BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Balancing selection is common in the extended MHC region but most alleles with opposite risk profile for autoimmune diseases are neutrally evolving

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

Abstract

Background: Several susceptibility genetic variants for autoimmune diseases have been identified. A subset of these polymorphisms displays an opposite risk profile in different autoimmune conditions. This observation open interesting questions on the evolutionary forces shaping the frequency of these alleles in human populations. We aimed at testing the hypothesis whereby balancing selection has shaped the frequency of opposite risk alleles.

Results: Since balancing selection signatures are expected to extend over short genomic portions, we focused our analyses on 11 regions carrying putative functional polymorphisms that may represent the disease variants (and the selection targets). No exceptional nucleotide diversity was observed for *ZSCAN23*, *HLA-DMB*, *VARS2*, *PTPN22*, *BAT3*, *C6orf47*, and *IL10*; summary statistics were consistent with evolutionary neutrality for these gene regions. Conversely, *CDSN/PSORS1C1*, *TRIM10/TRIM40*, *BTNL2*, and *TAP2* showed extremely high nucleotide diversity and most tests rejected neutrality, suggesting the action of balancing selection. For *TAP2* and *BTNL2* these signatures are not secondary to linkage disequilibrium with HLA class II genes. Nonetheless, with the exception of variants in *TRIM40* and *CDSN*, our data suggest that opposite risk SNPs are not selection targets but rather have accumulated as neutral variants.

Conclusion: Data herein indicate that balancing selection is common within the extended MHC region and involves several non-HLA loci. Yet, the evolutionary history of most SNPs with an opposite effect for autoimmune diseases is consistent with evolutionary neutrality. We suggest that variants with an opposite effect on autoimmune diseases should not be considered a distinct class of disease alleles from the evolutionary perspective and, in a few cases, the opposite effect on distinct diseases may derive from complex haplotype structures in regions with high genetic diversity.

Keywords: autoimmune disease, balancing selection, opposite risk profile, extended MHC region

Background

Genome-wide association studies (GWAS) have proved powerful in unravelling the genetic component of several common diseases and complex traits, although increasing evidences [1] suggest that rare variants, which are typically not analysed in GWASs, also contribute a considerable proportion of disease risk. Through GWASs and meta-analyses, a large number of single nucleotide polymorphisms (SNPs) have been associated with distinct autoimmune conditions including Crohn's disease (CD), ulcerative colitis (UC), multiple sclerosis (MS), type 1 diabetes (T1D), rheumatoid arthritis (RA), autoimmune thyroid disease (ATD), and ankylosing spondylitis (AS). A general concept emerging from these studies is that a portion of susceptibility alleles is shared among two or more diseases, suggesting that common molecular mechanisms and pathways are involved. This

* Correspondence: manuela.sironi@bp.lnf.it
[1]Scientific Institute IRCCS E. Medea, 23842 Bosisio Parini (LC), Italy
Full list of author information is available at the end of the article

may come as no surprise given the observation whereby clustering of distinct autoimmune diseases occurs within families (reviewed in [2]). However, recent evidences have also indicated that a subset of alleles displays an opposite risk profile in different autoimmune conditions with one allele predisposing to one disease while being protective for another [3,4]. The first described example concerns a nonsynonymous variant (R602W, rs2476601) in *PTPN22* (a tyrosine phosphatase expressed in T cells): the 602W allele protects from CD but predisposes to RA, SLE (systemic lupus erythematosus), T1D (reviewed in [5]), and vitiligo [6]. Similar observations have recently been extended to several SNPs [3,4], mostly located within the extended major histocompatibility complex (xMHC) region (Figure 1).

Besides opening interesting questions as to how immune balances are maintained and modulated, these data stimulate speculations on the evolutionary forces and selective pressures shaping the frequency of these alleles in human populations.

In general, variants associated with complex traits contribute little to the overall disease risk and are therefore thought to be subjected to mild purifying selection [7]. Yet, a portion of risk alleles may be regarded as deleterious, albeit mildly, from a medical standpoint but evolutionary neutral or even beneficial. Evolutionary studies of the MHC region have mainly focused on HLA class I and II genes, that are known to be characterized by extreme polymorphism levels maintained by natural selection (reviewed in [8]). Conversely, the evolutionary history of non-HLA genes has rarely been investigated.

Here we aimed at testing the hypothesis [4] whereby alleles with opposite risk profiles for autoimmune diseases have been maintained by balancing selection, possibly due to antagonistic pleiotropy, and to describe the evolutionary pattern of several non-HLA genes located in the human xMHC. Our data indicate that long-standing balancing selection has characterized the evolutionary history of non-HLA genes located in the xMHC but only a minority of alleles with opposite risk profile can be regarded as targets of natural selection in human populations.

Results

Identification of alleles with opposite risk profile and selection of candidate genomic regions

Two recent studies [3,4] have identified several alleles with opposite risk profiles for autoimmune diseases (Table 1). In order to search for additional instances, we analysed published GWAS http://www.genome.gov/26525384 and, in addition to the aforementioned variant in *PTPN22*, we identified rs744166 in *STAT3* and rs2201841 in *IL23R* (Table 1). The aim of our study is analysing the selective processes acting on variants with an opposite effect on two or more diseases and testing the hypothesis whereby balancing selection has shaped the frequency of a portion of these alleles. Balancing selection is a process that maintains genetic variability in human populations and its signatures, due to recombination and mutation, are expected to extend over relatively short genomic regions (reviewed in [9]). Since variants identified in association studies often represent genetic markers rather than causal polymorphisms, we analysed the SNPs reported in table 1 and their surrounding genomic regions for the presence of putative functional polymorphisms that may represent the causal variant and, possibly, the selection target. Details on functional annotation are reported in table 1 as well as in Figure 2 (and in additional file 1). Variants located in intergenic or relatively large intronic regions were not analysed due to the difficult inference of functional significance. Also, the polymorphism located in *MICA* was excluded as nucleotide diversity at this locus has been extensively investigated, although a clear picture of the selective or non-selective forces responsible for the presence of multiple alleles is still missing [10]. In the case of *TAP2*, the analysed region was extended so as to include a

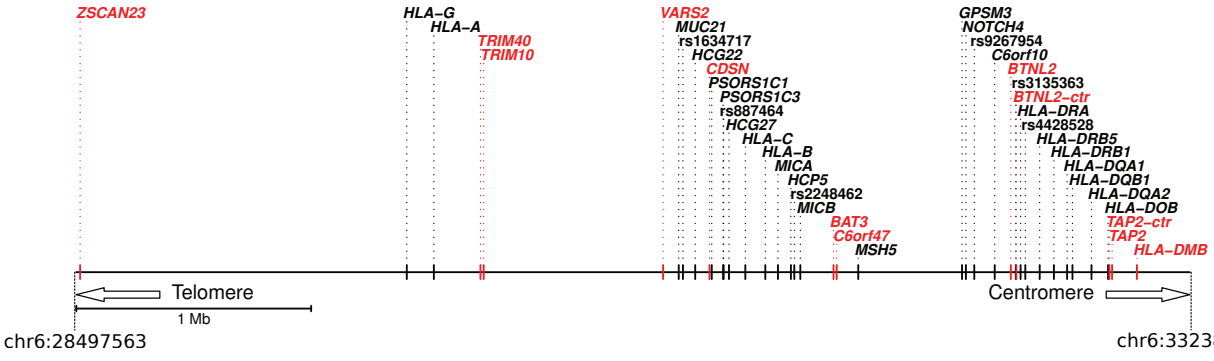

Figure 1 Schematic representation of a portion of the xMHC region. The location of all genes mentioned in the text is shown. Opposite risk SNPs are indicated with their ID if they are intergenic, otherwise the gene is indicated. The genes/gene regions we analysed are reported in red. Genomic coordinates for the region shown are chr6: 28497563-33238447 (NCBI Build 36.1).

Table 1 SNPs with opposite risk profiles in different autoimmune diseases

SNP	Gene	Chr[a]	A1[b]	Annotation	Diseases[c]	Ref
rs11752919	*ZSCAN23*	*6*	C	in small intron	RA-AS-T1D/ATD-MS	[3]
rs3130981	*CDSN/PSORS1C1*	*6*	A	Asp527Asn (in *CDSN*); less than 2.6 kb from rs1265048	RA-AS-T1D/ATD-MS	[3]
rs1265048		*6*	C	intergenic; 1.2kb upstream of *PSORS1C1*	RA-AS-T1D/ATD-MS	[3]
rs151719	*HLA-DMB*	*6*	G	in small intron	RA-AS-T1D/ATD-MS	[3]
rs10484565	*TAP2*	*6*	T	in 3' UTR	ATD-MS/RA-AS-T1D	[3]
0rs1264303	*VARS2*	*6*	G	in 5'UTR	RA-AS-T1D/ATD-MS	[3]
rs2076530	*BTNL2*	*6*	G	splice site altering/protein truncation	RA-AS-T1D/ATD-MS	[3]
rs3129953		*6*	T	intergenic; 0.7 kb downstream *BTNL2*	RA-AS-MS/T1D-ATD	[3]
rs757262	*TRIM40*	*6*	T	Thr183Met	RA-AS-T1D/ATD-MS	[3]
rs2517646	*TRIM10*	*6*	G	intronic; close to alternatively spliced exon	RA-AS-T1D-ATD/MS	[3]
rs2071286	*NOTCH4*	*6*	A	intronic	RA-AS-T1D/ATD-MS	[3]
rs2476601	*PTPN22*	*1*	T	Arg620TRP	CD/T1D-RA-SLE-GD	[6,60-64]
rs3024505	*IL10*	*1*	A	about 1kb downstream the transcription end site	T1D/CD-UC	[4]
rs917997	*IL18RAP*	*2*	T	about 1kb downstream the transcription end site	T1D/CD	[4]
rs9388489		*6*	G	intergenic	CD/T1D	[4]
rs4788084		*16*	T	intergenic	T1D/CD	[4]
rs1063635	*MICA*	*6*	A	Gln274Arg	T1D-ATD-MS/AS	[3]
rs1634717		*6*	A	intergenic	RA/T1D-ATD-MS	[3]
rs204991	*GPSM3*	*6*	C	intronic	RA-AS-MS/T1D-	[3]

Table 1 SNPs with opposite risk profiles in different autoimmune diseases *(Continued)*

rs2242655	*C6orf47*	*6*	C	Lys92Asn	ATD ATD-MS/RA-AS-T1D	[3]
rs2299851	*MSH5*	*6*	T	Intronic	ATD-MS/RA-AS-T1D	[3]
rs2248462		*6*	A	intergenic	AS/RA-T1D-ATD-MS	[3]
rs2844463	*BAT3*	*6*	T	intronic; close to alternatively spliced exon	RA-AS-ATD/T1D-MS	[3]
rs3135363		*6*	C	intergenic	RA-AS-MS/T1D-ATD	[3]
rs4428528		*6*	C	intergenic	RA-AS-MS/T1D-ATD	[3]
rs887464		*6*	A	intergenic (upstream PSORS1C3)	MS/RA-AS-T1D-ATD	[3]
rs9267954		*6*	T	intergenic	ATD-MS/RA-AS-T1D	[3]
rs2201841	*IL23R*	*1*	G	intronic	UC/psoriasis	[65,66]
rs744166	*STAT3*	*17*	G	intronic	CD/MS	[67,68]

[a] chromosome.

