Immunogenetics (2013) 65:479–483
DOI 10.1007/s00251-013-0693-x

BRIEF COMMUNICATION

Current perspectives on the intensity of natural selection of MHC loci

Yoshiki Yasukochi · Yoko Satta

Received: 29 December 2012 / Accepted: 5 March 2013 / Published online: 3 April 2013

Abstract Polymorphism of genes in the major histocompatibility complex (*MHC*) is believed to be maintained by balancing selection. However, direct evidence of selection has proven difficult to demonstrate. In 1994, Satta and colleagues estimated the selection intensity of the human *MHC* (human leukocyte antigen (*HLA*)) loci; however, at that time the number of *HLA* sequences was limited. By comparing five different methods, this study demonstrated the best way to calculate the selection coefficient, through a computer simulation study. Since the study, many *HLA* nucleotide sequences have been made available. Our new analysis takes advantage of these newly available sequences and compares new estimates with those of the previous study. Generally, our new results are consistent with those of the 1994 study. Our results show that, even after 20 years of exhaustive sequencing of human *HLA*, the number of dominant *HLA* alleles, on which our original estimate of selection intensity depended, appears to be conserved. Indeed, according to the frequency distribution for each *HLA* allele, most sequences in the database were minor or private alleles; therefore, we conclude that the selection intensities of *HLA* loci are at most 4.4 % even though the *HLA* is the prominent example on which the natural selection has been operating.

Keywords Allelic genealogy · Balancing selection · HLA · MHC · Selection intensity · Symmetric overdominant selection

Electronic supplementary material The online version of this article (doi:10.1007/s00251-013-0693-x) contains supplementary material, which is available to authorized users.

Y. Yasukochi (✉) · Y. Satta
Department of Evolutionary Studies of Biosystems, The Graduate University for Advanced Studies (SOKENDAI), Shonan Village, Hayama, Kanagawa 240-0193, Japan
e-mail: hyasukou@proof.ocn.ne.jp

The large extent of polymorphism of major histocompatibility complex (*MHC*) genes is believed to be maintained by balancing selection for the extent of the peptide binding repertoire between individuals (Hughes and Nei 1988, 1989; Takahata and Nei 1990; Hughes and Yeager 1998). A unique effect of balancing selection is the long persistence time of alleles in populations and, consequently, trans-species polymorphism (Klein 1987; Takahata 1990; Takahata et al. 1992; Klein et al. 1998, 2007). However, it is difficult to show direct evidence of such selection by experiments and to measure selection intensity directly. Satta et al. (1994) estimated the intensity of selection at the human *MHC* (human leukocyte antigen (*HLA*)) loci by using the available collection of allelic sequences and a simple model based on symmetric overdominant selection and the theory of allelic genealogy (Kimura and Crow 1964; Takahata 1990; Takahata and Nei 1990; Takahata et al. 1992).

In recent years, a number of *HLA* allelic nucleotide sequences have become available through IMGT/HLA database (http://www.ebi.ac.uk/imgt/hla/, Robinson et al. 2011). Currently (2012), the database contains 7,670 alleles. This large dataset of sequences provides an opportunity to estimate more reliable evolutionary parameters, such as natural selection intensity. Hence, we re-estimated the selection coefficient and compared the estimates with those in the previous study that was based on a limited number of sequences (Satta et al. 1994).

The large number of nucleotide sequences at the six functional *HLA* loci (*HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DQB1*, and *HLA-DPB1*), which play important roles in peptide presentation, was obtained from the IMGT/HLA database. In addition, nucleotide sequences of alleles at the *HLA* class II A (*DQA1* and *DPA1*) and class II B (*DRB3* and *DRB5*) loci were also used in this analysis. Because the inclusion of recombinants will lead a biased estimation of the selection intensity, possible recombinant alleles were excluded by using

the method described by Satta (1992). This method assumes that the relationship between the number of substitutions in a particular region and the number of substitutions in the entire region is binomially distributed. At the *HLA-B* locus, an exceptionally divergent *HLA-B*73:01* allele (Abi-Rached et al. 2011), which might have been transmitted to extant humans from a distinct *Homo* by interbreeding, was also excluded from this analysis. Applying the theory of allelic genealogy under symmetric overdominant selection to this analysis, we used only dominant alleles that have a frequency >1 % throughout various human populations (the NCBI dbMHC database, http://www.ncbi.nlm.nih.gov/gv/mhc, Meyer et al. 2007). We also excluded the nucleotide sequences with a wide range of undetermined nucleotides from this analysis (Table 1). Therefore, the number for alleles used in this analysis was limited to 9 *HLA-A* alleles, 19 *HLA-B*, 20 *HLA-C*, 25 *HLA-DRB1*, 13 *HLA-DQB1*, 10 *HLA-DPB1*, 6 *HLA-DQA1*, 3 *HLA-DPA1*, 13 *HLA-DRB3*, and 5 *HLA-DRB5*. These *HLA* alleles are listed in Online Resource 1. Interestingly, most of the enormously large numbers of nucleotide sequences in the current database are minor or private alleles.

