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

Rachele Cagliani1, Stefania Riva1, Mara Biasin2, Matteo Fumagalli1,4, Uberto Pozzoli1, Sergio Lo Caputo5, Francesco Mazzotta5, Luca Piacentini2, Nereo Bresolin1,6, Mario Clerici3,7,† and Manuela Sironi1,*†

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

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

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

INTRODUCTION

Cells expressing foreign proteins in association with a major histocompatibility complex (MHC) class I molecule are recognized and eliminated by CD8+ T cells; this pathway is specialized in the recognition of intracellular pathogens including viruses. The composition of the peptide repertoire displayed by the MHC I largely depends on the
specificity of the MHC peptide-binding groove, but also on the availability of suitable peptides generated by the antigen-processing pathway. Therefore, antigen processing is essential for assuring immune surveillance and for establishing immunodominance (i.e. the phenomenon whereby only a minority of epitopes are functionally immunogenic).

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

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

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

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

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

RESULTS

Nucleotide diversity and neutrality tests

ERAP1 and ERAP2 are located in a head-to-head orientation on the long arm of chromosome 5 (Fig. 1). Analysis of linkage disequilibrium (LD) in three HapMap populations, namely Yoruba (YRI), Europeans (CEU) and East Asians (EAS), indicated that the two genes lie in distinct LD blocks (Fig. 1 and Supplementary Material, Fig. S1). Throughout the manuscript, single-nucleotide polymorphisms (SNPs) are indicated using the NCBI notation for reference SNP cluster IDs (rs#) (http://www.ncbi.nlm.nih.gov/projects/SNP/), and derived alleles are defined through comparison with the chimpanzee reference sequence. In order to

Figure 1. Schematic diagram of the genomic region encompassing ERAP1 and ERAP2. Green lines represent the ERAP1 and ERAP2 gene regions. The direction of transcription is marked by the arrow. The position of SNPs mentioned in the text is shown. Tag SNPs are shown in red. The two regions we resequenced are denoted by the hatched boxes. LD (D’/Lod) refers to CEU, and data were derived from HapMap.
test the hypothesis whereby balancing selection has maintained nucleotide variability at ERAP1 and ERAP2, we resequenced two genomic regions internal to these genes in the same three populations. Specifically, an ∼3.9 kb region encompassing the Lys528Arg variant (rs30187) was analysed for ERAP1 (Fig. 1). This choice was motivated by the fact that, although multiple variants in the gene have been associated with ankylosing spondylitis (12), the Lys528Arg allele was also shown to decrease enzymatic activity (13), suggesting that rs30187 may represent a functional variant. The derived 528Arg allele has also been associated with essential hypertension (14) and haemolytic uraemic syndrome (15), and a previous analysis (16) indicated that rs30187 defines an expression QTL (eQTL) for ERAP1. This same observation applies to rs2247650, an SNP located within ERAP2 (16). This SNP lies relatively close to rs2549782 (Fig. 1), which determines the substitution of the highly conserved asparagine residue at codon 392 with a lysine; the polymorphism has been associated with both ankylosing spondylitis and pre-eclampsia (17,18). Thus, an ∼6.6 kb region within ERAP2 and covering both variants was resequenced (Fig. 1).

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

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

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

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

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

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

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

\( F_{ST} \) (25) measures variations in allele frequencies among populations (genetic differentiation) and largely depends on demographic history (which affects all loci equally). Yet, natural selection may drive allele frequencies to differ more or less than expected on the basis of demography alone. In particular, balancing selection may lead to a decrease in population differentiation compared with neutrally evolving loci (26). \( F_{ST} \) among YRI, CEU and EAS calculated for the
ERAP1 and ERAP2 regions amounted to 0.022 and 0.0088, respectively. Both values are lower than the genome average of 0.123 (27), and their percentile rank in the distribution of $F_{ST}$ calculated for 5 kb reference windows were 0.05 and 0.022, respectively, indicating that both regions display unusually low population genetic differentiation.