[b] minor allele in CEU.

[c] diseases are shown in order depending on the predisposing/protective effect of the minor allele.

region that undergoes haplotype-specific alternative splicing [11]. SNPs located in close physical proximity were analysed as a single region: this was the case for rs3130981 and rs1265048 (in *CDSN/PSORS1C1)*, rs2076530 and rs3129953 *(*in *BTNL2)*, and rs757262 and rs2517646 *(*in *TRIM10/TRIM40).*

As for rs917997, located downstream *IL18RAP*, the SNP was not considered as the gene has previously been shown to be subjected to balancing selection [12]. Finally, rs3024505 (downstream *IL10*) lies in a region resequenced by the SeattleSNPs program and data were therefore retrieved from their website. A signature of balancing selection at the promoter region of IL10 had previously been described [13]. Resequencing data for *STAT3* are also publicly available (from the SeattleSNPs program website) but the opposite-risk SNP was not analysed as it is located within a resequencing gap in the long intron 1.

Nucleotide diversity and neutrality tests

As reported in table 2, at least 2 kb encompassing each selected SNP(s) (Figure 2) were resequenced in 20 HapMap subjects with European ancestry (CEU), as most GWASs for autoimmune diseases have been performed in European cohorts. The number of segregating sites identified in each region is reported in table 2.

Common population genetic tests based on the site frequency spectrum (SFS) include Tajima's D (D_T) [14] and Fu and Li's D* and F* [15]. D_T tests the departure from neutrality by comparing two nucleotide diversity indexes: θ_W [16], an estimate of the expected per site heterozigosity, and π [17], the average number of pairwise sequence nucleotide differences. Positive values of D_T indicate an excess of intermediate frequency variants and are a signature of balancing selection. 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 [15]. As an empirical comparison, θ_W, π, as well as D_T, F* and D* were calculated for 5 kb windows (thereafter referred to as reference windows) deriving from 238 genes resequenced by the NIEHS program in CEU. Additionally, the statistical significance of neutrality tests was evaluated by performing coalescent simulations with a population genetic model that incorporates demographic scenarios [18].

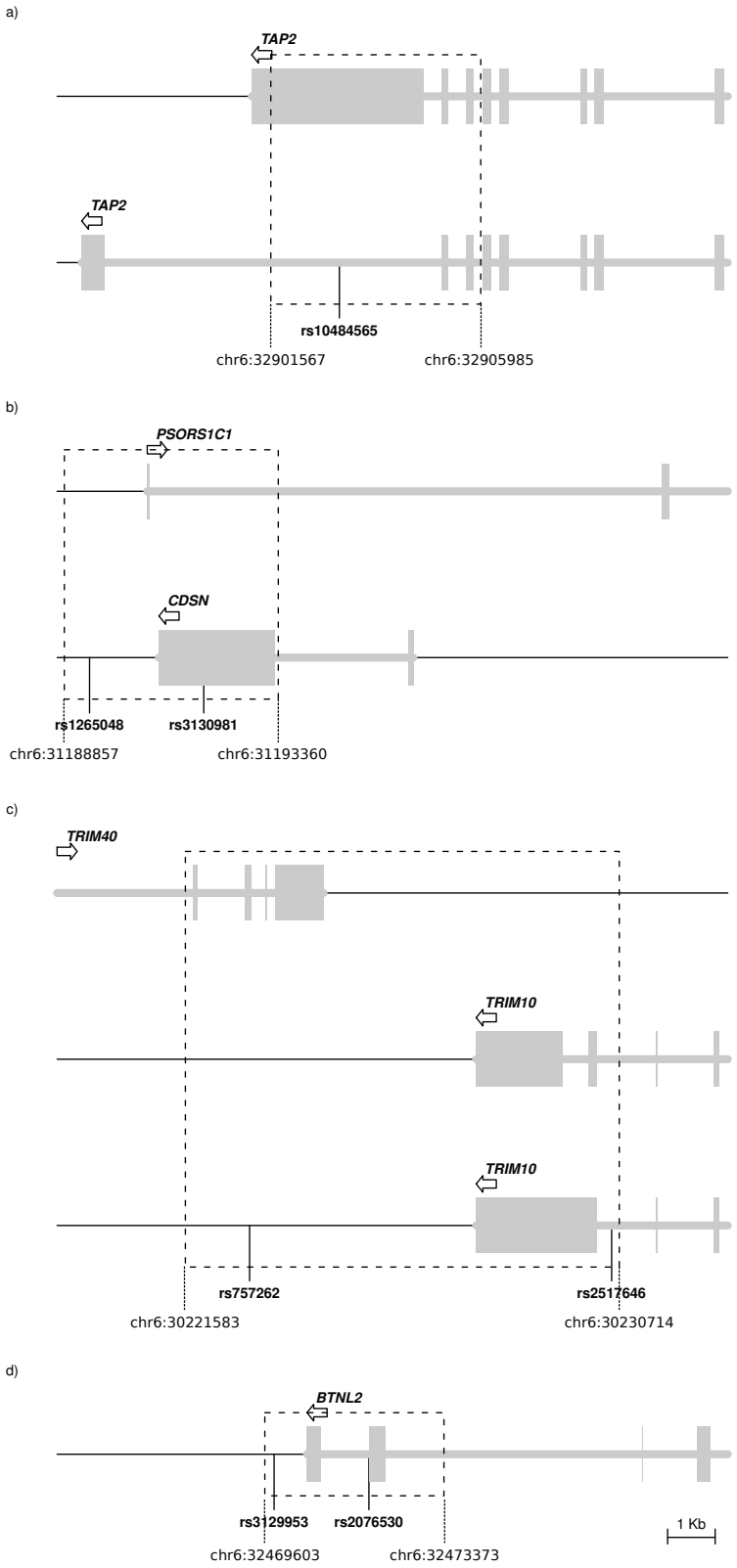

Figure 2 Schematic representation of the gene regions we resequenced in *TAP2*, *TRIM10/TRIM40*, *CDSN/PSORS1C1* and *BTNL2*. Transcribed regions are shown in grey; different transcripts either from the same or from different genes are shown separately. The direction of transcription is indicated by the arrows. The location of SNPs with opposite risk profile is reported.

Table 2 Nucleotide diversity and neutrality tests for the analysed gene regions

Gene	L[a]	P[b]	S[c]	θ_W[d]		π[e]		Tajima's D			Fu & Li's D*			Fu & Li's F*		
				value	rank[f]	value	rank[f]	value	rank[f]	p[g]	value	rank[f]	p[g]	value	rank[f]	p[g]
ZSCAN23	3.3	CEU	7	4.96	0.44	6.52	0.66	0.87	0.78	0.20	0.51	0.72	0.34	0.73	0.74	0.25
VARS2	3.2	CEU	10	7.30	0.69	10.78	0.86	1.43	0.89	0.09	0.24	0.65	0.40	0.73	0.74	0.25
HLA-DMB	3.1	CEU	14	10.60	0.91	12.12	0.89	0.45	0.66	0.30	-0.28	0.47	0.60	-0.050	0.54	0.51
PTPN22	3.5	CEU	7	4.70	0.44	4.57	0.41	-0.077	0.51	0.49	-0.98	0.23	0.80	-0.81	0.31	0.75
BAT3	2.0	CEU	3	3.47	0.21	3.14	0.27	-0.20	0.47	0.57	-0.35	0.46	0.64	-0.36	0.44	0.62
C6orf47	2.0	CEU	3	3.59	0.21	5.21	0.48	0.98	0.81	0.18	0.92	0.85	0.19	1.09	0.84	0.13
IL10	2.3	CEU	3	2.97	0.15	5.95	0.55	2.12	0.97	0.01	0.90	0.84	0.16	1.47	0.94	0.04
CDSN/ PSORSC1	4.5	CEU	59	30.80	>0.99	49.42	>0.99	2.17	0.97	0.01	1.40	0.95	0.03	1.98	0.97	0.01
		YRI	63	32.88	>0.99	46.20	>0.99	1.46	0.99	0.01	1.44	>0.99	<0.01	1.72	>0.99	<0.01
		EAS	64	33.41	>0.99	47.50	>0.99	1.52	0.91	0.08	1.45	0.98	001	1.75	0.98	0.02
TRIM40/ TRIM10	9.1	CEU	68	17.51	0.98	26.13	0.99	1.78	0.96	0.03	1.61	0.98	<0.01	1.98	0.97	<0.01
		YRI	64	16.48	0.96	15.74	0.94	-0.16	0.64	0.22	1.07	0.97	<0.01	0.76	0.93	0.02
		EAS	79	20.34	0.99	19.33	0.98	-0.18	0.46	0.31	-0.20	0.46	0.68	-0.23	0.45	0.71
BTNL2	3.8	CEU	53	33.04	>0.99	39.75	0.99	-0.33	0.42	0.36	1.32	0.93	0.03	0.88	0.78	0.15
		YRI	63	39.28	>0.99	30.01	>0.99	0.043	0.71	0.18	0.67	0.89	0.06	0.53	0.86	0.07
		EAS	94	58.60	>0.99	47.38	>0.99	-0.70	0.31	0.21	0.95	0.88	0.13	0.43	0.69	0.35
TAP2	4.4	CEU	33	17.56	0.98	29.07	0.99	2.27	0.98	0.01	1.59	0.98	0.01	2.16	0.98	<0.01
		YRI	54	28.73	>0.99	34.55	>0.99	0.73	0.94	0.05	1.63	>0.99	<0.01	1.56	0.99	<0.01
		EAS	41	21.81	0.99	32.88	>0.99	1.79	0.97	0.04	1.30	0.96	0.04	1.74	0.98	0.02
TAP2-ctr	2.6	CEU	12	10.71	0.91	12.18	0.89	0.42	0.65	0.32	0.46	0.70	0.29	0.52	0.69	0.18
BTNL2-ctr	2.0	CEU	15	17.47	0.98	16.73	0.96	-0.14	0.50	0.54	1.56	0.97	0.01	1.18	0.90	0.09

[a] length of analyzed sequenced region (kb).

[b] sampled population (for each population 40 chromosomes were resequenced).

[c] number of segregating sites.

[d] Watterson's theta estimation per site ($\times 10^{-4}$).

[e] nucleotide diversity per site ($\times 10^{-4}$).

[f] percentile rank relative to a distribution of 238 5kb segments from NIEHS genes.

[g] *p* value applying demographic coalescent simulations.

As shown in table 2, no exceptional nucleotide diversity was observed for *ZSCAN23*, *HLA-DMB*, *VARS2*, *PTPN22*, *BAT3*, *C6orf47*, and *IL10*. In line with this observation, summary statistics were consistent with evolutionary neutrality for these gene regions (Table 2). Conversely, the regions we analysed in *CDSN/ PSORS1C1*, *TRIM10/TRIM40*, *BTNL2* and *TAP2* showed extremely high nucleotide diversity, with both θ_W and π ranking above the 95th percentile in the distribution of 5kb reference windows.

For *TAP2, CDSN/PSORS1C1*, and *TRIM10/TRIM40* all tests rejected the null hypothesis of selective neutrality in CEU and ranks of D_T, F* and D* were higher than the 95th percentile. In the case of *BTNL2*, D*, but not D_T and F*, was close to statistical significance in the empirical comparison and rejected neutrality when coalescent simulations were performed.