According to the theory described in Takahata (1990) and Takahata et al. (1992), to estimate the selection coefficient s, two estimators, γ and K_B, must be calculated. The estimator γ is the ratio of the number of nonsynonymous substitutions per peptide-binding region (PBR) site to that of synonymous substitutions per site among given pairs of alleles, whereas K_B is the mean number of pairwise nonsynonymous substitutions in the PBR. The number of synonymous and nonsynonymous sites was estimated using the modified Nei–Gojobori method (Zhang et al. 1998) with the Jukes–Cantor correction (Jukes and Cantor 1969). Because of the relatively early ceiling in the number of nonsynonymous substitutions in the PBR due to acceleration of the nucleotide substitution rate by balancing selection, Satta et al. (1994) developed five methods for estimating K_B, and these methods were evaluated by computer simulations. Here, we used method II because this method minimized errors in the multiple-hit correction (Satta et al. 1994). In this method, selection coefficients can be adequately estimated by using only sets of sequences that are relatively closely related.

The estimated values of K_B and γ at the six major *HLA* loci described above are provided in Table 2. Using these values, we obtained other estimators, M and S, which were also necessary for estimating the selection coefficient, s (see Satta et al. 1994). Assuming that a long-term effective population size of humans is 10^5, the s values of *HLA-B* and *HLA-DRB1* loci (s=4.4 and 1.9 %, respectively) in the present study were the highest for the class I and class II loci, respectively. This result was consistent with that of the previous study (Satta et al. 1994). All s values were more or less similar to those of the previous study with the exception of *DQB1* and *DPB1* loci: the current estimate of *DQB1* was lower than the previous estimate and the value for *DPB1* was much higher than the previous estimate (Satta et al. 1994). One possible reason for this is the different set of nucleotides sequences used than the previous study. In fact, both for *DQB1* and *DPB1*, the number of dominant alleles used in the present analysis increases compared to that of the previous one.

Allelic genealogy predicts that K_B is approximately equal to the number of dominant alleles (n_a) in a population. In

Table 1 The number of alleles and dominant alleles in the database

HLA locus	No. of alleles				No. of PBR in different alleles	
	In the database	Whole	Nonrecombinant	Dominant[a, b]	Nonrecombinant[b]	Dominant[b]
A	1,594	156	50	27[c] (18)[d]	32	26
B	2,123	235	143	40[c] (21)[d]	113	39
C	1,102	143	129	20[c]	60	19
DRB1	975	64	56	26 (1)[d]	37	23
DQB1	144	(61)[e]	55	13	13	10
DPB1	145	(44)[e]	38	11 (1)[d]	23	11
DRB3	58	(13)[e]	13	–	6	–
DRB5	20	(5)[e]	5	–	4	–
DQA1	47	34	31	8 (2)[d]	9	5
DPA1	34	(11)[e]	9	5 (2)[d]	3	3

[a] The number of dominant alleles that have a high frequency (>1 %) throughout human populations worldwide (including possible recombinants)

[b] The number of amino acid sequences

[c] The number of dominant alleles that are detected in >100 chromosomes from >25 human populations

[d] The number of dominant alleles that are excluded due to a possible recombinant or short sequence

[e] Not whole coding sequence (see text)

Table 2 Estimates of the mean number of nonsynonymous substitutions, the relative nonsynonymous substitution rate in the PBR, and the selection coefficient (s)

