1 cellular protein", *Virology,* Vol. 422, no. 2, 2012, pp. 265-277.

141. Kantaputra P. N., Chinadet W., Ohazama A., Kono M., "Dyschromatosis symmetrica hereditaria with long hair on the forearms, hypo/hyperpigmented hair, and dental anomalies: Report of a novel ADAR1 mutation", *Am J Med Genet A,* Vol. 158A, no. 9, 2012, pp. 2258-2265.

142. Sharma R., Wang Y., Zhou P., Steinman R. A., Wang Q., "An essential role of RNA editing enzyme ADAR1 in mouse skin", *J Dermatol Sci,* Vol. 64, no. 1, 2011, pp. 70-72.

143. Paz-Yaacov N., Levanon E. Y., Nevo E., Kinar Y., Harmelin A., Jacob-Hirsch J., Amariglio N., Eisenberg E., Rechavi G., "Adenosine-to-inosine RNA editing shapes transcriptome diversity in primates", *Proc Natl Acad Sci U S A,* Vol. 107, no. 27, 2010, pp. 12174-12179.

144. Li J. B., Church G. M., "Deciphering the functions and regulation of brain-enriched A-to-I RNA editing", *Nat Neurosci,* Vol. 16, no. 11, 2013, pp. 1518-1522.

145. Liu Z., Pan Q., Ding S., Qian J., Xu F., Zhou J., Cen S., Guo F., Liang C., "The interferon-inducible MxB protein inhibits HIV-1 infection", *Cell Host Microbe,* 2013,.

146. Goujon C., Moncorge O., Bauby H., Doyle T., Ward C. C., Schaller T., Hue S., Barclay W. S., Schulz R., Malim M. H., "Human MX2 is an interferon-induced post-entry inhibitor of HIV-1 infection", *Nature,* 2013,.

147. Kane M., Yadav S. S., Bitzegeio J., Kutluay S. B., Zang T., Wilson S. J., Schoggins J. W., Rice C. M., Yamashita M., Hatziioannou T., et al, "MX2 is an interferon-induced inhibitor of HIV-1 infection", *Nature,* 2013,.

148. Mitchell P. S., Patzina C., Emerman M., Haller O., Malik H. S., Kochs G., "Evolution-guided identification of antiviral specificity determinants in the broadly acting interferon-induced innate immunity factor MxA", *Cell Host Microbe,* Vol. 12, no. 4, 2012, pp. 598-604.

149. Sasaki K., Tungtrakoolsub P., Morozumi T., Uenishi H., Kawahara M., Watanabe T., "A single nucleotide polymorphism of porcine MX2 gene provides antiviral activity against vesicular stomatitis virus", *Immunogenetics,* Vol. 66, no. 1, 2014, pp. 25-32.

150. 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.

151. Barrett J. H., Iles M. M., Harland M., Taylor J. C., Aitken J. F., Andresen P. A., Akslen L. A., Armstrong B. K., Avril M. F., Azizi E., et al,

"Genome-wide association study identifies three new melanoma susceptibility loci", *Nat Genet,* Vol. 43, no. 11, 2011, pp. 1108-1113.

152. Lambris J. D., Ricklin D., Geisbrecht B. V., "Complement evasion by human pathogens", *Nat Rev Microbiol,* Vol. 6, no. 2, 2008, pp. 132-142.

153. Pangburn M. K., Ferreira V. P., Cortes C., "Discrimination between host and pathogens by the complement system", *Vaccine,* Vol. 26 Suppl 8, 2008, pp. I15-21.

154. Serruto D., Rappuoli R., Scarselli M., Gros P., van Strijp J. A., "Molecular mechanisms of complement evasion: Learning from staphylococci and meningococci", *Nat Rev Microbiol,* Vol. 8, no. 6, 2010, pp. 393-399.

155. Sawyer S. L., Elde N. C., "A cross-species view on viruses", *Curr Opin Virol,* Vol. 2, no. 5, 2012, pp. 561-568.

156. Daugherty M. D., Malik H. S., "Rules of engagement: Molecular insights from host-virus arms races", *Annu Rev Genet,* Vol. 46, 2012, pp. 677-700.

157. Meri T., Amdahl H., Lehtinen M. J., Hyvarinen S., McDowell J. V., Bhattacharjee A., Meri S., Marconi R., Goldman A., Jokiranta T. S., "Microbes bind complement inhibitor factor H via a common site", *PLoS Pathog,* Vol. 9, no. 4, 2013, pp. e1003308.