TAP2 and *BTNL2* are located within the classical class II MHC sub-region (Figure 1) and flank a class II HLA gene cluster containing highly polymorphic genes subjected to balancing selection in humans and other primates [19,20]. We therefore wished to verify whether the selection signatures we identified at both genes might be secondary to linkage disequilibrium with HLA class II genes. Thus, intergenic regions flanking the class II HLA gene cluster were also resequenced: as shown in Figure 1, the TAP2-ctr region is telomeric to *TAP2*, while BTNL2-ctr is centromeric to *BTNL2*. These two control regions have very similar divergence to the *TAP2* and *BTNL2* regions we analysed (not shown). As reported in table 2, the TAP2-ctr region displayed no exceptional variability and all statistics were consistent with selective neutrality. As for BTNL2-ctr, a high nucleotide diversity was observed but both θ_W and π were about half the value observed at the *BTNL2* genic region; only F* displayed a significantly high value. Therefore, high nucleotide diversity at the *TAP2* locus is not secondary to LD with HLA class II genes. In the case of *BTNL2*, the lower diversity observed at the control compared to the genic region also suggest that the

gene is an independent target of balancing selection (see below).

In order to obtain a more comprehensive description of nucleotide diversity, the four regions covering *CDSN/PSORS1C1*, *TRIM10/TRIM40*, *BTNL2*, and *TAP2* were resequenced in two additional HapMap populations, namely Yoruba (YRI) and East Asians (EAS). For all regions, nucleotide diversity resulted extremely high in YRI and EAS, as well (Table 2). Neutrality tests and empirical comparison with resequenced regions rejected neutrality for these populations at the *CDSN/PSORS1C1* and *TAP2* regions. Similar results were obtained for *TRIM10/TRIM40* in YRI but not in EAS. This latter finding is due to the presence of several singletons that affect SFS-based tests in this population (see below). As for *BTNL2*, the values of SFS-based statistics were not exceptionally high in YRI and EAS.

A hallmark of balancing selection is an excess of polymorphism compared to neutral expectations. Indeed, our data (Table 2) indicate that nucleotide diversity indexes are extremely high for *CDSN/PSORS1C1*, *TRIM10/TRIM40*, *BTNL2* and *TAP2*. 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 [21]. The multi-locus HKA test was developed to verify this expectation [22]. We applied a multi-locus MLHKA (maximum-likelihood HKA) test by comparing polymorphisms and divergence levels at the *CDSN/PSORS1C1*, *TRIM10/TRIM40, BTNL2* and *TAP2* genomic regions with 16 NIEHS genes resequenced in YRI, CEU and EAS. The results are shown in table 3 and indicate that a significant excess of nucleotide diversity versus divergence is detectable in all populations for all loci. For *TAP2* and *TRIM10/TRIM40* the chimpanzee was used for inter-species divergence. Yet, divergence with chimpanzee is unusually low (0.5%) for the *CDSN/PSORS1C1* region and the reference sequence for orangutan is not available; as for macaque, divergence for *CDSN/PSORS1C1* (4.8%) is also lower than genome average but not markedly so. Therefore, for *CDSN/PSORS1C1* the MLHKA test was performed using macaque divergence data. Finally, in the case of *BTNL2*, no reference sequence for chimpanzee is available for the gene region we analysed. Since a partial sequence is available for orangutan, we sequenced the genomic DNA of one *Pongo pygmaeus* (see methods) and used the sequence we obtained for calculation of divergence; therefore the MLHKA test for *BTNL2* was performed with human/orangutan divergence data.

We next took advantage of the availability of data from the 1000 Genomes Pilot project [23] to validate our results in a larger poulation sample. The low-coverage 1000 Genomes approach, which generated whole genome sequencing data of 179 individuals with different ancestry (YRI, CEU and EAS), is estimated to have relatively low power to detect singleton SNPs or rare variants [23]. Thus, an empirical comaprison is needed to evaluate whether selected gene regions display diversity indexes or SFS-based staistics that are exceptionally high (i.e. that reject neutral expectations). To this aim we randomly selected 2,000 human genes and, from each of them, one 5 kb window was extracted. We next calculated diversity indexes and SFS-based statistics, as in table 2. All results were confirmed using this approach (Additional file 2). Finally, we used the 1000 Genomes data to perform a sliding-window analysis of an extended genomic regions covering *BTNL2*, the MHC class II gene cluster, and *TAP2* plus flanking regions. As shown in Figure 3, two peaks of diversity were observed at the *BTNL2* and *TAP2* regions we resequenced. These are separated from the MHC class II region (showing extreme diversity, as expected) by segments showing lower values of both θ_W and π, suggesting that they represent independent selection targets.

Table 3 MLHKA tests

Region	Pop.[a]	MLHKA	
		k[b]	*p*
TAP2	CEU	3.69	1.1×10^{-3}
	YRI	4.94	6.4×10^{-5}
	EAS	5.11	3.2×10^{-5}
CDSN/PSORS1C1	CEU	4.79	1.6×10^{-4}
	YRI	3.88	1.6×10^{-5}
	EAS	5.66	1.6×10^{-7}
TRIM10/TRIM40	CEU	3.82	5.5×10^{-4}
	YRI	2.23	2.1×10^{-2}
	EAS	4.21	3.9×10^{-5}
BTNL2	CEU	5.57	6.4×10^{-7}
	YRI	5.44	1.7×10^{-6}
	EAS	10.80	5.6×10^{-13}

[a] population.

[b]selection parameter (k>1 indicates an excess of polymorphism relative to divergence).

Haplotype analysis and TMRCA estimates

Further insight into the evolutionary history of a gene region can be gained by inferring haplotype genealogies. This has both a descriptive purpose (i.e. showing the relationship among alleles and their distribution in human populations) and can be used to test for selection. In particular, balancing selection is expected to result in two or more major haplotype clades with a deep coalescence time (TMRCA, time to the most

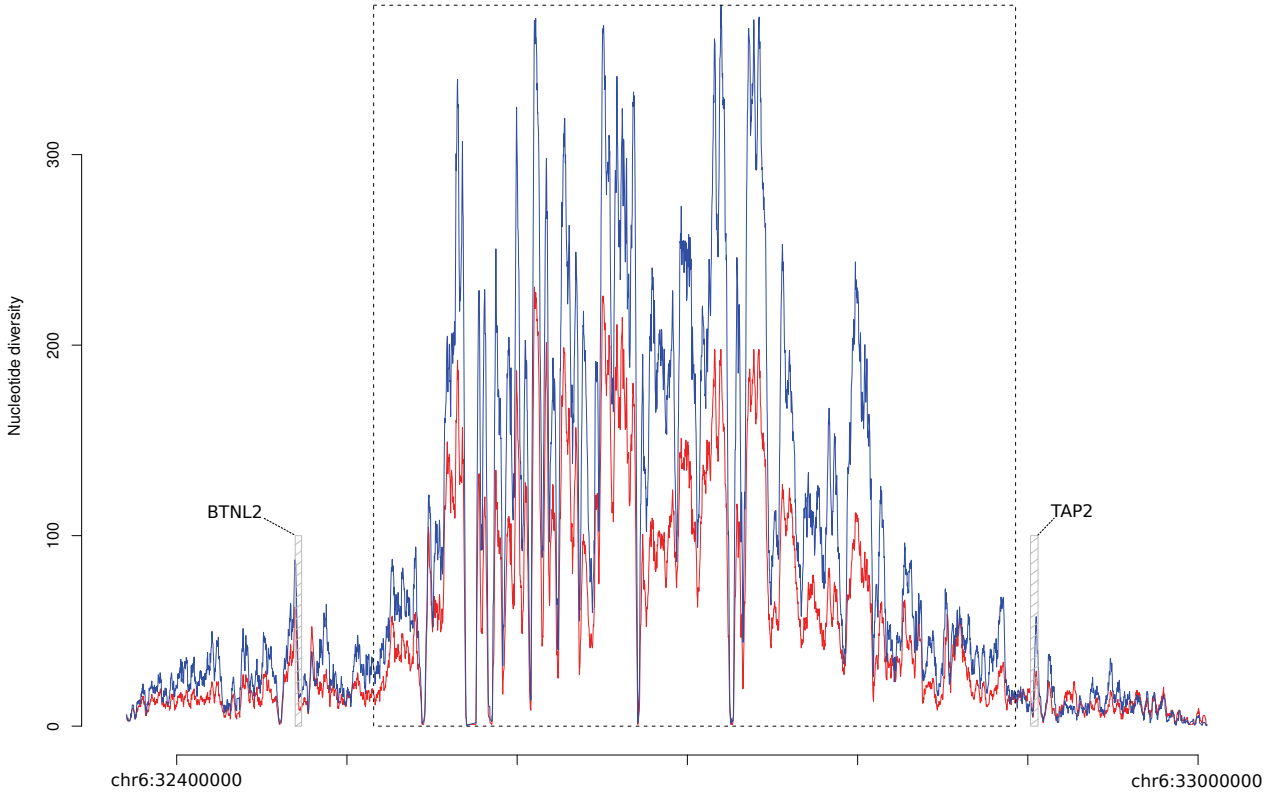

Figure 3 Sliding window analysis of the genomic region encompassing *BTNL2*, *HLA* class II genes and *TAP2*. Data were obtained from the low-coverage 1000 Genomes Project data and refer to CEU. θ_W (red) and π (blue) were calculated in 2 kb windows moving with a step of 150 bp. The two region we resequanced in TAP2 and BTNL2 are shown by the shaded boxes. The hatched region indicates the genomic portion occupied by HLA class II genes.

common recent ancestor). In fact, while neutrally evolving autosomal loci have TMRCAs ranging from 0.8 to 1.5 million years (MY) [24], gene regions under balancing selection may show coalescence times dating back more than 4 MY [25-27]. Here we constructed haplotype genealogies using two approaches: a neighbour-joining network and a maximum-likelihood coalescent method implemented in GENETREE. This latter assumes an infinite-site model without recombination, requiring the removal of variants and haplotypes that violate these assumptions. In order to obtain more reliable trees, we selected sub-regions based on LD for those genomic regions showing high recombination rates. Thus, for *TAP2* we used data from a 1.9 kb region with relatively high LD (Additional files 3 and 4). This region does not encompass the opposite risk SNP but includes a set of markers (rs241448, rs241447 and 241441, variants 19, 18 and 8, respectively in the network, Figure 4A) previously known to identify the two haplotypes that generate alternatively spliced isoforms [11]. As it is evident from both the network and GENETREE analysis (Figure 4), the *TAP2* genealogy is split into two major haplotype clades with an estimated time to the most recent common ancestor (TMRCA) of ~5.36 MY. The two clades differ at several variants including those affecting *TAP2* splicing.

Similarly, due to extensive recombination, a sub-region with stronger LD (Additional files 3 and 4) was analysed in the case of *TRIM10/TRIM40*. The rs2517646 variant (table 1) lies outside this region, whereas the second variant with an opposite risk profile (rs757262) is located on the major branch leading to clade B (Figure 5A). As evident in both the network and GENETREE analyses a single, highly divergent haplotype is observed in EAS (Figure 5); although several positions along the branch are recurrent and possibly originate from recombination between the two major clades or gene conversion, 13 SNPs are specific to this haplotype and represent singletons in EAS, therefore affecting SFS-based statistics (Table 2).