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

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

**Haplotype analysis and time to the most recent common ancestor estimation**

Another feature of balancing selection is the maintenance of two or more highly divergent haplotype clades; in cases of long-standing balancing selection, the coalescence time of
these genealogies is deeper than expected under neutrality (28). In order to analyse the haplotype structure of the two regions in \textit{ERAP1} and \textit{ERAP2}, we constructed median-joining networks (Fig. 2). Both haplotype networks display two major clades separated by deep branches. In line with the extended LD pattern (Fig. 1 and Supplementary Material, Fig. S1), the haplotype network for the \textit{ERAP2} gene region presented no reticulations and few recurrent mutations. Both the Asn392Lys variant (rs2549782, variant 12 in the network) separates the two major clades. In order to estimate the time to the most recent common ancestor (TMRCA) of the \textit{ERAP1} and \textit{ERAP2} haplotype genealogies, we applied a phylogeny-based method (29) based on the average pairwise difference between the haplotype clusters. Using mutation rates based on the number of fixed differences with chimpanzee and a separation time of 6 million years (MY) (30), we estimated TMRCAs of 4.12 MY (SD: 0.860 MY) and 5.08 MY (SD: 0.918 MY) for \textit{ERAP1} and \textit{ERAP2}, respectively.

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

### Table 2. MLHKA test for the two regions in \textit{ERAP1} and \textit{ERAP2}

<table>
<thead>
<tr>
<th>Gene</th>
<th>Fixed substitutions</th>
<th>MLHKA YRI ( k^a )</th>
<th>( P )-value</th>
<th>CEU ( k^b )</th>
<th>( P )-value</th>
<th>EAS ( k^c )</th>
<th>( P )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{ERAP1}</td>
<td>30</td>
<td>4.35</td>
<td>( 2.42 \times 10^{-4} )</td>
<td>6.46</td>
<td>( 5.19 \times 10^{-6} )</td>
<td>4.65</td>
<td>( 4.34 \times 10^{-4} )</td>
</tr>
<tr>
<td>\textit{ERAP2}</td>
<td>34</td>
<td>4.4</td>
<td>( 6.58 \times 10^{-4} )</td>
<td>6.09</td>
<td>( 2.86 \times 10^{-6} )</td>
<td>5.32</td>
<td>( 3.6 \times 10^{-5} )</td>
</tr>
</tbody>
</table>

*Selection parameter \((k > 1)\) indicates an excess of polymorphism compared with divergence; \((k < 1)\) indicates the opposite situation.

### \textit{ERAP1}/\textit{ERAP2} genotypes and resistance to HIV-1 infection

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

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

---

\[ e \]

\[ k \]

\[ a \]

\[ b \]

\[ c \]
the \( ERAP2 \) variant (rs2549782) was significantly different in the two cohorts (Bonferroni-corrected \( P = 0.004 \), Table 3), with the difference being mainly accounted for by an over-representation of GG (Lys/Lys) genotypes in the ESN sample (Bonferroni-corrected \( P \)-value for a recessive model = 0.00097, Table 3).

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

**DISCUSSION**

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

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

Comparison of genotype frequencies in the ESN and control samples indicated no significant difference for rs30187 in \( ERAP1 \) (Table 3). Conversely, the genotype distribution of

**Figure 3.** Estimated haplotype tree for the \( ERAP1 \) and \( ERAP2 \) gene regions we resequenced. The \( ERAP1 \) and \( ERAP2 \) trees are shown in (A) and (B), respectively. Mutations are represented as black dots and named for their physical position along the regions. The absolute frequency of each haplotype is also reported. Note that mutation numbering does not correspond to that reported in Figure 2.
and Asn392Lys (in ERAP2) may be regarded as possible genetic modifiers of viral infection susceptibility and we tested this hypothesis by analysing the genotype distribution of these variants in a cohort of Italian subjects who remained seronegative despite multiple exposures to HIV-1.

