

ORIGINAL ARTICLE

An evolutionary history of the selectin gene cluster in humans

M Fumagalli^{1,5}, M Fracassetti^{1,5}, R Cagliani¹, D Forni¹, U Pozzoli¹, GP Comi², F Marini¹, N Bresolin^{1,2}, M Clerici^{3,4} and M Sironi¹

Molecules involved in leukocyte trafficking have a central role in the development of inflammatory and immune responses. We performed F_{ST} analysis of the selectin cluster, as well as of *SELPLG*, *ICAM1* and *VCAM1*. Peaks of significantly high population genetic differentiation were restricted to two regions in *SELP* and one in *SELPLG*. Resequencing data indicated that the region covering *SELP* exons 11–13 displays high nucleotide diversity in Africans and Europeans (CEU), and a high level of within-species diversity compared with inter-specific divergence. Analysis of inferred haplotypes revealed a complex phylogeny with two deeply separated clades that coalesce at ~3.5 million years (MY) plus a minor clade with a TMRCA (time to the most recent common ancestor) of ~2.2 MY. A splicing assay indicated no haplotype-specific effect on *SELP* exon 14 inclusion. These data are consistent with a model of multiallelic balancing selection; single-nucleotide polymorphism analysis indicated that the Val640Leu variant represents a likely selection target. In populations of Asian ancestry a distinct haplotype, possibly carrying regulatory variants, has been driven to high frequency by positive selection. No deviation from neutrality was observed for the *SELPLG* region. Resequencing of *SELP* in chimpanzees revealed a haplotype phylogeny with extremely deep basal branches, suggesting either long-standing balancing selection or ancestral population structure. Thus, *SELP* has experienced a complex selective history, possibly as a result of local adaptation. Variants in the gene have been associated with autoimmune and cardiovascular diseases. Association studies would benefit from both taking the complex *SELP* haplotype structure into account and from analysis of possible regulatory variants in the gene.

Heredity (2012) 109, 117–126; doi:10.1038/hdy.2012.20; published online 2 May 2012

Keywords: selectin gene cluster; *SELP*; balancing selection; local adaptation

INTRODUCTION

Leukocyte interaction with the vascular endothelium has a central role in the development of inflammatory and immune responses. Thus, adhesion molecules involved in leukocyte trafficking have received considerable interest, especially in relation to the development of human diseases. The most well-studied molecules are the selectins, ICAM1 (intercellular adhesion molecule 1) and VCAM1 (vascular cell adhesion molecule 1).

Selectins are a family of C-type lectins, and humans express three family members, namely P-, E- and L-selectin, encoded by genes (*SELP*, *SELE* and *SELL*) located in cluster on chromosome 1. By binding to their ligands, selectins have a major role in leukocyte tethering and rolling on activated epithelia, thus favouring their extravasation and migration to inflamed tissues. The prefix for the three selectins carries a hint at their main expression site: L- and E-selectins are expressed on most leukocytes and on activated endothelial cells, respectively, whereas P-selectin is stored in platelet alpha granules and endothelial Weibel-Palade bodies (Ley, 2003). The three selectins have a common ligand, encoded by *SELPLG*. Exposure to different pro-inflammatory and pro-thrombotic stimuli determines

the rapid translocation of *SELP* to the cell surface where, through interaction with *SELPLG*, it promotes adhesion and rolling of both leukocytes and platelets, as well as their heterotypic aggregation (Ley, 2003). Inflammatory cytokines also induce E-selectin transcription, whereas L-selectin is constitutively expressed (Ley, 2003). Experiments in knockout mouse models, as well as data from rare human conditions where leukocyte binding to selectins is impaired, indicate that *SELP*, *SELL* and *SELE* have a central role in the trafficking of innate immunity cells, T lymphocytes, and platelets (Ley, 2003).

ICAM1 and VCAM1 belong to the immunoglobulin superfamily; inflammatory cytokines also trigger the expression of both ICAM1 and VCAM1 on endothelial cells to increase leukocyte adherence and migration (Golias *et al.*, 2007).

Selectins, as well as ICAM1 and VCAM1, also exist in soluble forms. The circulating amount of P-selectin and ICAM1 is influenced by polymorphisms located at their respective loci, and by variants in *ABO* (Miller *et al.*, 2004; Reiner *et al.*, 2008; Barbalic *et al.*, 2010).

Leukocyte extravasation and tissue infiltration are necessary for efficient immune response against invading pathogens; yet, these processes may also result in tissue damage and promote chronic

¹Scientific Institute IRCCS E. Medea, Bioinformatics Lab., Bosisio Parini (LC), Italy; ²Department of Neurological Sciences, Dino Ferrari Centre, University of Milan, Fondazione Ca' Granda IRCCS Ospedale Maggiore Policlinico, Milano, Italy; ³Department of Biomedical Sciences and Technologies LITA Segrate, University of Milan, Milano, Italy and ⁴Department of Molecular Medicine, Fondazione Don C. Gnocchi, IRCCS, Milano, Italy

⁵These authors equally contributed to this work.

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

E-mail: manuela.sironi@bp.inf.it

Received 31 October 2011; revised 7 February 2012; accepted 19 March 2012; published online 2 May 2012

inflammatory states. These observations suggest on one hand that genes encoding adhesion molecules may represent natural selection targets, as is the case of several immune response loci (Sironi and Clerici, 2010); on the other hand, genetic variation at these loci might protect from or predispose to autoimmune or chronic inflammatory conditions. Indeed, polymorphisms in *SELP* and *SELE* have been associated with susceptibility to systemic lupus erythematosus (Jacob *et al.*, 2007), whereas variants in *SELL* and *SELE* predispose to IgA nephropathy (Watanabe *et al.*, 2006). *SELP* also represents a risk locus for atopy (Bourgain *et al.*, 2003).

The increasing availability of large-scale genetic diversity data, which are being produced by next-generation sequencing platforms, opens unprecedented possibilities to the study of natural selection signatures along extended genomic regions. Still, the nature of these data also requires the development of new strategies to infer selective patterns. Here, we aimed to provide a description of the evolutionary history of the selectin gene cluster, by integrating low-coverage genome-wide resequencing data with classic population genetics analyses. Similarly, the genes encoding *ICAM1*, *VCAM1* and *SELPLG* were analyzed. Results indicate that a complex selective scenario in distinct human populations has shaped the genetic diversity of the *SELP* gene regions where autoimmune disease risk alleles are located.

MATERIALS AND METHODS

Samples and sequencing

Human genomic DNA from HapMap subjects (20 individuals of Asian descent) was obtained from the Coriell Institute for Medical Research. All analyzed regions were PCR-amplified and directly sequenced; primer sequences are available upon request. PCR products were treated with ExoSAP-IT (USB Corporation, Cleveland, OH, USA), directly sequenced on both strands with a Big Dye Terminator sequencing Kit (v3.1 Applied Biosystems, Foster City, CA, USA) 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. The genomic DNA of nine *Pan troglodytes* was obtained from the Gene Bank of Primates, Primate Genetics, Göttingen, Germany (<http://dpz.eu/index.php>). These samples have been shown to belong to the *Pan troglodytes verus* subspecies (Cagliani *et al.*, 2012). Primer pairs are available upon request.

Population genetic analyses

Tajima's *D* (Tajima, 1989), Fu and Li's *D** and *F** (Fu and Li, 1993) statistics, as well as diversity parameters θ_W (Watterson, 1975) and π (Nei and Li, 1979) were calculated as previously described (Fumagalli *et al.*, 2009). Calibrated coalescent simulations were performed using the *cost* package (Schaffner *et al.*, 2005) and its best-fit parameters for African American (AA), CEU and Asian (AS) populations with 10 000 iterations. Coalescent simulations were conditioned on mutation rate and recombination rate was derived from University of California Santa Cruz (UCSC) tables (<http://genome.ucsc.edu/>, [snpRecombRateHamap](http://genome.ucsc.edu/snpRecombRateHamap) table).