HLA locus	Length[a]	L_S[a]	L_B[a]	L_N[a]	No. of allele 1[b]	No. of allele 2[c]	K_B	γ	S	M	s
A	1,095 bp	295	123	674	27[d]	9	28.9 (26.0)	7.6 (6.3)	4,500 (3,000)	0.04 (0.09)	2.25 % (1.50 %)
B	1,086 bp	300	122	643	40[d]	19	35.9 (36.0)	9.7 (9.0)	8,825 (8,200)	0.01 (0.02)	4.41 % (4.20 %)
C	1,093 bp	301	125	665	20[d]	20	17.3 (15.0)	4.9 (3.4)	1,030 (530)	0.15 (0.29)	0.52 % (0.26 %)
DRB1	795 bp	223	53	521	26	25	23.2 (25.0)	10.2 (9.3)	3,890 (3,900)	0.01 (0.01)	1.94 % (1.90 %)
DQB1	687 bp	148	51	347	13	13	12.4 (20.0)	4.4 (6.0)	479 (1,700)	0.14 (0.08)	0.24 % (0.85 %)
DPB1	543 bp	146	53	344	11	10	11.9 (6.8)	9.2 (4.3)	918 (140)	0.01 (0.08)	0.46 % (0.07 %)
DRB3	549 bp	148	54	347	(13[e])	–	5.6 –	5.4 –	120 –	0.04 –	0.06 % –
DRB5	549 bp	148	53	348	(5[e])	–	8.0 –	7.9 –	360 –	0.01 –	0.18 % –
DQA1	765 bp	211	47	504	8	6	5.9 (13.0)	2.1 (4.5)	53 (550)	0.23 (0.14)	0.03 % (0.28 %)
DPA1	663 bp	190	42	428	5	3	4.8 –	3.3 –	54 –	0.10 –	0.03 % –

The numbers of sites of synonymous and nonsynonymous substitutions were estimated using the modified Nei–Gojobori model (R=1.04 for class I, R=1.14 for class II). The parameter values in parentheses were estimated on the basis of method II described in Satta (1992). The mutation rate per PBR per generation (u)=1.7×10^{-6} for class I loci and 7.5×10^{-7} for class II loci; effective population size (N_e)=10^5 (see Satta et al. 1994)

L_S the number of synonymous sites across the entire region, L_B the number of nonsynonymous sites at the PBR, L_N the number of nonsynonymous sites at the non-PBR

[a] The length or the number of sites used in this study (not in the previous study)

[b] The number of dominant alleles that have a high frequency (>1 %) throughout human populations worldwide (shown as n_a in text)

[c] The number of dominant alleles excluding possible recombinants

[d] The number of dominant alleles that are detected in >100 chromosomes from >25 human populations

[e] The number of alleles not derived from the dominant allele because of lack of information about allele frequencies in the human populations

fact, n_a showed good agreement with K_B in three class II B loci (Table 2). In class I loci, the *HLA-C* showed relatively good agreement between n_a and K_B, whereas for the *HLA-A* and *HLA-B* loci, the observed number of dominant alleles was less than the expected number. This discrepancy might indicate that the definition of dominant alleles is inappropriate for class I loci. Originally, we regarded an allele with a frequency of more than 1 % over all populations examined as a dominant allele. According to the dbMHC database, the number of chromosomes examined at all three class I loci was more than 10,000 in total, ranging from allele to allele. Thus, we defined 1 % (100 chromosomes) of 10,000 chromosomes as a class I dominant allele. In addition, the mean number of populations in which class II dominant alleles were observed was about 25. Therefore, for class I loci, we considered the alleles detected on >100 chromosomes through >25 populations as a dominant allele. Surprisingly, n_a of class I loci under this new definition showed good agreement with K_B (Table 2). This might imply that some dominant alleles, with <1 % allele frequency in the entire world population, were dominantly distributed throughout the human population until quite recently and that they have decreased in frequency because their alleles might be replaced by other alleles that had an advantage in the modern environments of some populations. The number of different dominant alleles in the PBR also shows good agreement with expectations (Table 1). After the exclusion of possible recombinants, the numbers at each locus were 26 at *HLA-A*, 39 at *HLA-B*, and 19 at *HLA-C*. However, when we included rare alleles, these numbers increased to 32, 113, and 60, respectively. The number of rare alleles which have de novo PBR nonsynonymous mutations is large and they may have emerged by a population expansion quite recently (Fu et al. 2013).

In addition to the above estimates, we further estimated the selection coefficients for *DRB3*, *DRB4*, *DRB5*, *DQA1*, and *DPA1* (Table 2). With the exception of *DRB4* (see below), all selection coefficient *s* of the four *HLA* class II loci were lower than those of the six major *HLA* loci (*HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DQB1* and *HLA-DPB1*), indicating that the six major loci have been strongly affected by balancing selection. The present *s* estimate of *DQA1* is lower than that of the previous one, but the present K_B value is similar to the n_a. We consider that the present estimate is close to the true value. For *DRB4*, 15 alleles were deposited in the database and they are identical at the PBR sites and nearly identical at the neutral (synonymous and non-PBR nonsynonymous) sites. Thus, inference of the γ and K_B values is difficult. The relatively recent emergence of *DRB4* (the per site nucleotide divergence from *DRB2* is 0.015~0.017: Satta et al. 1996) supports this observation. In addition, the small amount of nucleotide divergence at neutral sites for *DRB4* indicates the relatively small effective population size of *DRB4*. This suggests that the frequency of *DR53* haplotype on which *DRB4* resides is relatively lower than that of other *HLA* haplotypes. In addition, *DRB3* and *DRB5* also show the smaller effective size than that of other *HLA* loci (The estimated N_e values of *DRB3* and *DRB5* are quite smaller than 10^5). This is also because that *DRB3* and *DRB5* are located on a limited *DR* haplotype, whereas other *HLA* loci exist in all humans.