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

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

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

MATERIALS AND METHODS

HapMap samples and sequencing

Human genomic DNA from HapMap subjects (20 individuals for each population) was obtained from the Coriell Institute for Medical Research. All analysed regions were PCR-amplified and directly sequenced; primer sequences are available upon request. PCR products were treated with ExoSAP-IT (USB Corporation, Cleveland, OH, USA), directly sequenced on both strands with a Big Dye Terminator Sequencing Kit (v3.1 Applied Biosystems) and run on an Applied Biosystems ABI 3130 XL Genetic Analyzer (Applied Biosystems). Sequences were assembled using AutoAssembler version 1.4.0 (Applied Biosystems) and inspected manually by two distinct operators.

Human subjects and genotyping

Blood samples were collected from 69 Italian ESN subjects. Inclusion criteria were a history of multiple unprotected sexual episodes for more than 4 years at the time of the enrolment, with at least three episodes of at-risk intercourse within 4 months prior to study entry and an average of 30 (range 18 to >100) unprotected sexual contacts per year. These ESN subjects are part of a well-characterized cohort of serodiscordant heterosexual couples that has been followed since 1997 (reviewed in 50). As for controls, 218 Italian donors were also included in the study, irrespective of their HIV infection status. The study was reviewed and approved by the institutional review board of the S.M. Annunziata Hospital, Florence, Italy. Written informed consent was obtained from all subjects.

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

Data retrieval and haplotype construction

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

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

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

Statistical analysis

Tajima’s D (19), Fu and Li’s D* and F* (20) statistics, as well as diversity parameters θw (21) and π (22) were calculated using libsequence (54), a C++ class library providing an object-oriented framework for the analysis of molecular population genetic data. Calibrated coalescent simulations were performed using the cosi package (55) and its best-fit parameters for YRI, EU and EAS populations with 10 000 iterations. Coalescent simulations were conditioned on mutation rate and recombination rate. Estimates of the population recombination rate parameter θ were obtained from diploid data by a composite likelihood method (56), with the use of the Web application MAXDIP (http://genapps.uchicago.edu/maxdip/). The maximum-likelihood-ratio HKA test was performed using the MLHKA software (24), as proposed previously (57). Briefly, 16 reference loci were randomly selected among NIEHS loci shorter than 20 kb that have been resequenced in the three populations; the only criterion was that Tajima’s D did not suggest the action of natural selection (i.e. Tajima’s D is higher than the 5th and lower than the 95th percentiles in the distribution of NIEHS genes). The reference set was accounted for by the following genes: VN3, PLA2G2D, MB, MADC2L2, HRAS, CYP17A1, ATOX1, BNI3, CDC20, NGB, TUBA1, MT3, NUDT1, PRDX5, RETN and JUND.

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

Wittke-Thompson et al. (40) derived genotype frequencies for biallelic loci in cases and controls, assuming HWE in the general population. The equations are parametrized in q (susceptibility allele frequency), α (risk in non-susceptible homozygotes), β (homozygote relative risk), γ (homozygote relative risk) and Kp (trait prevalence in the general population). We obtained ML estimates for these parameters, minimizing the goodness-of-fit test statistic (as reported in 40) using the BFGS method.

Using an estimate of Kp, the procedure was repeated with a general model estimating q, β and γ, and for constrained specific models, estimating q and γ [dominant: β = γ; recessive: β = 1, γ > 1; additive: β = (γ + 1)/2, γ > 1 and multiplicative: β = sqrt(γ), γ > 1]. Given the different number of parameters in the general model, the Akaike Information Criterion was used for the best-fit model selection. A P-value was then calculated for the minimal value of the test statistic using a χ² distribution with 1 or 2 df for the general and constrained models, respectively.

All calculations were carried out in the R environment (58). Tag SNPs along ERAP2 were selected with Tagger (59) (http://www.broadinstitute.org/mpg/tagger/) using multi-marker predictors to capture alleles with minor allele frequency >0.20 and an r² threshold of 0.8. As allowed by the software, we specified to include rs2549782 among tag
SNPs. The input region cover the whole gene 5 kb upstream the transcription start site. Association analyses were performed using PLINK (60).