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

The DIND (derived intra-allelic nucleotide diversity) test was performed as previously proposed (Barreiro *et al.*, 2009). Significance thresholds for each MAF interval were computed via coalescent simulations incorporating a demographic model (Schaffner *et al.*, 2005).

Genotype data for 5 kb regions from 238 resequenced human genes were derived from the NIEHS SNPs Program web site (<http://egp.gs>

washington.edu). In particular, we selected genes that had been resequenced in populations of defined ethnicity including CEU, AA and AS (NIEHS panel 2). The iES (integrated extended haplotype homozygosity) score was calculated as previously described (Tang *et al.*, 2007).

Details on F_{ST} statistic calculation among continental groups in the HGDP-CEPH panel are available as Supplementary Methods. Genotype information was obtained for the *SELP* regions and for 2000 randomly selected RefSeq genes. Sliding-window analysis was performed on overlapping 20 single-nucleotide polymorphism (SNP) windows moving with a step of 5 SNPs.

Haplotype analysis and TMRCA calculation

Haplotypes were inferred using PHASE (version 2.1) (Stephens and Scheet, 2005). Details on haplotype analysis and TMRCA calculation are available as Supplementary Methods.

Minigene construction, transfection and transcript analysis

The genomic DNA of NA10851 HapMap subject (heterozygous for haplotypes belonging to the two main clades) was used as a template to PCR amplify a genomic fragment encompassing *SELP* exons 11–16. PCR amplifications were performed with high-fidelity polymerase (*Pfu* DNA Polymerase, Promega Corporation, Madison, WI, USA). The ~6500 bp PCR products were cloned in the pDisplay expression vector (Invitrogen, Carlsbad, CA, USA) to generate the pDisplay-*SELP*-HapA and pDisplay-*SELP*-HapB constructs. The two plasmids were checked by direct DNA sequencing using primers located on the pDisplay vector (HA-For: 5'-GATGTTCCAGATTATGCTG-3'; MYC-Rev: 5'-CAGATCCTCTTCTGAGATG-3').

HeLa cells were maintained in DMEM medium supplemented with 10% fetal-calf serum and transfected using Lipofectamine-2000 (Invitrogen) in accordance with the manufacturer's instructions. Cells were harvested 48 h after transfection and total RNA was extracted by TRIzol reagent (Invitrogen), following the manufacturer's protocols. Purified RNA was reverse transcribed to cDNA using random hexamers and Ready-To-Go You-Prime First-Strand Beads (GE Healthcare, Little Chalfont, Buckinghamshire, UK). RT-PCR was performed with *SELP* primers located in exon 12 and 16. PCR products were separated on a 2% agarose gel.

RESULTS

Sliding-window analysis of the selectin cluster

In humans, three selectin genes (*SELP*, *SELL* and *SELE*) are located in a ~150 kb region on chromosome 1. In order to analyze population genetic differentiation along the selectin cluster, we exploited data from the 1000 Genomes Pilot project deriving from the low-coverage whole-genome sequencing of 179 individuals with different ancestry (1000 Genomes Project Consortium *et al.*, 2010). In particular, individuals from three distinct ethnic groups have been sequenced: CEU, Yoruba from Nigeria (YRI) and Japanese plus Chinese (East Asian, AS). Using these data we calculated F_{ST} (Wright, 1950) in sliding-windows moving along the chromosome 1 genomic region where *SELE*, *SELL* and *SELP* are located. As an empirical comparison, the same procedure was applied to 2000 randomly selected human genes, so as to calculate percentiles of genetic differentiation for pairwise population comparisons. As shown in Figure 1, prominent peaks in F_{ST} occur within the *SELP* gene region. Specifically, AS/YRI population genetic differentiation exceeds the 95th percentile in two regions covering exons 9–10 and 11–13. In these two regions F_{ST} is also high for CEU/AS and YRI/CEU comparisons (Figure 1). The 1000 Genomes Pilot project data have an advantage over other data sets (for example, HapMap and the HGDP-CEPH panel) because they do not rely on the genotyping of selected SNPs, resulting in a less-marked bias towards intermediate frequency variants. Still, these data suffer from lack of power for rare variants (1000 Genomes Project Consortium *et al.*, 2010), and possible biases within repetitive elements (Cagliani *et al.*, 2012). Thus, we repeated F_{ST} analysis using

both HapMap data and genotype information from the HGDP-CEPH panel: we obtained results very similar to those from the 1000 Genomes Pilot project data (Figure 1 and Supplementary Figure 1). These were also confirmed by F_{ST} analysis using Sanger resequencing data for SELP, SELL and SELE from the SeattleSNPs Variation Discovery Resource (Supplementary Figure 2). Finally, we calculated the iES (Tang *et al.*, 2007) along the selectin cluster without identifying regions with unusual score (Supplementary Figure 3). These results warrant further analysis on the selection pattern of the two SELP regions, hereafter referred to as $SELP_{ex9-10}$ and $SELP_{ex11-13}$.

Sliding-window analysis of SELPLG, ICAMI and VCAMI

The same approach described above was applied to analyze F_{ST} along SELPLG, ICAMI and VCAMI. No unusual F_{ST} value was observed in ICAMI and VCAMI using data from the 1000 Genomes Pilot project; conversely a region of high genetic differentiation between YRI and AS was observed in the 5' portion of SELPLG (hereafter referred to as $SELPLG_{5'}$) (Figure 2). High genetic differentiation was not detected using the HGDP-CEPH panel data (Figure 2), possibly because of the small number of SNPs genotyped in SELPLG; similarly, analysis of HapMap SNPs (Supplementary Figure 4) revealed no unusual F_{ST} value in the gene, nor did data derived from SeattleSNPs (Supplementary Figure 5). Yet, in this latter case the analysis was restricted to CEU and AA, as no AS individual was resequenced by the Program.

Population genetics of SELP and SELPLG in human populations

In order to perform further population genetics analyses we turned to Sanger resequencing data. As mentioned above, the SELP and SELPLG transcription units have been fully resequenced in CEU and AA by the

Seattle SNPs Variation Discovery Resource. As for AS, we resequenced $SELP_{ex9-10}$, $SELP_{ex11-13}$ and $SELPLG_{5'}$ in 20 HapMap AS individuals. These data were used to calculate θ_W , an estimate of the expected per site heterozygosity (Watterson, 1975) and π , the average number of pairwise sequence nucleotide differences between haplotypes (Nei and Li, 1979). In order to compare the values we obtained for the SELP and SELPLG regions, we calculated θ_W and π for 5 kb windows (hereafter referred to as reference windows) deriving from 238 genes resequenced by the NIEHS Program in the same populations; the percentile rank corresponding to the SELP and SELPLG regions in the distribution of reference windows is reported in Table 1. No exceptional θ_W or π value was obtained for $SELP_{ex9-10}$ and $SELPLG_{5'}$ in the three populations. Conversely, both indexes were extremely high in $SELP_{ex11-13}$ in AA and CEU, although in this latter population the rank of π did not reach the 95th percentile. Peaks of nucleotide diversity in this region were also obtained for YRI using a sliding-window approach with the 1000 Genomes Pilot Project data (Supplementary Figure 6)