158. Fortin A., Stevenson M. M., Gros P., "Susceptibility to malaria as a complex trait: Big pressure from a tiny creature", *Hum Mol Genet,* Vol. 11, no. 20, 2002, pp. 2469-2478.

159. Ngampasutadol J., Ram S., Blom A. M., Jarva H., Jerse A. E., Lien E., Goguen J., Gulati S., Rice P. A., "Human C4b-binding protein selectively interacts with neisseria gonorrhoeae and results in species-specific infection", *Proc Natl Acad Sci U S A,* Vol. 102, no. 47, 2005, pp. 17142-17147.

160. Barber M. F., Elde N. C., "Nutritional immunity. escape from bacterial iron piracy through rapid evolution of transferrin", *Science,* Vol. 346, no. 6215, 2014, pp. 1362-1366.

161. Meri S., Jordens M., Jarva H., "Microbial complement inhibitors as vaccines", *Vaccine,* Vol. 26 Suppl 8, 2008, pp. I113-7.

162. Jefferies J. M., Clarke S. C., Webb J. S., Kraaijeveld A. R., "Risk of red queen dynamics in pneumococcal vaccine strategy", *Trends Microbiol,* Vol. 19, no. 8, 2011, pp. 377-381.

163. Little T. J., Allen J. E., Babayan S. A., Matthews K. R., Colegrave N., "Harnessing evolutionary biology to combat infectious disease", *Nat Med,* Vol. 18, no. 2, 2012, pp. 217-220.

164. Al-Daghri N. M., Cagliani R., Forni D., Alokail M. S., Pozzoli U., Alkharfy K. M., Sabico S., Clerici M., Sironi M., "Mammalian NPC1 genes may undergo positive selection and human polymorphisms

associate with type 2 diabetes", *BMC Med,* Vol. 10, 2012, pp. 140-7015-10-140.

165. Wang H., Shi Y., Song J., Qi J., Lu G., Yan J., Gao G. F., "Ebola viral glycoprotein bound to its endosomal receptor niemann-pick C1", *Cell,* Vol. 164, no. 1-2, 2016, pp. 258-268.

166. Ng M., Ndungo E., Kaczmarek M. E., Herbert A. S., Binger T., Kuehne A. I., Jangra R. K., Hawkins J. A., Gifford R. J., Biswas R., et al, "Filovirus receptor NPC1 contributes to species-specific patterns of ebolavirus susceptibility in bats", *Elife,* Vol. 4, 2015, pp. 10.7554/eLife.11785.

167. Taylor D. J., Ballinger M. J., Zhan J. J., Hanzly L. E., Bruenn J. A., "Evidence that ebolaviruses and cuevaviruses have been diverging from marburgviruses since the miocene", *Peerj,* Vol. 2, 2014, pp. E556.

168. Taylor D. J., Dittmar K., Ballinger M. J., Bruenn J. A., "Evolutionary maintenance of filovirus-like genes in bat genomes", *BMC Evol Biol,* Vol. 11, 2011, pp. 336-2148-11-336.

169. Taylor D. J., Leach R. W., Bruenn J., "Filoviruses are ancient and integrated into mammalian genomes", *BMC Evol Biol,* Vol. 10, 2010, pp. 193-2148-10-193.

170. Wilson J. A., Hevey M., Bakken R., Guest S., Bray M., Schmaljohn A. L., Hart M. K., "Epitopes involved in antibody-mediated protection

from ebola virus", *Science,* Vol. 287, no. 5458, 2000, pp. 1664-1666.

171. Azarian T., Lo Presti A., Giovanetti M., Cella E., Rife B., Lai A., Zehender G., Ciccozzi M., Salemi M., "Impact of spatial dispersion, evolution, and selection on ebola zaire virus epidemic waves", *Sci Rep,* Vol. 5, 2015, pp. 10170.

172. Ladner J. T., Wiley M. R., Mate S., Dudas G., Prieto K., Lovett S., Nagle E. R., Beitzel B., Gilbert M. L., Fakoli L., et al, "Evolution and spread of ebola virus in liberia, 2014-2015", *Cell Host Microbe,* Vol. 18, no. 6, 2015, pp. 659-669.