Haplotype analysis and TM RCA calculation

Median-joining networks to infer haplotype genealogy were constructed using NETWORK 4.5 (29). The estimate of the TM RCA was obtained using a phylogeny-based approach implemented in NETWORK 4.5 using a mutation rate based on the number of fixed differences between chimpanzee and humans. Additional TM RCA estimates derived from the application of a maximum-likelihood coalescent method implemented in GENETREE (31,32). Again, the mutation rate $\mu$ was obtained on the basis of the divergence between human and chimpanzee and under the assumption that both the species separation occurred 6 MY ago (30) and with a generation time of 25 years. The migration matrix was derived from previous estimated migration rates (55). Using this $\mu$ and $\theta$ maximum-likelihood ($\theta_{ML}$), we estimated the effective population size parameter ($N_e$). With these assumptions, the coalescence time, scaled in $2N_e$ units, was converted into years. For the coalescence process, $10^8$ simulations were performed.

In order to obtain TM RCA estimates that take recombination into account, we used the InterRho program (35,36), kindly provided by Ying Wang. Genealogies are related considering $N_e$ population size parameter ($\theta_{ML}$). The estimate of the TM RCA was obtained on the basis of the divergence between chimpanzee and humans. Additional TM RCA estimates derived from the application of a maximum-likelihood coalescent method implemented in GENETREE (31,32). Again, the mutation rate $\mu$ was obtained on the basis of the divergence between human and chimpanzee and under the assumption that both the species separation occurred 6 MY ago (30) and with a generation time of 25 years. The migration matrix was derived from previous estimated migration rates (55). Using this $\mu$ and $\theta$ maximum-likelihood ($\theta_{ML}$), we estimated the effective population size parameter ($N_e$). With these assumptions, the coalescence time, scaled in $2N_e$ units, was converted into years. For the coalescence process, $10^8$ simulations were performed.

In order to obtain TM RCA estimates that take recombination into account, we used the InterRho program (35,36), kindly provided by Ying Wang. Genealogies are related through an ancestral recombination graph, and an algorithm based on a full-likelihood Bayesian Markov Chain Monte Carlo method estimates background mutation rates and hotspots. The times, in $4N_e$ units, was converted into years. For the coalescence process, $10^8$ simulations were performed.

In order to obtain TM RCA estimates that take recombination into account, we used the InterRho program (35,36), kindly provided by Ying Wang. Genealogies are related through an ancestral recombination graph, and an algorithm based on a full-likelihood Bayesian Markov Chain Monte Carlo method estimates background mutation rates and hotspots. The times, in $4N_e$ units, was converted into years. For the coalescence process, $10^8$ simulations were performed.

**SUPPLEMENTARY MATERIAL**

Supplementary Material is available at HMG online.

**ACKNOWLEDGEMENTS**

We wish to thank Dr Ying Wang (University of Chicago), who provided the InterRho program and helped in the calculation of recombination rates and TM RCA estimates. Also, we are grateful to Dr Franca Rosa Guerini (Fondazione Don C. Gnocchi, IRCCS, Milan) for helpful advice on genotyping.

**Conflict of Interest statement.** None declared.

**FUNDING**

M.C. is supported by grants from Istituto Superiore di Sanita’ ‘Programma Nazionale di Ricerca sull’ AIDS’, the EMPRO and AVIP EC WP6 Projects, the nGIN EC WP7 Project, the Japan Health Science Foundation, 2008 Research Finalizzata (Italian Ministry of Health), 2008 Ricerca Corrente (Italian Ministry of Health), Progetto FIRB RETI: Rete Italiana Chimica Farmaceutica CHEM-PROFARMA-NET (RBPR05NWWC) and Fondazione CARIPLO.

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

**REFERENCES**