Our findings show that although the number of sequences in the database has greatly increased in the past 20 years, most of the accumulated sequences are minor or private alleles and the number of dominant alleles does not change largely since the previous estimation. Therefore, most of selection coefficients in the six major *HLA* loci estimated in the present study were similar to those of the previous study. One may consider that application of symmetrical overdominance is too strict for the actual data. However, the simulation study by Takahata and Nei (1990) reveals that the asymmetrical overdominance model does not fit the mode of polymorphism for actual data: under a given selection coefficient of asymmetrical model, the number of alleles and the average heterozygosity become smaller than those under symmetrical overdominance model. In fact, the number of dominant alleles at all *HLA* loci was consistent with the K_B values under symmetrical overdominance, suggesting the consistency between our assumed model and the actual data. Therefore, the overdominance model is appropriate to the present estimation. Through this analysis, we confirmed that the selection intensity (selection coefficient, *s*) of *HLA* loci in modern humans is at most 4.4 %, even though *HLA* is the prominent example on which natural selection acts.

Acknowledgments This work was supported by Grant-in-Aid for Scientific Research on Innovative Areas from the Ministry of Education, Culture, Sports, Science and Technology of Japan (22133007). The authors thank John A. Eimes for the critical checking of the English language of this manuscript. We owe special thanks to Naoyuki Takahata for providing valuable comments.

References

Abi-Rached L, Jobin MJ, Kulkarni S et al (2011) The shaping of modern human immune systems by multiregional admixture with archaic humans. Science 334:89–94

Fu W, O' Connor TD, Jun G et al (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493:216–220

Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167–170

Hughes AL, Nei M (1989) Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc Natl Acad Sci U S A 86:958–962

Hughes AL, Yeager M (1998) Natural selection at major histocompatibility complex loci of vertebrates. Annu Rev Genet 32:415–435

Jukes T, Cantor C (1969) Evolution of protein molecules. In: Munro H (ed) Mammalian protein metabolism. Academic, New York, pp 21–132

Kimura M, Crow JF (1964) The number of alleles that can be maintained in a finite population. Genetics 49:725–738

Klein J (1987) Origin of major histocompatibility complex polymorphism: the trans-species hypothesis. Hum Immunol 19:155–162

Klein J, Sato A, Nagl S, O'hUigin C (1998) Molecular trans-species polymorphism. Annu Rev Ecol Systemat 29:1–21

Klein J, Sato A, Nikolaidis N (2007) MHC, TSP, and the origin of species: from immunogenetics to evolutionary genetics. Annu Rev Genet 41:281–304

Meyer D et al (2007) Single locus polymorphism of classical HLA genes. In: Hansen JA (ed) Immunobiology of the human MHC: Proceedings of the 13th International Histocompatibility Workshop and Conference, vol. I. IHWG, Seattle, WA, pp 653–704

Robinson J, Mistry K, McWilliam H, Lopez R, Parham P, Marsh SGE (2011) The IMGT/HLA database. Nucleic Acids Res 39:D1171–1176

Satta Y (1992) Balancing selection at HLA loci. In: Takahata N (ed) In The Proceedings of the 17th Taniguchi Symposium. Japan Science Society, Tokyo, pp 111–131

Satta Y, O'hUigin C, Takahata N, Klein J (1994) Intensity of natural selection at the major histocompatibility complex loci. Proc Natl Acad Sci U S A 91:7184–7188

Satta Y, Mayer WE, Klein J (1996) Evolutionary relationship of HLA-DRB genes inferred from intron sequences. J Mol Evol 42:648–657

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

Takahata N, Nei M (1990) Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics 124:967–978

Takahata N, Satta Y, Klein J (1992) Polymorphism and balancing selection at major histocompatibility complex loci. Genetics 130:925–938

Zhang J, Rosenberg HF, Nei M (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci U S A 95:3708–3713