High nucleotide diversity might be suggestive of balancing selection, as neutral variants tend to be maintained with the selected allele(s). In order to formally rule out the possibility that the excess of nucleotide diversity is due to higher local mutations rates, we applied a MLHKA test (Wright and Charlesworth, 2004). This is based on the concept whereby, under neutral evolution, the amount of within- and between-species diversity is expected to be similar for all loci in the genome. Therefore, the test compares polymorphism and divergence levels for a region of interest with those calculated for other neutrally evolving genomic segments. Sixteen reference loci were used for the analysis (Fumagalli *et al.*, 2009), and the results for $SELP_{ex11-13}$ indicated a significant excess

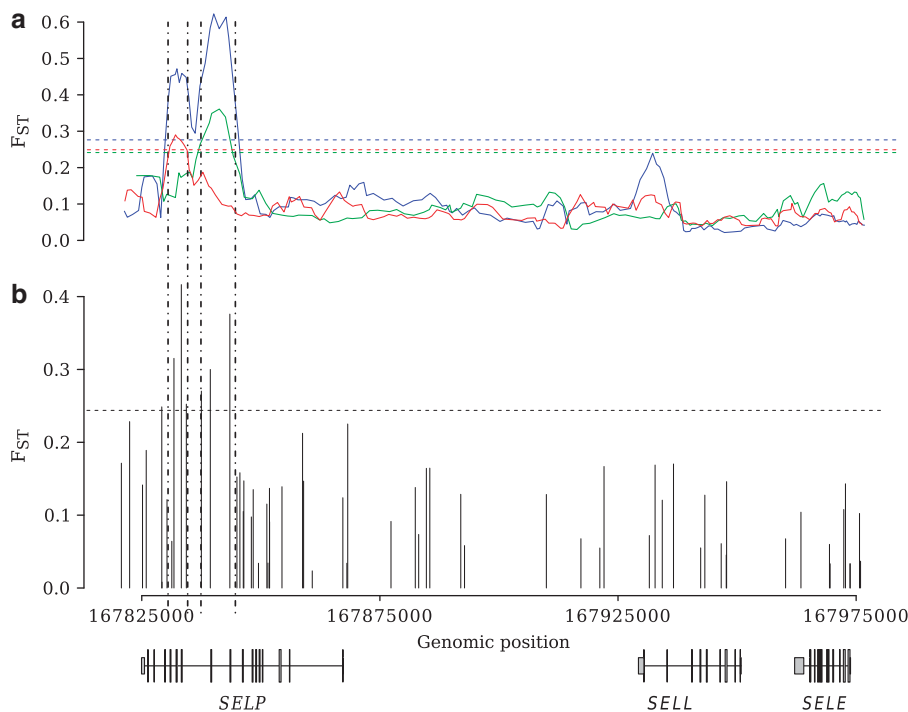


Figure 1 Sliding-window analysis of F_{ST} along the selectin cluster. (a) Data from the 1000 Genomes Pilot Project were used to calculate F_{ST} in sliding windows of 20 SNPs moving along the selectin gene cluster with a step of 5 SNPs. Color codes refer to population comparisons: red, YRI/CEU; blue, YRI/AS; green, CEU/AS. Horizontal dashed lines represent the 95th percentile in the distribution of F_{ST} calculated for sliding windows deriving from 2000 randomly selected human genes. (b) The location of SNPs genotyped in the HGDP-CEPH panel is reported with the value of F_{ST} calculated across continental groups. The horizontal line corresponds to the 95th percentile in the distribution of all SNPs genotyped in the panel.

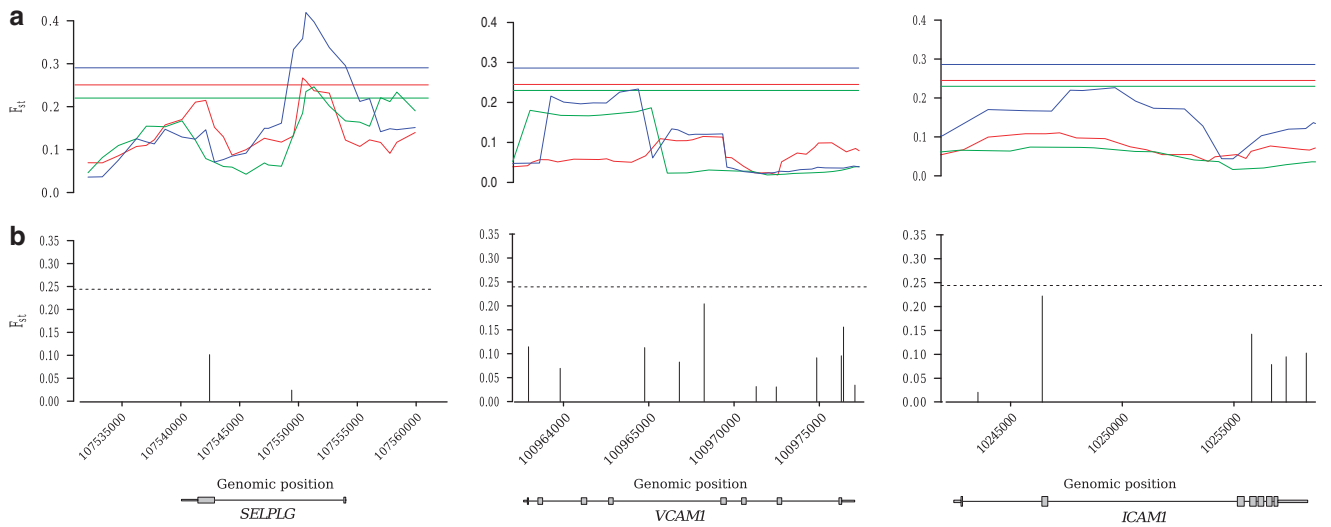


Figure 2 Sliding-window analysis of F_{ST} for *SELPLG*, *ICAMI* and *VCAMI*. F_{ST} calculation and color codes as in Figure 1. (a) Data from the 1000 Genomes Pilot Project. (b) Data from the HGDP-CEPH panel.

Table 1 Nucleotide diversity and MLHKA test for the two analyzed *SELP* and *SELPLG* regions

Gene region	Size (Kb)	Pop.	N ^a	S ^b	$\theta_W (\times 10^{-4})$		$\pi (\times 10^{-4})$		MLHKA	
					Value	Rank ^c	Value	Rank ^c	k ^d	P-value
<i>SELP_{ex9-10}</i>	6.6	AA	48	37	12.55	0.81	11.13	0.81	1.92	0.11
		CEU	46	18	6.51	0.64	11.72	0.88	1.24	0.10
		AS	40	15	4.97	0.42	1.55	0.07	1.04	0.13
<i>SELP_{ex11-13}</i>	4.1	AA	48	52	28.88	>0.99	27.52	0.99	5.95	2.3×10^{-6}
		CEU	46	27	15.14	0.98	11.75	0.88	4.21	4.2×10^{-4}
		AS	40	15	8.69	0.87	6.12	0.63	2.83	0.012
<i>SELPLG₅</i>	2.7	AA	48	18	15.06	0.91	10.54	0.79	2.46	0.062
		CEU	46	13	10.99	0.90	9.64	0.83	2.57	0.064
		AS	40	7	6.11	0.74	7.51	0.75	1.59	0.84

Abbreviations: AA, African American; AS, Asian; CEU, Europeans; MLHKA, maximum-likelihood-ratio Hudson-Kreitman-Aguadé; Pop., population.

^aSample size (chromosomes).

^bNumber of segregating sites.

^cPercentile rank relative to a distribution of 5 kb windows from NIEHS genes.

^dSelection parameter.

of polymorphism compared with divergence in all populations (Table 1). No deviation from neutrality was observed for the *SELP_{ex9-10}* and *SELPLG₅* regions (Table 1).