In the case of *CDSN/PSORS1C1*, again we selected a 2 kb region of relative LD (Additional files 3 and 4) which includes one of the two variants with opposite risk-profile (variants 13 in the network, Figure 6A). The two major clades of the genealogy have a TMRCA of 4.18 MY and clade B is split into two main haplogroups that

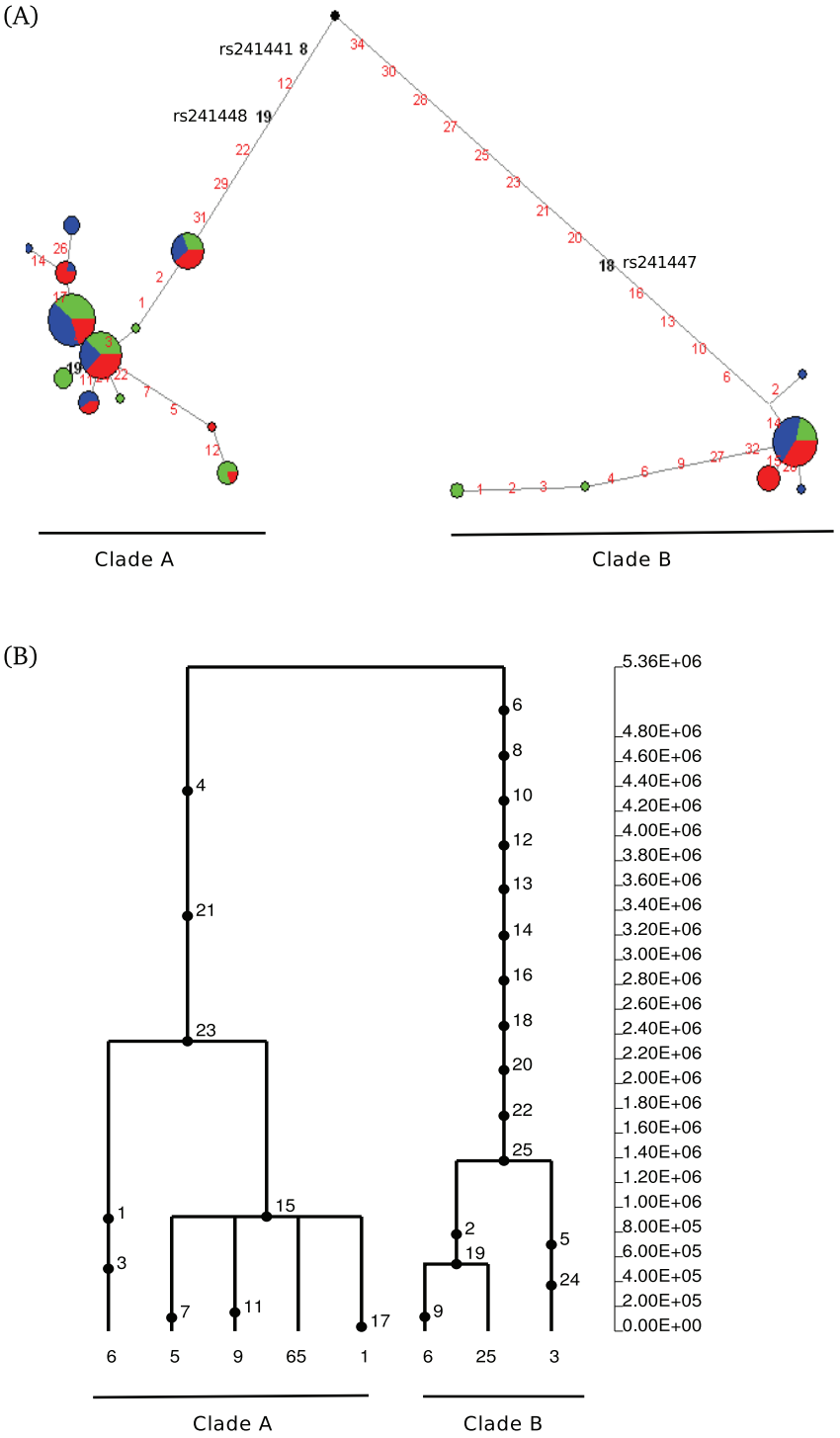

Figure 4 Haplotype genealogy of the analysed *TAP2* region. (A) Network analysis. Each node represents a different haplotype, with the size of the circle proportional to frequency. Nucleotide differences between haplotypes are indicated on the branches of the network. Circles are colour-coded according to population (green: YRI, blue: CEU, red: EAS). The most recent common ancestor is also shown (black circle). The relative position of mutations along a branch is arbitrary. (B) GENETREE analysis. 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 (A).

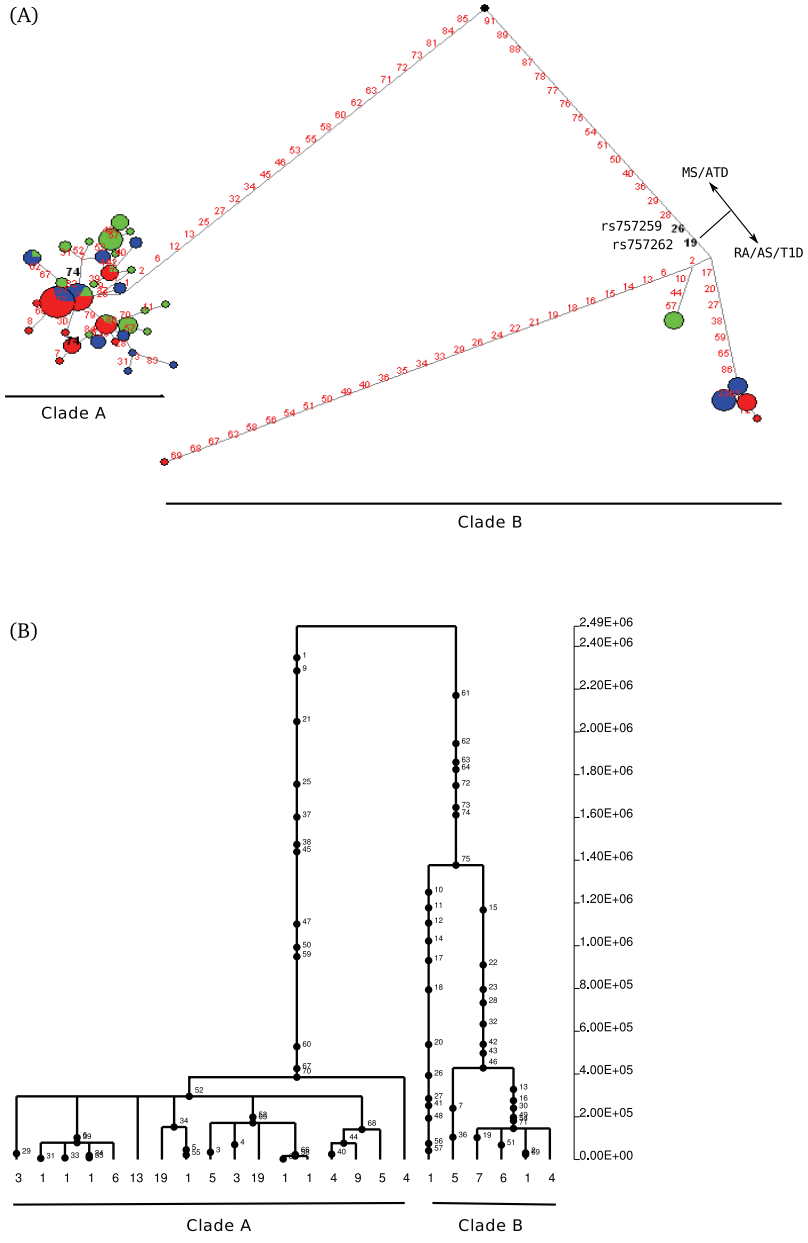

Figure 5 Haplotype genealogy of the analysed *TRIM10/TRIM40* region. Legend is as for figure 3. In addition an indication is given as to which haplotypes predispose to RA/AS/T1D or MS/ATD.

coalesce at 2.2 MY (Figure 6B), possibly suggesting a multiallelic balancing selection regime. It is worth noting that, for the same reasons reported above, the coalescence time was calibrated on the basis of a mutation rate calculated from human/macaque divergence. The opposite-risk SNP defines a subset of haplotypes in clade B (Figure 6A).

Finally, in the case of *BTNL2*, the gene portion we analysed is in relative tight LD and the network and GENETREE analyses were performed over the entire region. Three major haplotype clades are evident with the most distantly related haplotype cluster being present in EAS only (Figure 7).

Conversely, chromosomes from all populations contribute to the two remaining clades, although with extremely different frequency. The TMRCA for the whole genealogy amounts to ~ 6 MY, while the two more closely related clades coalesce at ~ 2.8 MY (Figure 7B). The branches leading to the two lower-frequency clades share some variants and their relatively close physical proximity suggests that these apparent homoplasies are due to a recombination or gene conversion event.

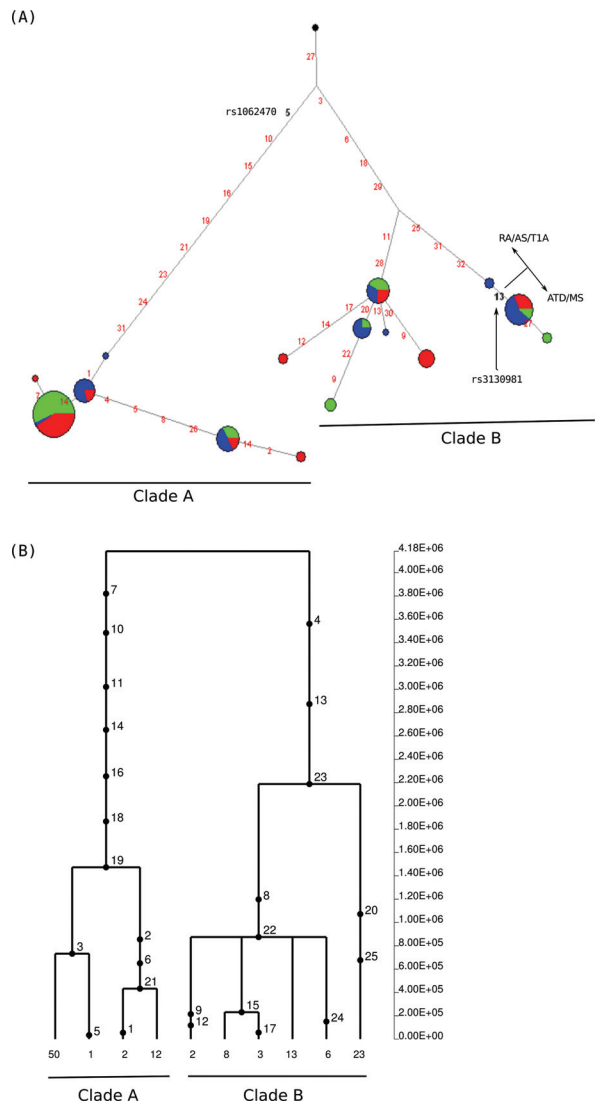

Figure 6 Haplotype genealogy of the analysed *CDSN/PSORS1C1* region. Legend is as for figure 4.