173. Park D. J., Dudas G., Wohl S., Goba A., Whitmer S. L., Andersen K. G., Sealfon R. S., Ladner J. T., Kugelman J. R., Matranga C. B., et al, "Ebola virus epidemiology, transmission, and evolution during seven months in sierra leone", *Cell,* Vol. 161, no. 7, 2015, pp. 1516-1526.

174. Li Y. H., Chen S. P., "Evolutionary history of ebola virus", *Epidemiol Infect,* Vol. 142, no. 6, 2014, pp. 1138-1145.

175. Casadevall A., Pirofski L. A., "The ebola epidemic crystallizes the potential of passive antibody therapy for infectious diseases", *PLoS Pathog,* Vol. 11, no. 4, 2015, pp. E1004717.

176. Liu W., Li Y., Learn G. H., Rudicell R. S., Robertson J. D., Keele B. F., Ndjango J. B., Sanz C. M., Morgan D. B., Locatelli S., et al, "Origin of the human malaria parasite plasmodium falciparum in gorillas", *Nature,* Vol. 467, no. 7314, 2010, pp. 420-425.

6. SCIENTIFIC PRODUCTS

1. Biasin M., Sironi M., Saulle I., Pontremoli C., Garziano M., Cagliani R., Trabattoni D., Lo Caputo S., Vichi F., Mazzotta F., et al, "A 6-amino acid insertion/deletion polymorphism in the mucin domain of TIM-1 confers protections against HIV-1 infection", *Microbes Infect,* 2016,.

2. Cagliani R., Forni D., Filippi G., Mozzi A., De Gioia L., Pontremoli C., Pozzoli U., Bresolin N., Clerici M., Sironi M., "The mammalian complement system as an epitome of host-pathogen genetic conflicts", *Mol Ecol,* Vol. 25, no. 6, 2016, pp. 1324-1339.

3. Pontremoli C., Forni D., Cagliani R., Filippi G., De Gioia L., Pozzoli U., Clerici M., Sironi M., "Positive selection drives evolution at the host-filovirus interaction surface", *Mol Biol Evol,* Vol. 33, no. 11, 2016, pp. 2836-2847.

4. Forni D., Mozzi A., Pontremoli C., Vertemara J., Pozzoli U., Biasin M., Bresolin N., Clerici M., Cagliani R., Sironi M., "Diverse selective regimes shape genetic diversity at ADAR genes and at their coding targets", *RNA Biol,* Vol. 12, no. 2, 2015, pp. 149-161.

5. Forni D., Pontremoli C., Cagliani R., Pozzoli U., Clerici M., Sironi M., "Positive selection underlies the species-specific binding of plasmodium falciparum RH5 to human basigin", *Mol Ecol,* Vol. 24, no. 18, 2015, pp. 4711-4722.

6. Mozzi A., Pontremoli C., Forni D., Clerici M., Pozzoli U., Bresolin N.,

Cagliani R., Sironi M., "OASes and STING: Adaptive evolution in concert", *Genome Biol Evol,* Vol. 7, no. 4, 2015, pp. 1016-1032.

7. Pontremoli C., Mozzi A., Forni D., Cagliani R., Pozzoli U., Menozzi G., Vertemara J., Bresolin N., Clerici M., Sironi M., "Natural selection at the brush-border: Adaptations to carbohydrate diets in humans and other mammals", *Genome Biol Evol,* Vol. 7, no. 9, 2015, pp. 2569-2584.

8. Sironi M., Biasin M., Pontremoli C., Cagliani R., Saulle I., Trabattoni D., Vichi F., Lo Caputo S., Mazzotta F., Aguilar-Jimenez W., et al, "Variants in the CYP7B1 gene region do not affect natural resistance to HIV-1 infection", *Retrovirology,* Vol. 12, 2015, pp. 80-015-0206-0.

9. Sironi M., Biasin M., Cagliani R., Gnudi F., Saulle I., Ibba S., Filippi G., Yahyaei S., Tresoldi C., Riva S., Pontremoli C., et al, "Evolutionary analysis identifies an MX2 haplotype associated with natural resistance to HIV-1 infection", *Mol Biol Evol,* 2014,.

10. Sironi M., Cagliani R., Pontremoli C., Rossi M., Migliorino G., Clerici M., Gori A., "The CCR5Delta32 allele is not a major predisposing factor for severe H1N1pdm09 infection", *BMC Res Notes,* Vol. 7, 2014, pp. 504-0500-7-504.