As mentioned above, natural selection acting on specific gene regions can determine a distortion in the site frequency spectrum. Common neutrality tests based on the SFS include Tajima's D (D_T) (Tajima, 1989), and Fu and Li's D^* and F^* (Fu and Li, 1993). D_T tests the departure from neutrality by comparing θ_W and π , and positive values indicate an excess of intermediate frequency variants. Fu and Li's F^* and D^* are also based on the site frequency spectrum (SFS) and differ from D_T in that they also take into account whether mutations occur in external or internal branches of a genealogy. As population history, in addition to selective processes, is known to affect the SFS, we evaluated the significance of neutrality tests by performing coalescent simulations that incorporate demographic

scenarios (see methods) (Schaffner *et al.*, 2005). As explained above, we also applied an empirical comparison by calculating the percentile rank of D_T , F^* and D^* for the *SELP* and *SELPLG* gene regions relative to reference windows. No deviation from neutrality was observed in AA, whereas D_T and F^* were significantly high in *SELP_{ex9-10}* for CEU (Table 2). In this population, D^* was extremely high in *SELP_{ex11-13}*. As for AS, SFS-based statistics tended to display low values in both *SELP* regions (Table 2). These data were confirmed by running coalescent simulations with different demographic models (Supplementary Table 1). No departure from neutrality was observed for *SELPLG₅*.

SELP haplotype analysis and selective pattern in distinct populations

The power of SFS-based statistics to reject neutrality depends on several factors such as the timing and strength of the selective event,

Table 2 SFS-based statistics for the analyzed *SELP* and *SELPLG* gene regions

Gene region	Pop. ^a	Tajima's D			Fu and Li's D*			Fu and Li's F*			Fay and Wu's H	
		Value	Rank ^a	P value ^b	Value	Rank ^a	P value ^b	Value	Rank ^a	P value ^b	Value	P value ^b
<i>SELP</i> _{ex9-10}	AA	-0.38	0.58	0.35	-0.77	0.38	0.22	-0.75	0.41	0.31	1.64	0.44
	CEU	2.57	0.98	0.004	1.28	0.93	0.03	2.04	0.98	0.005	-0.54	0.39
	AS	-2.17	0.005	0.001	-1.65	0.15	0.048	-2.14	0.09	0.023	-17.83	<0.001
<i>SELP</i> _{ex11-13}	AA	-0.16	0.71	0.28	-1.66	0.14	0.12	-1.33	0.24	0.22	1.39	0.38
	CEU	-0.75	0.28	0.19	1.78	0.99	<0.001	1.07	0.83	0.11	-13.71	<0.001
	AS	-0.94	0.23	0.11	-3.19	0.02	0.005	-2.89	0.03	0.010	-6.93	0.007
<i>SELPLG</i> ₅	AA	-0.95	0.30	0.28	-1.13	0.28	0.25	-1.26	0.27	0.23	2.01	0.06
	CEU	-0.37	0.42	0.36	-1.49	0.15	0.12	-1.32	0.19	0.15	-1.79	0.12
	AS	0.63	0.70	0.30	0.51	0.75	0.32	0.64	0.75	0.29	-0.16	0.35

Abbreviations: AA, African American; AS, Asian; CEU, Europeans; Pop., population.

^aPercentile rank relative to a distribution of 5 kb windows from NIEHS genes.

^bP-value obtained through coalescent simulations.

and complex selective patterns may leave signatures that are difficult to detect. Thus, to gain further insight into the evolutionary history of *SELP*_{ex11-13}, we performed haplotype analysis using two approaches: a neighbor-joining network (Bandelt *et al.*, 1999) and a maximum-likelihood coalescent method implemented in GENETREE (Griffiths and Tavare, 1995). This latter indicated that the haplotype genealogy of *SELP*_{ex11-13} is split into two major clades (clades A and B) with a coalescence time of 3.51 million years (MY) (SD: 0.569 MY) (Figure 3). Both clades are further split into relatively deep subclades, one of them, in clade A, showing a TMRCA of more than 2.2 MY. As shown in the haplotype network (Figure 3), no AS chromosome contributes to clade B, whereas all other major clade A haplotypes are represented across all populations. These analyses, and the deep coalescence time of *SELP*_{ex11-13} haplotypes suggest that this region has been a target of multiallelic balancing selection in CEU and AA. Conversely, the low diversity levels observed in AS for the two *SELP* regions indicate that a selective sweep may have occurred in this population. Indeed, calculation of Fay and Wu's H revealed significantly negative values for both *SELP*_{ex9-10} ($H = -17.83$, $P < 0.01$, Table 2) and *SELP*_{ex11-13} ($H = -6.93$, $P = 0.01$, Table 2). In order to further explore this possibility, we applied the DIND test (Barreiro *et al.*, 2009). This is based on the ratio of intra-allelic diversity associated with the ancestral and derived alleles ($i\pi_A/i\pi_D$) plotted against the frequency of the derived allele (Figure 4): a high value of $i\pi_A/i\pi_D$ for variants with high frequency of the derived allele is suggestive of positive selection, as the neutral diversity associated with the derived allele is limited despite its high frequency in the population. We applied the DIND test to *SELP*_{ex9-10} and *SELP*_{ex11-13}; again, statistical significance was calculated by coalescent simulations that incorporate a demographic model (see methods). Significantly, high values of DIND were obtained in *SELP*_{ex9-10} for rs3917750, rs3917751, rs3766126, rs2223303, rs2205894 and rs3917777 (Figure 4). These SNPs define a haplotype with a frequency of 0.95, 0.45 and 0.10 in AS, CEU and AA, respectively. Visualization of haplotypes in this region indicated that the derived alleles of the 5 SNPs reported above define a set of highly homogeneous haplotypes in both CEU and AS, as also assessed by calculation of nucleotide diversity for the distinct haplotype groups (Figure 4).

The soluble form of P-selectin is thought to result both from alternative splicing of the transmembrane domain and from shedding

of the membrane-bound form (Vestweber and Blanks, 1999). Given that the transmembrane domain is encoded by exon 14, we wished to verify whether the two major haplogroups we observed for *SELP*_{ex11-13} carry variants with an effect on exon 14 inclusion in *SELP* transcripts. To this aim, we constructed *SELP* minigenes. In particular, DNA fragments spanning exons 11–16 were cloned in an expression vector. Two distinct fragments were cloned, deriving from chromosomes carrying clade A or clade B haplotypes. RT-PCR amplification of *SELP* transcripts isolated from HeLa cells transfected with both minigenes generated similar amounts of two products: a 433 bp transcript deriving from the full-length *SELP* and a 323 bp PCR fragment corresponding to exon 14 skipping. Albeit not quantitative, these results suggest that exon 14 skipping is not associated with a specific haplotype.