The two variants with opposite effect are not located on the major branches of the genealogy but define haplotye subsets within the major clade (Figure 7A).

This haplotype genealogy explains the results we obtained for SFS-based statistics in this region; the presence of two (CEU and YRI) or three (EAS) deeply separated haplotype clades introduces a high number of polymorphic sites (resulting in high nucleotide diversity indexes). Yet, clade 2 haplotypes are present at low frequency in all populations and the same applies to clade 1 in EAS. Therefore, the frequency spectrum is not markedly skewed towards intermediate frequency alleles and, consequently, most SFS-based statistics fail to reject the null hypothesis of neutrality.

Nonetheless, the haplotype genealogy and TMRCAs we obtained are not consistent with neutral evolution in an unstructured population but rather suggest the action of balancing selection and/or a possible contribution of archaic population structure in Asia. An alternative possibility is that, as discussed below, this pattern is due to LD with HLA class II genes.

Discussion

Autoimmune diseases are common in industrialized societies, collectively reaching a prevalence of 5% in populations with European ancestry [28]. Epidemiological studies have indicated that the incidence of these disorders has been steadily increasing during the last decades in the industrialized world. Therefore, much scientific debate has addressed the role of human evolutionary history and adaptation in shaping the genetic predisposition to the development of autoimmunity [29]. For any single autoimmune condition, more than 50% of the disease risk is heritable [30] and GWASs have unveiled the role of several common risk variants, possibly reflecting an allelic architecture for autoimmune disease that matches a "common variant/common disease" model more closely than observed for other traits [30]. From the evolutionary perspective, this raises interesting questions on the forces responsible for the maintenance of disease alleles in populations (reviewed in [29]). Since the evolutionary history of autoimmune-related alleles is only beginning to be investigated, our knowledge is still relatively limited in this field. Specifically, a subset of risk alleles for CD and UC has previously been shown to have evolved in response to pathogen-driven selective pressures, the underlying scenario for some of them being balancing selection [12]. More recently, several disease alleles for autoimmune conditions were reported to display signatures of directional selection in favour of the risk alleles [31]. Finally, two common variants for T1D, an early-onset, potentially lethal disease, have been described as neutrally evolving [32]. Albeit limited to a small number of variants, these data do not support the notion whereby autoimmune phenomena have acted as a selective pressure strong enough to affect the frequency distribution of risk alleles, although some authors have speculated that balancing selection at innate immunity genes might stem from a tuning of response to pathogens and to self [33]. The identification of a set of variants with an opposite risk profile for different autoimmune conditions prompted the speculation that these variants might be targets of balancing selection possibly deriving from antagonistic pleiotropy. Therefore the question is: if neither allele can be considered medically favourable, are they both evolutionary beneficial under specific circumstances? Our data indicate that 4 out of ten gene regions we analysed have been subjected to balancing selection. Additionally, *IL18RAP* has previously been

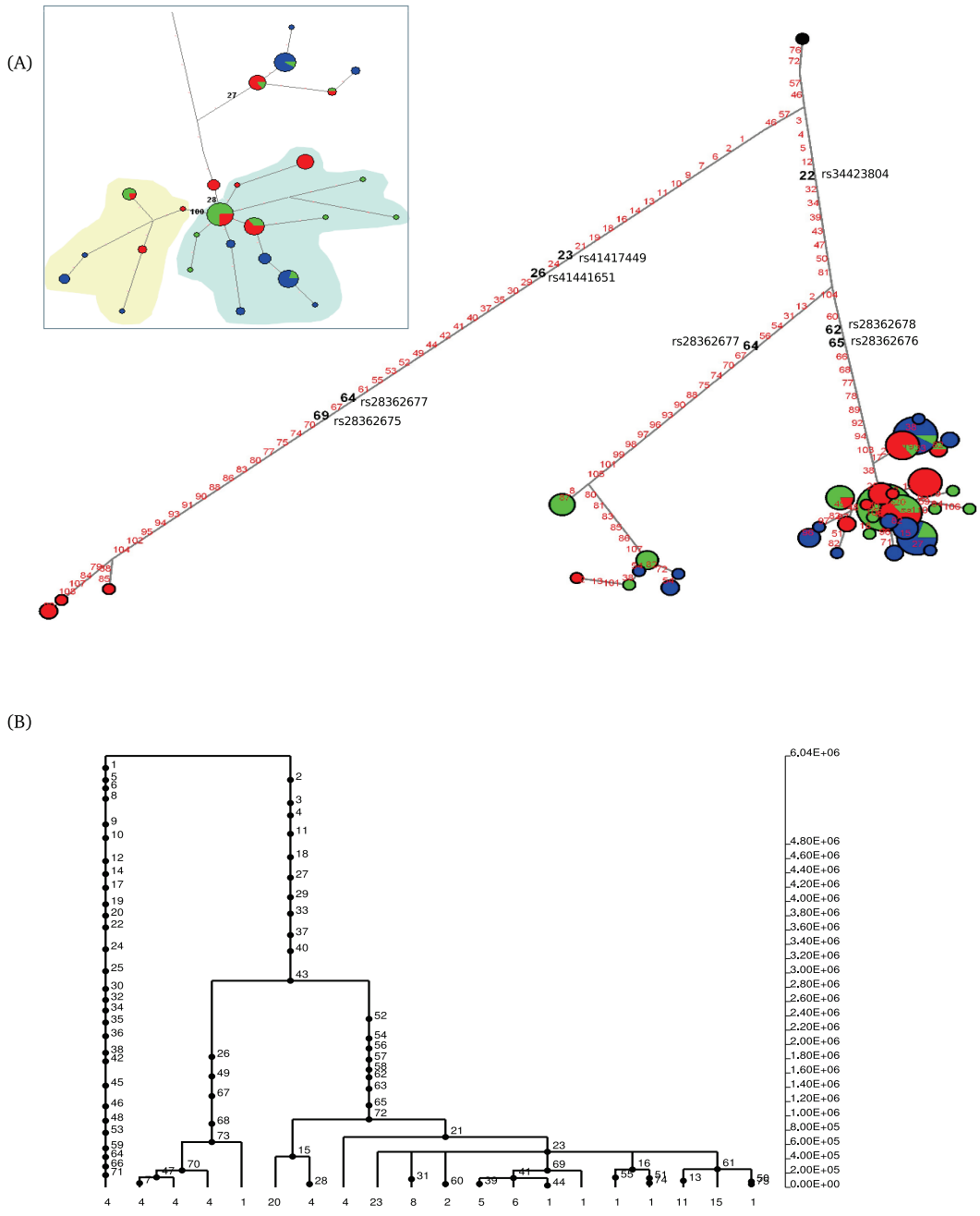

Figure 7 Haplotype genealogy of the analysed *BTNL2* region. Legend is as for figure 3. In (A) positions in bold refer to the following SNPs: 22, rs34423804 (Val189Asp); 23, rs41417449 (Met201Val); 26, rs41441651 (Asp242Asn); 62, rs28362678 (Pro285Leu); 64, rs28362677 (Met286Ile); 65, rs28362676 (Pro299Gln); 69, rs28362675 (Gly360Cys). An enlargement of the major haplotype cluster is also shown (left-upper panel): susceptibility haplotypes for ATD and MS are circled in yellow and cyan, respectively. SNPs numbering is as follows: 27, rs9268480; 28, rs2076530; 100, 3129953.

indicated as a possible selection target [12]. Similarly, a signature of balancing selection has previously been described at the promoter region of *IL10* [13], suggesting that the selected variant might be different from the opposite risk allele which is located downstream the transcription end site. The evidences we report herein for the *TAP2*, *TRIM10/TRIM40*, and *CDSN/PSORS1C1* gene regions are all consistent with the hypothesis whereby genetic variability is maintained at these loci by long-standing balancing selection. In the case of *BTNL2* and *TAP2*, we addressed the possibility that LD with HLA class II genes influences the results we obtained,

as genetic hitch-hiking can potentially affect neutral diversity over long genomic regions. Yet, the sliding window analysis of nucleotide diversity across the region encompassing these two genes and HLA class II loci indicated that peaks at *BTNL2* and *TAP2* are flanked by regions with lower diversity, suggesting that these two genes represent independent (from class II loci) selection targets.

Conversely, we identified no selection signature for the remaining 6 genes, namely *ZSCAN23*, *PTPN22*, *HLA-DMB*, *VARS2*, *C6orf47*, and *BAT3*. In all these cases nucleotide diversity was within average values and no test significantly deviated from the null hypothesis of neutral evolution. One possibility is that opposite risk SNPs in these regions do not represent causal variants but genetic markers and, therefore, that natural selection might be acting in a region different from the one we resequenced. Yet, this is unlikely to be the case for *PTPN22*, as the opposite risk variant has been shown to be functional. The derived 602W allele, which segregates at low frequency in Caucasian populations and is extremely rare outside Europe, confers to the phosphatase a stronger ability to inhibit the T cell receptor signalling pathway [34]; this allele has been associated with T1D, RA and other autoimmune manifestations [35]. Conversely, the ancestral allele has been associated with a higher risk of developing CD. Our analysis of the region carrying the R602W variant did not unveil any molecular signature of natural selection, as all tests were consistent with neutrality. Nonetheless, the power of most tests is strongly influenced by the timing and strength of the selective pressure; it has been suggested that the 602W allele has risen in frequency in some European populations as a result of natural selection, as its frequency seems to increase with latitude [35]. An extremely recent selective event in these populations would not be detected using our approach.

The four targets of balancing selection we identified in this study are all located within the xMHC and have different molecular functions. *TAP2* is a central component of the antigen processing pathway; the protein products of *TAP1* and *TAP2* interact to form a transporter complex (TAP) that translocates antigenic peptides to the endoplasmic reticulum (ER) where loading onto MHC class I molecules occurs. Therefore, inhibition of TAP has been exploited by different viruses as an immune evasion strategy (reviewed in [36]). Specifically, proteins encoded by herpesviruses (HSV), human cytomegalovirus (HCMV) and Epstein-Barr virus can block TAP function, limit the supply of peptides to the endoplasmic reticulum and therefore inhibit MHC class I maturation. Interestingly, the ability of HSV- and HCMV-encoded proteins to block TAP function is species-specific, suggesting that viral products have co-evolved with the TAP molecules of their host species. This observation also suggests that *TAP1* and *TAP2* may be subjected to a selective pressure exerted by viruses to avoid binding of inhibitory proteins. The *TAP2* gene portion we analysed herein is subjected to a haplotype-specific alternative splicing event that results in two molecular forms differing at the C-terminus; our data show that the two haplotypes associated with alternative splicing are maintained by balancing selection. Whether the two protein products are differentially sensitive to viral inhibitors is presently unknown, but a previous report has indicated that they display marked differences in the translocation efficiency of specific peptides. This difference may have an effect on the susceptibility to specific viral infections and the Ala665Thr variant (rs241447), which is associated with alternative splicing, and located on the major branch leading to clade b (Figure 4A) has been associated to altered susceptibility to HIV-1 infection [37].