Population genetics of *SELP* in chimpanzee

To analyze *SELP* evolutionary history in *Pan troglodytes*, we resequenced *SELP*_{ex9-10} and *SELP*_{ex11-13} in nine unrelated chimpanzees. No polymorphic variant shared with humans was observed. The total number of segregating sites was 39 and 33 in *SELP*_{ex9-10} and *SELP*_{ex11-13}, respectively; their analysis indicated that most nonsynonymous substitutions (five out of six) occur in exons 12 and 13 (Supplementary Table 2). In order to assess whether the two *SELP* regions show unusual levels of nucleotide variability, we resequenced 16 genomic regions in these same individuals to be used as an empirical comparison (Cagliani *et al.*, 2012). Calculation of θ_W that *SELP*_{ex9-10} and *SELP*_{ex11-13} show by far the highest values of diversity compared with the 16 reference regions (Figure 5); as for π , the value for *SELP*_{ex11-13} was also the highest. This result was confirmed by application of the MLHKA test, which yielded significant results ($k = 3.81$ and 7.72 , $P = 0.0021$ and 5.41×10^{-6} for *SELP*_{ex9-10} and *SELP*_{ex11-13}, respectively). Calculation of SFS statistics indicated that both *SELP* regions display the lowest negative values for D_T , D^* and F^* (Supplementary Table 3). Further analysis indicated that variants in *SELP*_{ex9-10} and *SELP*_{ex11-13} are in tight linkage disequilibrium in the chimpanzee sample (Supplementary Figure 7), and one single individual carries a haplotype, which is highly divergent from those observed among the remaining eight primates. This translates in a haplotype genealogy with extremely long basal branches and a deep coalescent time. GENETREE analysis yielded a maximum-likelihood

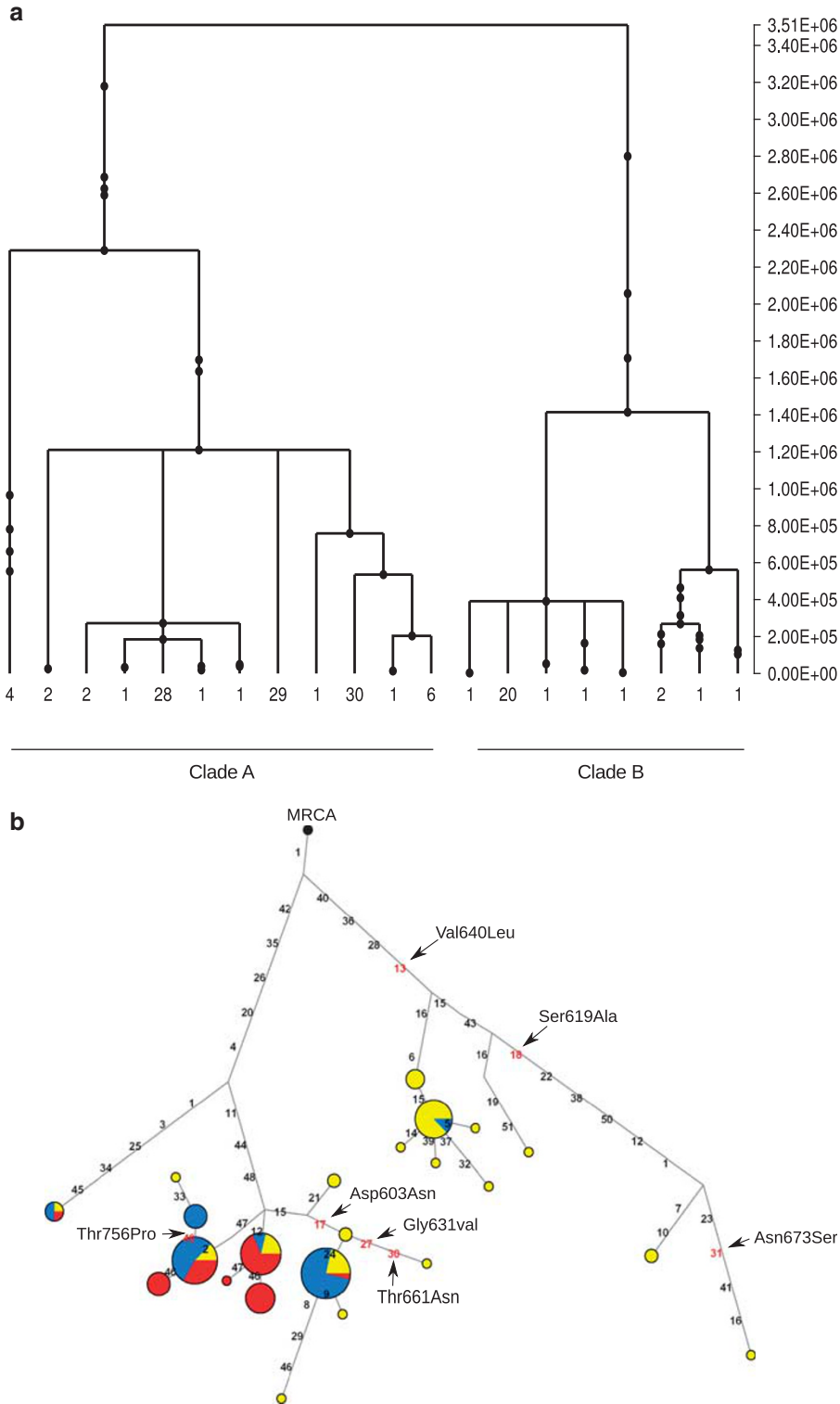


Figure 3 Haplotype analysis for *SELP_{ex11-13}*. **(a)** GENETREE: mutations are represented as black dots. The absolute frequency of each haplotype is reported. **(b)** Median joining network: each node represents a different haplotype, with the size of the circle proportional to frequency. Branch lengths are proportional to the number of nucleotide differences. Circles are color-coded according to population (yellow: AA, blue: CEU, red: AS). The most recent common ancestor (MRCA) is also shown (black circle). Nonsynonymous substitutions are shown in red and indicated by the arrows.

estimate of θ (θ_{ML}) of 9.2, resulting in an effective population size (N_e) of 14081. Using this method, the TMRCA of the chimpanzee *SELP* haplotype lineages amounted to 3.52 MY (SD: 424 KY).

DISCUSSION

In this study, we analyzed the evolutionary history of genes encoding major molecules involved in leukocyte–epithelium interaction. A

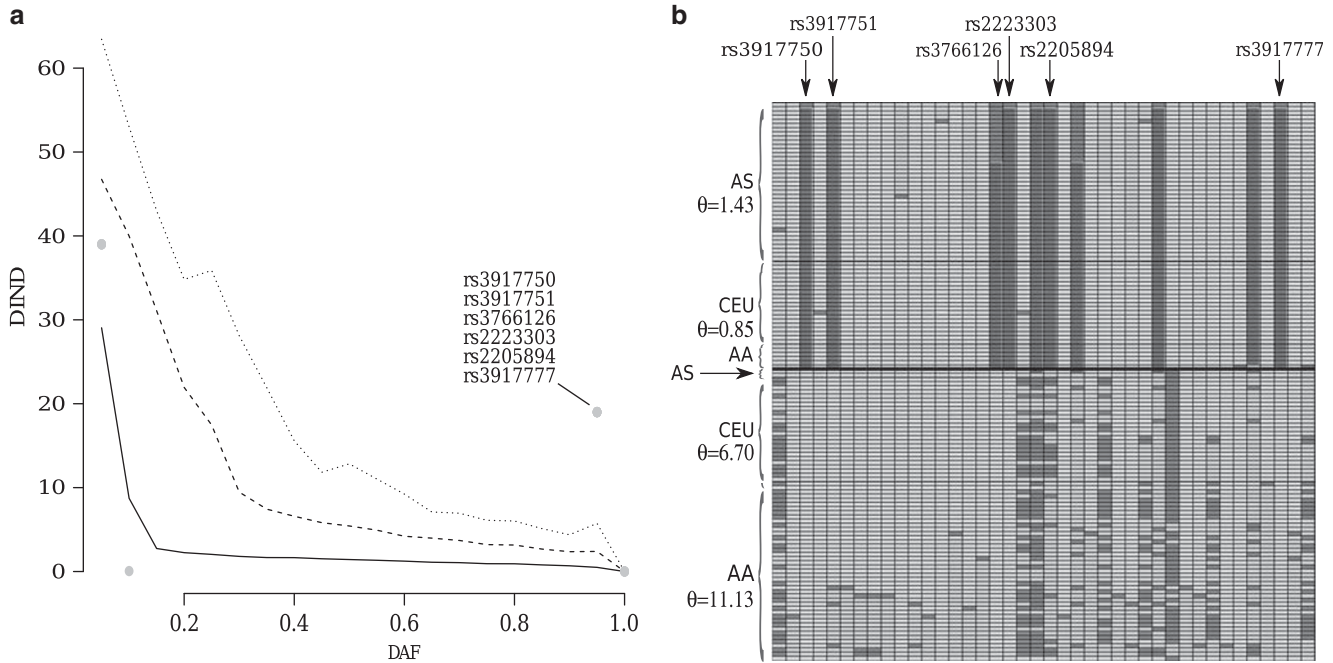


Figure 4 Haplotype analysis for *SELP_{ex9-10}*. (a) The DIND test was performed for AS. *P*-values were calculated through coalescent simulations that incorporate demographic scenarios (Schaffner *et al.*, 2005). The continuous, dashed and dotted lines indicate the median, the 95th and the 99th percentiles, respectively. (b) Visualization of *SELP_{ex9-10}* haplotypes; polymorphic positions are color-coded according to their allelic state (dark gray: derived, light gray: ancestral). The position of the 5 SNPs identified through the DIND test are shown and inferred haplotypes are ordered on the basis of their allelic status.

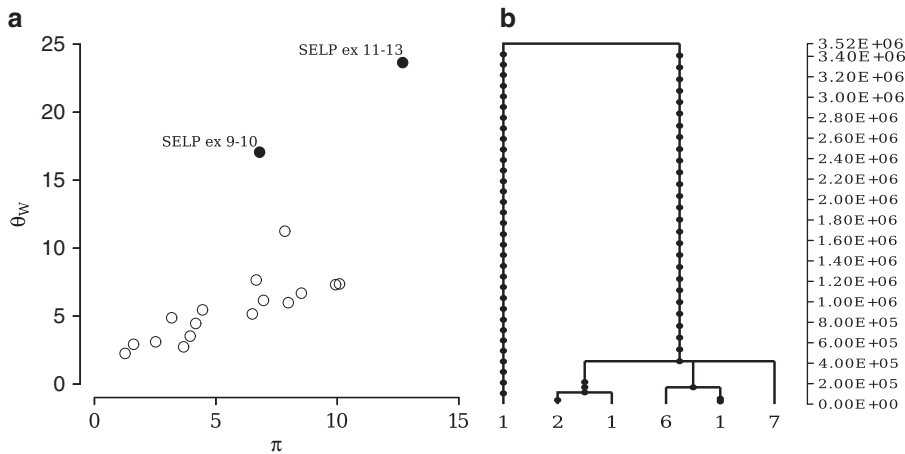


Figure 5 Nucleotide diversity and haplotype analysis in chimpanzee. (a) Plot of θ_W and π values for the 16 regions we resequenced (white circles) and for the two *SELP* regions (black). (b) GENETREE analysis of the *SELP* region covering exons 9–13: mutations are represented as black dots and the absolute frequency of each haplotype is reported.

previous analysis of immune response genes with available sequences in public databases (Casals *et al.*, 2011) included *ITGAL* and *ITGA2*, encoding leukocyte integrin molecules also involved in white blood cell trafficking; the two genes were found to be neutrally evolving. To our knowledge, no other study has addressed the evolutionary history of adhesion molecules involved in leukocyte trafficking in human populations. Here, we exploited the availability of the 1000 Genomes Project low-coverage data to perform a sliding-window analysis of F_{ST} along a ~150 kb region encompassing *SELL*, *SELE* and *SELP*. Peaks of significantly high population genetic differentiation were identified

through an empirical approach based on the comparison with a large number of human genes, and were restricted to two regions in *SELP* where several amino-acid changes are located. A region in *SELP* also showed high population genetic differentiation. Extremely high values of F_{ST} strongly suggest the action of local adaptation, as allele frequencies are driven to differ more than that would be expected on the basis of drift or demography alone. Thus, we focused on these regions, and, in order to overcome the possible biases in the SFS, which are introduced by next-generation sequencing approaches (1000 Genomes Project Consortium *et al.*, 2010), we analyzed Sanger

sequencing data to infer selection patterns. Results indicated that the genomic region covering *SELP*_{ex11–13} displays high nucleotide diversity in Africans and CEU, and a higher level of within-species diversity compared with inter-specific divergence. Analysis of inferred haplotypes in this region revealed the presence of two deeply separated clades with a TMRCA of ~ 3.5 MY. Coalescence times for neutrally evolving autosomal human loci range between 0.8 and 1.5 MY (Tishkoff and Verrelli, 2003); thus, the TMRCA of the *SELP* haplotype clades is much deeper than expected under neutrality. Indeed, clade A is further split into two lineages that coalesce at ~ 2.2 MY (Figure 3). Overall, these data are consistent with a model of multiallelic balancing selection and may explain the values of SFS-based statistics (Tajima's *D* and Fu and Li's *F** and *D**), which are not strikingly positive in *SELP*_{ex11–13}, as the skew toward intermediate frequency variants tends to be less marked in a multiallelic selection model than in the case of biallelic selection (Navarro and Barton, 2002). In this respect, it is worth mentioning that, although *SELP*_{ex9–10} showed high values for *D_T* and *F**, these features are not accompanied by a local increase in nucleotide diversity (which is a hallmark of balancing selection (Charlesworth, 2006)), suggesting that the balancing selection target(s) is located in *SELP*_{ex11–13}, and that linkage disequilibrium might cause the distortion of the SFS in exons 9–10. As for *SEPLG*, despite the high population genetic differentiation identified in the initial scan, no test suggested departure from neutrality, confirming the previous indication that high *F_{ST}* alone is not sufficient to demonstrate natural selection (Xue *et al.*, 2009).

Network analysis of *SELP*_{ex11–13} indicated that distinct haplotypes have different frequencies in the three human populations we analyzed, with AS chromosomes being absent from clade B; this explains the limited nucleotide diversity observed for this population, and suggests a different selective scenario for *SELP* in Asia (see below). Analysis of SNPs located on the branches of the haplotype network indicated that the Val640Leu (rs6133) polymorphism separates the two major haplotype clades and might therefore represent a selection target. The derived 640Leu allele, which is common in populations of African descent, has previously been associated with atopy, thrombo-embolic stroke and systemic lupus erythematosus, suggesting that it might affect some properties of SELP, which are relevant to immunologic/inflammatory functions (Bourgain *et al.*, 2003; Zee *et al.*, 2004; Morris *et al.*, 2009). The variant has also been shown to modulate circulating levels of soluble P-selectin in some studies but not in others (Reiner *et al.*, 2008; Barbalic *et al.*, 2010). As for the two subclades within A group haplotypes, all variants separating the two branches are non-coding, suggesting that the selection target might be represented by one or more regulatory variants. Two nonsynonymous variants, namely Thr756Pro (rs6136) and Asp603Asn (rs6127), differentiate major haplotypes within clade A (Figure 3). The position of these variants in the haplotype genealogy indicates that they do not represent selection targets or, at least, that the selective pressure acting on them is extremely weak or recent. Nonetheless, these variants have been the subjects of intense investigation, as they were reported to affect cardiovascular functions in humans, although results have often been inconsistent among studies and in populations with different ethnic origin (Tregouet *et al.*, 2002; Miller *et al.*, 2004; Volcik *et al.*, 2007). As already suggested by some authors (Tregouet *et al.*, 2002), results herein clearly show that haplotype-based association analysis, rather than single SNP genotyping, should be applied when *SELP* genetic variation is to be associated with human phenotypic traits. Indeed, the *SELP* haplotype structure is complex and coding variants (in addition to putative regulatory variants) define distinct major

haplotypes. These have extremely different frequencies in distinct human populations, also resulting in diverse statistical power to detect associations depending on sample ethnicity.