Two genes encoding ER aminopeptidases that trim peptides translocated by TAP have recently been shown to be subjected to long-standing balancing selection with polymorphic variants conferring resistance against HIV-1 [26]. It is therefore tempting to speculate that maintenance of genetic diversity at genes in the antigen processing pathway derives from the differential activity of diverse alleles for specific peptides, eventually leading to distinctive repertoires of antigens presented to $CD8^+$ T cells and, possibly, altered susceptibility to specific infections. Nonetheless, it is worth noting that the allele with opposite risk profile (rs10484565) has a low MAF (frequency of minor allele < 0.10) in all populations and our data suggest that it is not (and is not in linkage with) the balancing selection target.

TRIM10 and *TRIM40* code for two members of a large family of tripartite motif-containing (TRIM) proteins. While several TRIM proteins have been shown to act as antiviral factors, the role of TRIM10 and TRIM40 is virtually unknown, although TRIM10 may have a role in erythropoiesis [38]. A recent GWAS for host genetic factors involved in HIV control identified a SNP (rs9468692) located in 3'UTR of *TRIM10*, within the region we analysed. Our resequencing data indicate that this SNP is triallelic with T/G alleles segregating in CEU and YRI, and A/G in EAS. This is clearly shown in the network analysis: variant 74 defines two different haplotype clusters one containing African and European chromosomes and the other Asian haplotypes only. The variant involves a CpG dinucleotide suggesting that the triallelic status is determined by a two-hit mutation process on different haplotypic backgrounds and involving deamination of 5-methyl cytosine in the case of the A/G alleles. While the description of a triallelic variant might have a relevance for future association studies in populations with different ancestry, its location in the network suggests that the site is neutrally evolving. Conversely, the Thr183Met variant

(rs757262, variant 19 in Figure 5A) with opposite risk profile (Table 1) is in strong LD with another nonsynonymous variant (Glu215Lys, rs757259, variant 26 in Figure 5A) and both SNPs separate the two major haplotype clades, suggesting that they may represent the selection target(s). The limited knowledge on the biological function of *TRIM10* and *TRIM40* does not allow extensive speculation on the selective pressure responsible for maintaining nucleotide diversity at these genes and further functional studies will be required to determine whether it is virus-driven or not.

The third gene region which we found to be subjected to balancing selection covers part of *CDSN* and *PSORS1C1*. A previous work using a different dataset has also suggested that CDSN might be a balancing selection target [39]. *CDSN* and *PSORS1C1* the two genes are transcribed in the opposite direction and the whole coding region of corneodesmosin (*CDSN*) is comprised within the first intron of *PSORS1C1* (also known as *SEEK1*). Both transcripts are widely expressed http://biogps.gnf.org and have been associated with psoriasis in several studies. Specifically, *CDSN* is up-regulated in psoriatic lesisons [40] and one psoriasis-associated allele (rs1062470, variant 5 in Figure 6A) affects the binding of an unknown cellular factor, resulting in increased transcript stability [40]. This SNP is located on the network branch separating the two major haplotype clades, suggesting that it may represent the selected variant. Similarly, one of the two SNPs (rs3130981, Asp527Asn) with an opposite effect on autoimmune diseases defines a major haplotype group within clade b, indicating that it may represent (or be in linkage with) a selection target (in a multiallelic selection regime).

The role of the protein product of *PSORS1C1* is presently unknown, although one SNP in *PSORS1C1* immediately downstream the region we analysed has been associated through GWAS with white blood cell counts [41] and a second downstream variant with delayed AIDS progression [42]. Conversely the role of *CDSN* is well studied as the gene encodes corneodesmosin, an adhesive protein which participates in the stabilization of corneodesmosomes in the skin and other cornified squamous epithelia. Loss of coreneodesmosin in humans results in generalised peeling skin disease, a condition characterized by skin barrier defects, pruritus and atopy, with patients also showing *Staphylococcus aureus* superinfections [43]. Mice lacking *CDSN* display a severe phenotype and the skin barrier defect is accompanied by a 10-fold increase in transepithelial water loss. These observations highlight the central role played by the epidermal barrier in protection from infections and in water homeostasis, two processes that are though to have been targeted by natural selection during human evolutionary history. Therefore, SNPs in *CDSN* may have been maintained by balancing selection due to their modulation of the skin barrier properties. Still, the observation that humans with genetic defects in *CDSN* also display food allergies [43] suggests that the protein may play a role in processes unrelated to skin function.

Finally, *BTNL2* encodes a butrophilin-like, widely expressed protein with still poorly understood function. Experiments in mice have suggested that *BTNL2* might function as a negative co-stimulatory molecule with inhibitory activity on T cell activation [44,45]. In particular, a possible modulatory role for the protein in intestine inflammation has been proposed in this animal model, in line with the observation that a SNP (rs9268480) in the region we analysed has also been associated with UC [46].

The haplotype genealogy of the *BTNL2* region we resequenced is peculiar with one deeply separated clade limited to EAS subjects. Previous studies have suggested that such tree topologies might originate from admixture of anatomically modern humans with archaic hominin populations (reviewed in [47]) rather than by a selective force; yet, a TMRCA > 5 MY may be unlikely even under ancient admixture scenarios [47]. Therefore, we suggest that the distantly related clade is maintained by a selective process acting on *BTNL2* or on the nearby class II MHC loci. As reported above, the coalescence time of the two more closely related clades is also deep and several nonsynonymous variants are located on the major branches (Figure 7A). Specifically, 3D modelling of BTNL2 [48] indicated that the Pro285Leu, Met286Ile and Pro299Gln variants are all located in close physical proximity within the IgC domain and adjacent to cystein residues involved in disulfide bonding. These variants are postulated to be functional [48] and can be regarded as potential selection targets. Conversely both the opposite risk alleles and the UC susceptibility variant identify haplotype subsets within the major clade (Figure 7A).

Therefore, even in those regions where we did identify balancing selection signatures, the opposite risk alleles do not always represent (or are in tight linkage with) the selected variants. Rather, our data suggest that, with the exclusion of rs757262 and rs3130981 in *TRIM40* and *CDSN*, opposite risk SNPs are likely to have accumulated as neutral variants, possibly maintained through hitchhiking with the balanced polymorphisms. An alternative possibility is that, as in the case of the gene regions that we described as neutrally evolving, weaker and more recent selective events have been maintaining the opposite risk alleles. In this respect it is worth noting that the haplotype structure of the gene regions we analysed is often complex, raising the possibility that the opposite risk profile identified for some of these SNPs is secondary to association with another causal variant which is not analysed in the association study. In the case of *BTNL2*, for example, we were able to include two opposite risk

alleles in the network analysis; based on the risk profiles of variants 28 and 100 (rs2076530 and rs3129953, respectively, Table 1) we were able to identify a set of haplotypes that should confer increased risk for ATD (they carry 2 predisposing alleles) and a group of haplotypes that predispose to MS (again with two risk alleles) (Figure 7A). This means that rs2076530 and rs3129953 may simply define haplotype subsets that carry causal variants for MS and ATD but when these variants are not typed in the association study, the haplotype structure is such that the derived allele of at position 100 will be under-represented among MS patients for example, resulting in an apparent protective effect. A similar situation might apply to *CDSN*, as well (Figure 6A).

Conclusion

Data herein indicate that balancing selection is common within the xMHC region and involves several non-HLA loci. Yet, the evolutionary history of most SNPs with an opposite effect for autoimmune diseases is consistent with evolutionary neutrality. We suggest that variants with an opposite effect on autoimmune diseases should not be considered a distinct class of disease alleles from the evolutionary perspective. Moreover, in a few cases, the opposite effect on distinct diseases may derive from complex haplotype structures in regions where genetic diversity is high and typing in association studies only captures limited information.

Methods

HapMap samples and sequencing

Human genomic DNA from HapMap subjects 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 Ohio, USA), directly sequenced on both strands with a Big Dye Terminator sequencing Kit (v3.1 Applied Biosystems) and run on an Applied Biosystems ABI 3130 XL Genetic Analyzer (Applied Biosystems). Sequences were assembled using AutoAssembler version 1.4.0 (Applied Biosystems), and inspected manually by two distinct operators. Orangutan (*Pongo Pygmaeus*) genomic DNA (Cell line name: EB185_JC) was obtained from European Collection of Cell Cultures. In order to fill gaps in the reference sequence, we PCR-amplified two genomic region using the following primer sets: F1 (5'-TGGAGTGCAGTGGCATGATC-3')/R1 (5'-TCAGTCTGCCCTCGTCAATG-3') and F2 (5'-CTTGTCAGAGTGGGAGAAGAT-3')/R2 (5'-CTCAGAGGAGTAGAATCCCTG-3'). The PCR products F1/R1 and F2/R2 were sequenced with seqF1 (5'-CTCGTCAGAGTGGGAGAAGAT-3') and R2, respectively.

Data retrieval and haplotype construction

Information on SNPs identified in GWAS were retrieved from the National Human Genome Institute website (A Catalog of Published Genome-Wide Association Studies; http://www.genome.gov/26525384) updated to September the 1st, 2010. Data on gene expression in different human tissues were derived from the BioGPS website http://biogps.gnf.org. 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 Europeans (CEU), Yoruba (YRI) and East Asians (EAS) (NIEHS panel 2).

Haplotypes were inferred using PHASE version 2.1 [49,50]. Linkage disequilibrium analyses were performed using the Haploview (v. 4.1) [51] and blocks were identified through an algorithm implemented in the software.

Data from the Pilot 1 phase of the 1000 Genomes Project were retrieved from the dedicated website http://www.1000genomes.org/. Low coverage SNP genotypes were organized in a MySQL database. A set of programs was developed to retrieve genotypes from the database and to analyse them according to selected regions/populations. These programs were developed in C++ using the GeCo++ [52] and the LibSequence [53] libraries. Genotype information was obtained for all analysed region and for 2,000 regions (5kb in size) randomly derived from an equal number of RefSeq genes.

Statistical analysis

Tajima's D [14], Fu and Li's D* and F* [15] statistics, as well as diversity parameters θ_W [16] and π [17] were calculated using *libsequence* [53]. Calibrated coalescent simulations were performed using the *cosi* package [18] and its best-fit parameters for YRI, EU and EAS populations with 10,000 iterations. Briefly, this model supposes a series of population size changes occurring at each population, along with migratory events, with intensity and duration estimated by fitting demographic parameters with genome-wide real data (see[18] for further details including parameters of the best-fitting model). A major advantage of this model is that it can generates simulated data closely resembling empirical data, rather than accurately infer historical scenarios for human populations. Coalescent simulations were performed conditioned on the local mutation rate, estimated from the number of fixed differences with an out-group species, and on the local recombination rate, retrieved by UCSC Genome Browser tables (ucsc.genome.gov). The maximum-likelihood-ratio HKA test was performed using the MLHKA software [22], as previously proposed [54]. Briefly, 16 reference loci were randomly selected

among NIEHS loci shorter than 20 kb that have been resequenced in the 3 populations; the only criterion was that Tajima's D did not suggest the action of natural selection (i.e. Tajima's D is higher than the 5th and lower than the 95th percentiles in the distribution of NIEHS genes). The reference set was accounted for by the following genes: *VNN3, PLA2G2D, MB, MAD2L2, HRAS, CYP17A1, ATOX1, BNIP3, CDC20, NGB, TUBA1, MT3, NUDT1, PRDX5, RETN* and *JUND*. The chimpanzee, orangutan or macaque sequence was used as the out-group as detailed in the text.