Given that a fraction of soluble P-selectin is thought to result from alternative splicing of exon 14 (Vestweber and Blanks, 1999), we also tested whether the two major haplotypes encompassing exons 11–13 might affect exon inclusion. *SELP* transcript levels in lymphoblastoid cell lines were low (not shown), not allowing quantification of the alternatively spliced product. For this reason we adopted a minigene approach with transfection in HeLa cells. On the one hand, this allows splicing to occur for the two haplotypes in the same conditions and genetic background, therefore removing possible confounding effects. On the other hand, a major drawback of this approach is that the minigene is expressed from an exogenous promoter, and increasing evidences suggest that polymerase elongation rate and promoter elements have an impact on alternative splicing regulation (Kornblihtt *et al.*, 2004). Under the experimental conditions we used, the two *SELP* haplotypes yielded similar amounts of full-length and alternatively spliced transcripts, suggesting no major effect of *SELP*_{ex11–13} variants in exon 14 inclusion. However, further analyses will be required to determine whether promoter elements, environmental cues or biochemical stimuli affect the haplotype-specific splicing of *SELP*.

As mentioned above, striking differences in allele/haplotype frequencies, which are reflected in high *F_{ST}* values, are commonly regarded as a signature of local adaptation. Our analysis strongly suggest that while genetic diversity has long been maintained in African and CEU populations owing to a balancing selection regime, a *SELP* haplotype has been driven to high frequency in AS populations as a result of positive selection. Indeed, AS populations show reduced diversity at the two analyzed *SELP* regions, especially *SELP*_{ex9–10}, and neutrality tests highlight an excess of low-frequency alleles. Calculation of Fay and Wu (2000), which is sensitive to the presence of high-frequency-derived alleles, yielded a strongly negative result in the region encompassing exons 9–10. Consistently, application of the DIND test confirmed the hypothesis of positive selection for *SELP* and identified a haplotype, which has been driven to a frequency of 0.95 in AS whereas not reaching 0.5 in either CEU or AA. In both AS and CEU populations, a set of highly homogeneous haplotypes is defined by the derived alleles of the 5 SNPs we identified through the DIND test, whereas haplotypes carrying ancestral alleles at these loci display much higher diversity (Figure 4). It is worth mentioning that the two *SELP* regions we analyzed are located nearby, and therefore cannot be thought of evolving independently. In principle, the loss of the balanced polymorphism at *SELP*_{ex11–13} in AS might explain the reduced diversity observed at *SELP*_{ex9–10} in this population; yet, this scenario is not expected to result in the presence of a set of highly homogeneous haplotypes. Thus, we suggest that the selection targets in the two *SELP* regions have been competing, and that the selective sweep in the region encompassing exons 9–10 has determined the loss of the balanced SNP(s) in AS and reduced its frequency in CEU. In this respect, it is also worth mentioning that no significant *iES* score was detected in the region encompassing *SELP*_{ex9–10}; yet, simulation analyses have indicated that DIND has more power than haplotype-based tests for frequencies of the selected haplotype ranging from 10 to 98% (Barreiro *et al.*, 2009). Variants defining the *SELP*_{ex9–10} high frequency haplotype are all non-coding and inspection of available resequencing databases (for example, 1000 Genome Project; NIEHS Exome Project) indicated that they are not in linkage disequilibrium with any coding variant. Thus, the selection target in AS populations might be represented by a regulatory variant that possibly affects *SELP* expression.

Overall, these data suggest that distinct human *SELP* variants have been targeted by natural selection possibly as a response to different environmental pressures. In an attempt to analyze the selective history of P-selectin in our closest living relatives, we sequenced the region spanning exons 9–13 in nine unrelated *Pan troglodytes*. Our results indicated that *SELP* displays extremely high diversity when compared with other loci resequenced in the same individuals. Yet, this feature is due to the presence of a single highly divergent haplotype that originates a gene phylogeny with extremely deep basal branches that coalesce at 3.5 MY. This estimated TMRCA largely pre-dates the split of the chimpanzee and bonobo lineages, as well as the dispersal of modern chimpanzee populations (Bequet *et al.*, 2007). Thus, these data lend themselves to at least two possible explanations: either distinct *SELP* haplotypes have been maintained in these primates as a result of long-standing balancing selection or the ancestral primate population was structured. Indeed, archaic population structure with migration can determine the occasional sampling of loci with deep TMRCA from extant populations. Unfortunately, the still limited knowledge of chimpanzee demographic history, and the small number of sequenced individuals complicate the interpretation of this result. Nonetheless, it is interesting to notice that in both humans and chimpanzees several amino-acid replacements tend to cluster in the region covering exons 11–13, although in either species the MK test was not significant when orangutan or macaque were used as outgroups (not shown). This gene region encodes the *SELP* sushi domains (also known as complement control protein modules, or short consensus repeats), which are thought to be relevant for the interaction between P-selectin and leukocytes (Ley, 2003). As mentioned above, the binding of *SELP* to SELPLG has a central role in mediating leukocyte tethering, rolling and extravasation to sites of acute and chronic inflammation. Thus, the selective pressure(s) acting on *SELP* may be related to its central role in the elicitation of immune and inflammatory responses, as infectious diseases are commonly considered a powerful selective force. *SELP* has been involved in the pathogenesis of severe malaria, a disease that is commonly regarded as the strongest selective pressure acting on the human genome (Pozzoli *et al.*, 2010). P-selectin contributes to the cytoadherence of infected erythrocytes to the brain microvasculature (Rowe *et al.*, 2009) and mice that specifically lack endothelial expression of *Selp* are protected against cerebral malaria (Combes *et al.*, 2004). The parasite ligand for P-selectin is thought to be represented by PfEMP1 (*Plasmodium falciparum* erythrocyte membrane protein 1), as the two proteins bind *in vitro* (Senczuk *et al.*, 2001). PfEMP1 is a highly variable protein which has a central role in parasite immune evasion by changing binding tropism. Thus, an interesting possibility is that the diversity of human (and possibly chimpanzee) *SELP* results, at least partially, from a host–pathogen genetic conflict with *Plasmodium* species that also accounts for local changes in allele frequency of human polymorphisms.

Immune and inflammatory responses are vital to organism survival, but need fine tuning to prevent autoimmunity and tissue damage. Interestingly, the interaction of *SELP* with SELPLG has been shown to contribute to the generation of tolerogenic dendritic cells (Urzainqui *et al.*, 2007), and P-selectin binding to pentraxin 3 activates a negative feed-back loop that decreases leukocyte recruitment at inflammation sites (Deban *et al.*, 2010). In line with these observations and with its role in immune responses, polymorphisms in *SELP* have been associated with autoimmune manifestations or chronic inflammatory diseases. Risk alleles for autoimmune conditions have been shown to be common targets of natural selection in humans (Sironi and Clerici, 2010), possibly because of their

conferring a selective advantage against ancestral infections. This hypothesis might apply to the Val640Leu SNP in *SELP*, which is a likely selection target and has been associated with atopy and systemic lupus erythematosus (Bourgain *et al.*, 2003; Morris *et al.*, 2009). Nonetheless, data herein suggest that further genetic association studies for autoimmune diseases or other phenotypic traits would benefit from taking the complex *SELP* haplotype structure into account and from the analysis of possible regulatory variants in the gene, as they might have represented additional targets of natural selection.

Data archiving

Data have been deposited at Dryad: 10.5061/dryad.211vn8m4.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by the Italian Ministry of Health (grant: RF-IEM-2007-633627). We wish to thank Dr Christian Roos from the Gene Bank of Primates, Primate Genetics (Germany), for sending chimpanzee genomic DNA.