Haplotype analysis and TMRCA calculation

Median-joining networks to infer haplotype genealogy was constructed using NETWORK 4.5 [55]. Estimate of the time to the most common ancestor (TMRCA) derived from application of a maximum-likelihood coalescent method implemented in GENETREE [56,57]. Again, the mutation rate μ was obtained on the basis of the divergence between human and chimpanzee or orangutan or macaque and under the assumption both that the species separation occurred 6, 13 and 23 MY ago, respectively [58] and of a generation time of 25 years. The migration matrix was derived from previous estimated migration rates [18]. Using this μ and θ maximum likelihood (θ_{ML}), we estimated the effective population size parameter (N_e). With these assumptions, the coalescence time, scaled in $2N_e$ units, was converted into years. For the coalescence process, 10^6 simulations were performed. Details on the GENETREE analyses are available in Additional file 5.

All calculations were carried out in the R environment [59].

Additional material

Additional file 1: Schematic representation of the gene regions we resequenced in *ZSCAN23*, *HLA-DMB*, *VARS2*, *C6orf47*, *BAT3*, *PTPN22* and *IL10*.

Additional file 2: Nucleotide diversity indexes and summary statistics calculated from the low-coverage 1000 Genomes data.

Additional file 3: LD map of the genomic regions we resequenced in *TAP2* region (chr6:32901567-32905985), *TRIM40/TRIM10* region (chr6:30221583-30230714), *CDSN/PSORS1C1* region (chr6:31188857-31193360).

Additional File 4: Nucleotide diversity estimates and summary statistics for the sub-regions included and excluded from GENETREE analysis.

Additional file 5: Table with details on GENETREE analyses.

Abbreviations

SNP: Single Nucleotide Polymorphism; MY: million years; YRI: Yorubans; CEU: Europeans; EAS: East Asians; TMRCA: Time to the Most Recent Common Ancestor; D_T, Tajima's D; GWAS: genome-wide association study; CD: Crohn's disease; UC: ulcerative colitis; MS: multiple sclerosis; T1D: type 1 diabetes; RA: rheumatoid arthritis; ATD: autoimmune thyroid disease; AS: and ankylosing spondylitis; SLE: systemic lupus erythematosus.

Acknowledgements

MC is supported by grants from The Eli and Edythe Broad Foundation (IBD-0294), Istituto Superiore di Sanita' "Programma Nazionale di Ricerca sull' AIDS", the EMPRO and AVIP EC WP6 Projects, the nGIN EC WP7 Project, the Japan Health Science Foundation, 2008 Ricerca Finalizzata [Italian Ministry of Health], 2008 Ricerca Corrente [Italian Ministry of Health], Progetto FIRB RETI: Rete Italiana Chimica Farmaceutica CHEM-PROFARMA-NET [RBPR05NWWC], and Fondazione CARIPLO.

Author details

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

Authors' contributions

RC and SR performed all resequencing experiments and analysed the data; MF and UP performed most population genetics analyses; MS, MF, RC, GPC and UP analysed and interpreted the data; NB participated in the study coordination. MS and MC wrote the paper. MS and MC conceived and coordinated the study. All authors read and approved the final manuscript.

Received: 18 March 2011 Accepted: 17 June 2011
Published: 17 June 2011

References

1. Cirulli ET, Goldstein DB: **Uncovering the roles of rare variants in common disease through whole-genome sequencing.** *Nat Rev Genet* 2010, **11(6)**:415-425.
2. Zenewicz LA, Abraham C, Flavell RA, Cho JH: **Unraveling the genetics of autoimmunity.** *Cell* 2010, **140(6)**:791-797.
3. Sirota M, Schaub MA, Batzoglou S, Robinson WH, Butte AJ: **Autoimmune disease classification by inverse association with SNP alleles.** *PLoS Genet* 2009, **5(12)**:e1000792.
4. Wang K, Baldassano R, Zhang H, Qu HQ, Imielinski M, Kugathasan S, Annese V, Dubinsky M, Rotter JI, Russell RK, Bradfield JP, Sleiman PM, Glessner JT, Walters T, Hou C, Kim C, Frackelton EC, Garris M, Doran J, Romano C, Catassi C, Van Limbergen J, Guthery SL, Denson L, Piccoli D, Silverberg MS, Stanley CA, Monos D, Wilson DC, Griffiths A, Grant SF, Satsangi J, Polychronakos C, Hakonarson H: **Comparative genetic analysis of inflammatory bowel disease and type 1 diabetes implicates multiple loci with opposite effects.** *Hum Mol Genet* 2010, **19(10)**:2059-2067.
5. Lettre G, Rioux JD: **Autoimmune diseases: insights from genome-wide association studies.** *Hum Mol Genet* 2008, **17(R2)**:R116-21.
6. Jin Y, Birlea SA, Fain PR, Gowan K, Riccardi SL, Holland PJ, Mailloux CM, Sufit AJ, Hutton SM, Amadi-Myers A, Bennett DC, Wallace MR, McCormack WT, Kemp EH, Gawkrodger DJ, Weetman AP, Picardo M, Leone G, Taieb A, Jouary T, Ezzedine K, van Geel N, Lambert J, Overbeck A, Spritz RA: **Variant of TYR and autoimmunity susceptibility loci in generalized vitiligo.** *N Engl J Med* 2010, **362(18)**:1686-1697.
7. Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, Bustamante CD, Teshima KM, Przeworski M: **Natural selection on genes that underlie human disease susceptibility.** *Curr Biol* 2008, **18(12)**:883-889.
8. Meyer D, Thomson G: **How selection shapes variation of the human major histocompatibility complex: a review.** *Ann Hum Genet* 2001, **65(Pt 1)**:1-26.
9. Charlesworth D: **Balancing selection and its effects on sequences in nearby genome regions.** *PLoS Genet* 2006, **2(4)**:e64.
10. Choy MK, Phipps ME: **MICA polymorphism: biology and importance in immunity and disease.** *Trends Mol Med* 2010, **16(3)**:97-106.
11. Qu HQ, Lu Y, Marchand L, Bacot F, Frechette R, Tessier MC, Montpetit A, Polychronakos C: **Genetic control of alternative splicing in the TAP2 gene: possible implication in the genetics of type 1 diabetes.** *Diabetes* 2007, **56(1)**:270-275.

12. Fumagalli M, Pozzoli U, Cagliani R, Comi GP, Riva S, Clerici M, Bresolin N, Sironi M: **Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions.** *J Exp Med* 2009, **206(6)**:1395-1408.
13. Wilson JN, Rockett K, Keating B, Jallow M, Pinder M, Sisay-Joof F, Newport M, Kwiatkowski D: **A hallmark of balancing selection is present at the promoter region of interleukin 10.** *Genes Immun* 2006, **7(8)**:680-683.
14. Tajima F: **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.** *Genetics* 1989, **123(3)**:585-595.
15. Fu YX, Li WH: **Statistical tests of neutrality of mutations.** *Genetics* 1993, **133(3)**:693-709.
16. Watterson GA: **On the number of segregating sites in genetical models without recombination.** *Theor Popul Biol* 1975, **7(2)**:256-276.
17. Nei M, Li WH: **Mathematical model for studying genetic variation in terms of restriction endonucleases.** *Proc Natl Acad Sci USA* 1979, **76(10)**:5269-5273.
18. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D: **Calibrating a coalescent simulation of human genome sequence variation.** *Genome Res* 2005, **15(11)**:1576-1583.
19. Salamon H, Klitz W, Easteal S, Gao X, Erlich HA, Fernandez-Vina M, Trachtenberg EA, McWeeney SK, Nelson MP, Thomson G: **Evolution of HLA class II molecules: Allelic and amino acid site variability across populations.** *Genetics* 1999, **152(1)**:393-400.
20. Loisel DA, Rockman MV, Wray GA, Altmann J, Alberts SC: **Ancient polymorphism and functional variation in the primate MHC-DQA1 5′ cis-regulatory region.** *Proc Natl Acad Sci USA* 2006, **103(44)**:16331-16336.
21. Kimura M: *The Neutral Theory of Molecular Evolution* Cambridge: Cambridge University Press; 1983.
22. Wright SI, Charlesworth B: **The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model.** *Genetics* 2004, **168(2)**:1071-1076.
23. 1000 Genomes Project Consortium, Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467(7319)**:1061-1073.
24. Tishkoff SA, Verrelli BC: **Patterns of human genetic diversity: implications for human evolutionary history and disease.** *Annu Rev Genomics Hum Genet* 2003, **4**:293-340.
25. Cagliani R, Fumagalli M, Biasin M, Piacentini L, Riva S, Pozzoli U, Bonaglia MC, Bresolin N, Clerici M, Sironi M: **Long-term balancing selection maintains trans-specific polymorphisms in the human TRIM5 gene.** *Hum Genet* 2010.
26. Cagliani R, Riva S, Biasin M, Fumagalli M, Pozzoli U, Lo Caputo S, Mazzotta F, Piacentini L, Bresolin N, Clerici M, Sironi M: **Genetic diversity at endoplasmic reticulum aminopeptidases is maintained by balancing selection and is associated with natural resistance to HIV-1 infection.** *Hum Mol Genet* 2010.
27. Gongora R, Figueroa F, Klein J: **The HLA-DRB9 gene and the origin of HLA-DR haplotypes.** *Hum Immunol* 1996, **51(1)**:23-31.
28. Jacobson DL, Gange SJ, Rose NR, Graham NM: **Epidemiology and estimated population burden of selected autoimmune diseases in the United States.** *Clin Immunol Immunopathol* 1997, **84(3)**:223-243.
29. Sironi M, Clerici M: **The hygiene hypothesis: an evolutionary perspective.** *Microbes Infect* 2010, **12(6)**:421-427.
30. Plenge R: **GWASs and the age of human as the model organism for autoimmune genetic research.** *Genome Biol* 2010, **11(5)**:212.
31. Barreiro LB, Quintana-Murci L: **From evolutionary genetics to human immunology: how selection shapes host defence genes.** *Nat Rev Genet* 2010, **11(1)**:17-30.
32. Fumagalli M, Cagliani R, Riva S, Pozzoli U, Biasin M, Piacentini L, Comi GP, Bresolin N, Clerici M, Sironi M: **Population genetics of IFIH1: ancient population structure, local selection and implications for susceptibility to type 1 diabetes.** *Mol Biol Evol* 2010.
33. Ferrer-Admetlla A, Bosch E, Sikora M, Marques-Bonet T, Ramirez-Soriano A, Muntasell A, Navarro A, Lazarus R, Calafell F, Bertranpetit J, Casals F: **Balancing selection is the main force shaping the evolution of innate immunity genes.** *J Immunol* 2008, **181(2)**:1315-1322.
34. Vang T, Congia M, Macis MD, Musumeci L, Orru V, Zavattari P, Nika K, Tautz L, Tasken K, Cucca F, Mustelin T, Bottini N: **Autoimmune-associated lymphoid tyrosine phosphatase is a gain-of-function variant.** *Nat Genet* 2005, **37(12)**:1317-1319.
35. Chung SA, Criswell LA: **PTPN22: its role in SLE and autoimmunity.** *Autoimmunity* 2007, **40(8)**:582-590.
36. Hansen TH, Bouvier M: **MHC class I antigen presentation: learning from viral evasion strategies.** *Nat Rev Immunol* 2009, **9(7)**:503-513.
37. Liu C, Carrington M, Kaslow RA, Gao X, Rinaldo CR, Jacobson LP, Margolick JB, Phair J, O'Brien SJ, Detels R: **Association of polymorphisms in human leukocyte antigen class I and transporter associated with antigen processing genes with resistance to human immunodeficiency virus type 1 infection.** *J Infect Dis* 2003, **187(9)**:1404-1410.
38. Harada H, Harada Y, O'Brien DP, Rice DS, Naeve CW, Downing JR: **HERF1, a novel hematopoiesis-specific RING finger protein, is required for terminal differentiation of erythroid cells.** *Mol Cell Biol* 1999, **19(5)**:3808-3815.
39. Andres AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, Boyko AR, Gutenkunst RN, White TJ, Green ED, Bustamante CD, Clark AG, Nielsen R: **Targets of balancing selection in the human genome.** *Mol Biol Evol* 2009, **26(12)**:2755-2764.
40. Capon F, Allen MH, Ameen M, Burden AD, Tillman D, Barker JN, Trembath RC: **A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups.** *Hum Mol Genet* 2004, **13(20)**:2361-2368.
41. Kamatani Y, Matsuda K, Okada Y, Kubo M, Hosono N, Daigo Y, Nakamura Y, Kamatani N: **Genome-wide association study of hematological and biochemical traits in a Japanese population.** *Nat Genet* 2010, **42(3)**:210-215.
42. Limou S, Le Clerc S, Coulonges C, Carpentier W, Dina C, Delaneau O, Labib T, Taing L, Sladek R, Deveau C, Ratsimandresey R, Montes M, Spadoni JL, Lelievre JD, Levy Y, Therwath A, Schachter F, Matsuda F, Gut I, Froguel P, Delfraissy JF, Hercberg S, Zagury JF, ANRS Genomic Group: **Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02).** *J Infect Dis* 2009, **199(3)**:419-426.
43. Oji V, Eckl KM, Aufenvenne K, Natebus M, Tarinski T, Ackermann K, Seller N, Metze D, Nurnberg G, Folster-Holst R, Schafer-Korting M, Hausser I, Traupe H, Hennies HC: **Loss of corneodesmosin leads to severe skin barrier defect, pruritus, and atopy: unraveling the peeling skin disease.** *Am J Hum Genet* 2010, **87(2)**:274-281.
44. Nguyen T, Liu XK, Zhang Y, Dong C: **BTNL2, a butyrophilin-like molecule that functions to inhibit T cell activation.** *J Immunol* 2006, **176(12)**:7354-7360.
45. Arnett HA, Escobar SS, Gonzalez-Suarez E, Budelsky AL, Steffen LA, Boiani N, Zhang M, Siu G, Brewer AW, Viney JL: **BTNL2, a butyrophilin/B7-like molecule, is a negative costimulatory molecule modulated in intestinal inflammation.** *J Immunol* 2007, **178(3)**:1523-1533.
46. Asano K, Matsushita T, Umeno J, Hosono N, Takahashi A, Kawaguchi T, Matsumoto T, Matsui T, Kakuta Y, Kinouchi Y, Shimosegawa T, Hosokawa M, Arimura Y, Shinomura Y, Kiyohara Y, Tsunoda T, Kamatani N, Iida M, Nakamura Y, Kubo M: **A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population.** *Nat Genet* 2009, **41(12)**:1325-1329.
47. Garrigan D, Hammer MF: **Reconstructing human origins in the genomic era.** *Nat Rev Genet* 2006, **7(9)**:669-680.
48. Valentonyte R, Hampe J, Huse K, Rosenstiel P, Albrecht M, Stenzel A, Nagy M, Gaede KI, Franke A, Haesler R, Koch A, Lengauer T, Seegert D, Reiling N, Ehlers S, Schwinger E, Platzer M, Krawczak M, Muller-Quernheim J, Schurmann M, Schreiber S: **Sarcoidosis is associated with a truncating splice site mutation in BTNL2.** *Nat Genet* 2005, **37(4)**:357-364.
49. Stephens M, Smith NJ, Donnelly P: **A new statistical method for haplotype reconstruction from population data.** *Am J Hum Genet* 2001, **68(4)**:978-989.
50. Stephens M, Scheet P: **Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation.** *Am J Hum Genet* 2005, **76(3)**:449-462.
51. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps.** *Bioinformatics* 2005, **21(2)**:263-265.
52. Cereda M, Sironi M, Cavalleri M, Pozzoli U: **GeCo++: a C++ library for genomic features computation and annotation in the presence of variants.** *Bioinformatics* 2011, **27(9)**:1313-1315.
53. Thornton K: **Libsequence: a C++ class library for evolutionary genetic analysis.** *Bioinformatics* 2003, **19(17)**:2325-2327.

54. Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi GP, Menozzi G, Bresolin N, Sironi M: **Widespread balancing selection and pathogen-driven selection at blood group antigen genes.** *Genome Res* 2009, **19(2)**:199-212.
55. Bandelt HJ, Forster P, Rohl A: **Median-joining networks for inferring intraspecific phylogenies.** *Mol Biol Evol* 1999, **16(1)**:37-48.
56. Griffiths RC, Tavare S: **Unrooted genealogical tree probabilities in the infinitely-many-sites model.** *Math Biosci* 1995, **127(1)**:77-98.
57. Griffiths RC, Tavare S: **Sampling theory for neutral alleles in a varying environment.** *Philos Trans R Soc Lond B Biol Sci* 1994, **344(1310)**:403-410.
58. Glazko GV, Nei M: **Estimation of divergence times for major lineages of primate species.** *Mol Biol Evol* 2003, **20(3)**:424-434.
59. R Development Core Team: *R: A Language and Environment for Statistical Computing* Vienna, Austria; 2008.
60. Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, Plagnol V, Bailey R, Nejentsev S, Field SF, Payne F, Lowe CE, Szeszko JS, Hafler JP, Zeitels L, Yang JH, Vella A, Nutland S, Stevens HE, Schuilenburg H, Coleman G, Maisuria M, Meadows W, Smink LJ, Healy B, Burren OS, Lam AA, Ovington NR, Allen J, Adlem E, Leung HT, Wallace C, Howson JM, Guja C, Ionescu-Tirgoviste C, Genetics of Type 1 Diabetes in Finland, Simmonds MJ, Heward JM, Gough SC, Wellcome Trust Case Control Consortium, Dunger DB, Wicker LS, Clayton DG: **Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes.** *Nat Genet* 2007, **39(7)**:857-864.
61. Hakonarson H, Grant SF, Bradfield JP, Marchand L, Kim CE, Glessner JT, Grabs R, Casalunovo T, Taback SP, Frackelton EC, Lawson ML, Robinson LJ, Skraban R, Lu Y, Chiavacci RM, Stanley CA, Kirsch SE, Rappaport EF, Orange JS, Monos DS, Devoto M, Qu HQ, Polychronakos C: **A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene.** *Nature* 2007, **448(7153)**:591-594.
62. Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, Ding B, Liew A, Khalili H, Chandrasekaran A, Davies LR, Li W, Tan AK, Bonnard C, Ong RT, Thalamuthu A, Pettersson S, Liu C, Tian C, Chen WV, Carulli JP, Beckman EM, Altshuler D, Alfredsson L, Criswell LA, Amos CI, Seldin MF, Kastner DL, Klareskog L, Gregersen PK: **TRAF1-C5 as a risk locus for rheumatoid arthritis–a genomewide study.** *N Engl J Med* 2007, **357(12)**:1199-1209.
63. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ, NIDDK IBD Genetics Consortium, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot JP, de Vos M, Vermeire S, Louis E, Belgian-French IBD Consortium, Wellcome Trust Case Control Consortium, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ: **Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease.** *Nat Genet* 2008, **40(8)**:955-962.
64. Kyogoku C, Langefeld CD, Ortmann WA, Lee A, Selby S, Carlton VE, Chang M, Ramos P, Baechler EC, Batliwalla FM, Novitzke J, Williams AH, Gillett C, Rodine P, Graham RR, Ardlie KG, Gaffney PM, Moser KL, Petri M, Begovich AB, Gregersen PK, Behrens TW: **Genetic association of the R620W polymorphism of protein tyrosine phosphatase PTPN22 with human SLE.** *Am J Hum Genet* 2004, **75(3)**:504-507.
65. McGovern DP, Gardet A, Torkvist L, Goyette P, Essers J, Taylor KD, Neale BM, Ong RT, Lagace C, Li C, Green T, Stevens CR, Beauchamp C, Fleshner PR, Carlson M, D'Amato M, Halfvarson J, Hibberd ML, Lordal M, Padyukov L, Andriulli A, Colombo E, Latiano A, Palmieri O, Bernard EJ, Deslandres C, Hommes DW, de Jong DJ, Stokkers PC, Weersma RK, NIDDK IBD Genetics Consortium, Sharma Y, Silverberg MS, Cho JH, Wu J, Roeder K, Brant SR, Schumm LP, Duerr RH, Dubinsky MC, Glazer NL, Haritunians T, Ippoliti A, Melmed GY, Siscovick DS, Vasiliauskas EA, Targan SR, Annese V, Wijmenga C, Pettersson S, Rotter JI, Xavier RJ, Daly MJ, Rioux JD, Seielstad M: **Genome-wide association identifies multiple ulcerative colitis susceptibility loci.** *Nat Genet* 2010, **42(4)**:332-337.
66. Nair RP, Duffin KC, Helms C, Ding J, Stuart PE, Goldgar D, Gudjonsson JE, Li Y, Tejasvi T, Feng BJ, Ruether A, Schreiber S, Weichenthal M, Gladman D, Rahman P, Schrodi SJ, Prahalad S, Guthery SL, Fischer J, Liao W, Kwok PY, Menter A, Lathrop GM, Wise CA, Begovich AB, Voorhees JJ, Elder JT, Krueger GG, Bowcock AM, Abecasis GR, Collaborative Association Study of Psoriasis: **Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways.** *Nat Genet* 2009, **41(2)**:199-204.
67. Jakkula E, Leppa V, Sulonen AM, Varilo T, Kallio S, Kemppinen A, Purcell S, Koivisto K, Tienari P, Sumelahti ML, Elovaara I, Pirttila T, Reunanen M, Aromaa A, Oturai AB, Sondergaard HB, Harbo HF, Mero IL, Gabriel SB, Mirel DB, Hauser SL, Kappos L, Polman C, De Jager PL, Hafler DA, Daly MJ, Palotie A, Saarela J, Peltonen L: **Genome-wide association study in a high-risk isolate for multiple sclerosis reveals associated variants in STAT3 gene.** *Am J Hum Genet* 2010, **86(2)**:285-291.
68. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ, NIDDK IBD Genetics Consortium, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot JP, de Vos M, Vermeire S, Louis E, Belgian-French IBD Consortium, Wellcome Trust Case Control Consortium, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ: **Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease.** *Nat Genet* 2008, **40(8)**:955-962.

doi:10.1186/1471-2148-11-171
Cite this article as: Cagliani *et al.*: **Balancing selection is common in the extended MHC region but most alleles with opposite risk profile for autoimmune diseases are neutrally evolving.** *BMC Evolutionary Biology* 2011 **11**:171.