- 1000 Genomes Project Consortium Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Durbin RM *et al.* (2010). A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073.
- Bandelt HJ, Forster P, Rohl A (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16**: 37–48.
- Barbalic M, Dupuis J, Dehghan A, Bis JC, Hoogeveen RC, Schnabel RB *et al.* (2010). Large-scale genomic studies reveal central role of ABO in sP-selectin and sICAM-1 levels. *Hum Mol Genet* **19**: 1863–1872.
- Barreiro LB, Ben-Ali M, Quach H, Laval G, Patin E, Pickrell JK *et al.* (2009). Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet* **5**: e1000562.
- Bequet C, Patterson N, Stone AC, Przeworski M, Reich D (2007). Genetic structure of chimpanzee populations. *PLoS Genet* **3**: e66.
- Bourgain C, Hoffjan S, Nicolae R, Newman D, Steiner L, Walker K *et al.* (2003). Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus. *Am J Hum Genet* **73**: 612–626.
- Cagliani R, Guerini FR, Fumagalli M, Riva S, Agliardi C, Galimberti D *et al.* (2012). A *trans*-specific polymorphism in *ZC3HAV1* is maintained by long-standing balancing selection and may confer susceptibility to multiple sclerosis. *Mol Biol Evol* **29**: 1599–1613.
- Casals F, Sikora M, Laayouni H, Montanucci L, Muntasell A, Lazarus R *et al.* (2011). Genetic adaptation of the antibacterial human innate immunity network. *BMC Evol Biol* **11**: 202.
- Charlesworth D (2006). Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet* **2**: e64.
- Combes V, Rosenkranz AR, Redard M, Pizzolato G, Lepidi H, Vestweber D *et al.* (2004). Pathogenic role of P-selectin in experimental cerebral malaria: importance of the endothelial compartment. *Am J Pathol* **164**: 781–786.
- Deban L, Russo RC, Sironi M, Moalli F, Scanziani M, Zambelli V *et al.* (2010). Regulation of leukocyte recruitment by the long pentraxin PTX3. *Nat Immunol* **11**: 328–334.
- Fay JC, Wu CI (2000). Hitchhiking under positive Darwinian selection. *Genetics* **155**: 1405–1413.
- Fu YX, Li WH (1993). Statistical tests of neutrality of mutations. *Genetics* **133**: 693–709.
- Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi GP, Menozzi G *et al.* (2009). Widespread balancing selection and pathogen-driven selection at blood group antigen genes. *Genome Res* **19**: 199–212.
- Golias C, Tsoutsis E, Matziris A, Makridis P, Batistatou A, Charalabopoulos K (2007). Review. Leukocyte and endothelial cell adhesion molecules in inflammation focusing on inflammatory heart disease. *In Vivo* **21**: 757–769.
- Griffiths RC, Tavaré S (1995). Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Math Biosci* **127**: 77–98.
- Jacob CO, Reiff A, Armstrong DL, Myones BL, Silverman E, Klein-Gitelman M *et al.* (2007). Identification of novel susceptibility genes in childhood-onset systemic lupus erythematosus using a uniquely designed candidate gene pathway platform. *Arthritis Rheum* **56**: 4164–4173.
- Kornblihtt AR, de la Mata M, Fededa JP, Munoz MJ, Nogues G (2004). Multiple links between transcription and splicing. *RNA* **10**: 1489–1498.
- Ley K (2003). The role of selectins in inflammation and disease. *Trends Mol Med* **9**: 263–268.

- Miller MA, Kerry SM, Dong Y, Strazzullo P, Cappuccio FP (2004). Association between the Thr715Pro P-selectin gene polymorphism and soluble P-selectin levels in a multiethnic population in South London. *Thromb Haemost* **92**: 1060–1065.
- Morris DL, Graham RR, Erwig LP, Gaffney PM, Moser KL, Behrens TW *et al.* (2009). Variation in the upstream region of P-Selectin (SELP) is a risk factor for SLE. *Genes Immun* **10**: 404–413.
- Navarro A, Barton NH (2002). The effects of multilocus balancing selection on neutral variability. *Genetics* **161**: 849–863.
- Nei M, Li WH (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci USA* **76**: 5269–5273.
- Pozzoli U, Fumagalli M, Cagliani R, Comi GP, Bresolin N, Clerici M *et al.* (2010). The role of protozoa-driven selection in shaping human genetic variability. *Trends Genet* **26**: 95–99.
- Reiner AP, Carlson CS, Thyagarajan B, Rieder MJ, Polak JF, Siscovick DS *et al.* (2008). Soluble P-selectin, SELP polymorphisms, and atherosclerotic risk in European-American and African-American young adults: the Coronary Artery Risk Development in Young Adults (CARDIA) Study. *Arterioscler Thromb Vasc Biol* **28**: 1549–1555.
- Rowe JA, Claessens A, Corrigan RA, Arman M (2009). Adhesion of Plasmodium falciparum-infected erythrocytes to human cells: molecular mechanisms and therapeutic implications. *Expert Rev Mol Med* **11**: e16.
- Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* **15**: 1576–1583.
- Senczuk AM, Reeder JC, Kosmala MM, Ho M (2001). Plasmodium falciparum erythrocyte membrane protein 1 functions as a ligand for P-selectin. *Blood* **98**: 3132–3135.
- Sironi M, Clerici M (2010). The hygiene hypothesis: an evolutionary perspective. *Microbes Infect* **12**: 421–427.
- Stephens M, Scheet P (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet* **76**: 449–462.
- Tajima F (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.
- Tang K, Thornton KR, Stoneking M (2007). A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol* **5**: e171.
- Tishkoff SA, Verrelli BC (2003). Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu Rev Genomics Hum Genet* **4**: 293–340.
- Tregouet DA, Barbaux S, Escolano S, Tahri N, Golmard JL, Tiret L *et al.* (2002). Specific haplotypes of the P-selectin gene are associated with myocardial infarction. *Hum Mol Genet* **11**: 2015–2023.
- Urzainqui A, Martinez del Hoyo G, Lamana A, de la Fuente H, Barreiro O, Olazabal IM *et al.* (2007). Functional role of P-selectin glycoprotein ligand 1/P-selectin interaction in the generation of tolerogenic dendritic cells. *J Immunol* **179**: 7457–7465.
- Vestweber D, Blanks JE (1999). Mechanisms that regulate the function of the selectins and their ligands. *Physiol Rev* **79**: 181–213.
- Volcik KA, Ballantyne CM, Coresh J, Folsom AR, Boerwinkle E (2007). Specific P-selectin and P-selectin glycoprotein ligand-1 genotypes/haplotypes are associated with risk of incident CHD and ischemic stroke: the Atherosclerosis Risk in Communities (ARIC) study. *Atherosclerosis* **195**: e76–e82.
- Watanabe Y, Inoue T, Okada H, Kotaki S, Kanno Y, Kikuta T *et al.* (2006). Impact of selectin gene polymorphisms on rapid progression to end-stage renal disease in patients with IgA nephropathy. *Intern Med* **45**: 947–951.
- Watterson GA (1975). On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* **7**: 256–276.
- Wright S (1950). Genetical structure of populations. *Nature* **166**: 247–249.
- Wright SI, Charlesworth B (2004). The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. *Genetics* **168**: 1071–1076.
- Xue Y, Zhang X, Huang N, Daly A, Gillson CJ, Macarthur DG *et al.* (2009). Population differentiation as an indicator of recent positive selection in humans: an empirical evaluation. *Genetics* **183**: 1065–1077.
- Zee RY, Cook NR, Cheng S, Reynolds R, Erlich HA, Lindpaintner K *et al.* (2004). Polymorphism in the P-selectin and interleukin-4 genes as determinants of stroke: a population-based, prospective genetic analysis. *Hum Mol Genet* **13**: 389–396.

Supplementary Information accompanies the paper on Heredity website (<http://www.nature.com/hdy>)