insertions and their associations with *HLA* alleles in a Chinese population from Malaysia. *Tissue Antigens* 2007: **70**: 136–43.
15. Tian W, Wang F, Cai JH, Li LX. Polymorphic insertions in 5 Alu loci within the major histocompatibility complex class I region and their linkage disequilibria with *HLA* alleles in four distinct populations in mainland China. *Tissue Antigens* 2008: **72**: 559–67.
16. Yao Y, Shi L, Shi L et al. The association between *HLA-A, -B* alleles and major histocompatibility complex class I polymorphic *Alu* insertions in four populations in China. *Tissue Antigens* 2009: **73**: 575–81.
17. Dunn DS, Inoko H, Kulski JK. The association between non-melanoma skin cancer and a young dimorphic *Alu* element within the major histocompatibility complex class I genomic region. *Tissue Antigens* 2006: **68**: 127–34.
18. Stewart CA, Horton R, Allcock RJ et al. Complete *MHC* haplotype sequencing for common disease gene mapping. *Genome Res* 2004: **14**: 1176–87.
19. Moriyama Y, Kato K, Mura T, Juji T. Analysis of *HLA* gene frequencies and *HLA* haplotype frequencies for bone marrow donors in Japan (in Japanese). *MHC* 2006: **12**: 83–201.
20. Kulski JK, Shigenari A, Shiina T et al. Human endogenous retrovirus (HERVK9) structural polymorphism with haplotypic *HLA-A* allelic associations. *Genetics* 2008: **180**: 445–57.
21. Marsh SG. WHO Nomenclature Committee for Factors of the *HLA* System. Nomenclature for factors of the *HLA* system, update July 2000. *Tissue Antigens* 2000: **56**: 476–7.
22. Andersson G. Evolution of the *HLA*-DR region. *Front Biosci* 1998: **3**: d739–45.
23. Dorak MT, Lawson T, Machulla HK, Mills KI, Burnett AK. Increased heterozygosity for *MHC* class II lineages in newborn males. *Genes Immun* 2002: **3**: 263–9.
24. Raymond M, Rousset F. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J Hered* 1995: **86**: 248–9.
25. Sasieni PD. From genotypes to genes: doubling the sample size. *Biometrics* 1997: **53**: 1253–61.
26. Meirmans PG, Van Tienderen PH. GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Mol Ecol Notes* 2004: **4**: 792–4.
27. Gaunt TR, Rodriguez S, Carlos Zapata C, Day INM. MIDAS: software for analysis and visualisation of interallelic disequilibrium between multiallelic markers. *BMC Bioinformatics* 2006: **7**: 227–38.
28. Excoffier L, Laval G, Schneider S. Arlequin ver. 3.0: an integrated software package for population genetics data analysis. *Evol Bioinform Online* 2005: **1**: 47–50.
29. Perneger TV. What is wrong with Bonferroni adjustments. *Br Med J* 1998: **136**: 1236–8.
30. Weir BS, Cockerham CC. Estimating F-Statistics for the analysis of population structure. *Evolution* 1984: **38**: 1358–70.
31. Horton R, Gibson R, Coggill P et al. Variation analysis and gene annotation of eight *MHC* haplotypes: the *MHC* haplotype project. *Immunogenetics* 2008: **60**: 1–18.
32. Dawkins R, Leelayuwat C, Gaudieri S et al. Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease. *Immunol Rev* 1999: **167**: 275–304.
33. Ahmad T, Neville M, Marshall SE et al. Haplotype-specific linkage disequilibrium patterns define the genetic topography of the human *MHC*. *Hum Mol Genet* 2003: **12**: 647–56.
34. Raymond CK, Kas A, Paddock M et al. Ancient haplotypes of the *HLA* class II region. *Genome Res* 2005: **15**: 1250–7.
35. Imanishi T, Akaza T, Kimura A, Tokunaga K, Gojobori T. Allele and haplotype frequencies for *HLA* and complement loci in various ethnic groups. In: Tsuji K, Aizawa M, Sasazuki T, eds. *HLA 1991, Vol I. Oxford, UK: Oxford University Press*, 1992, 1065–204.
36. Khitrinskaya I, Stepanov VA, Puzyrev VP. *Alu* repeats in the human genome. *Mol Biol* 2003: **37**: 325–33.
37. Bharadwaj U, Khan F, Srivastava S, Goel H, Agrawal S. Phylogenetic application of *HLA* class II loci. *Int J Hum Genet* 2007: **7**: 123–31.

Supporting Information

The following supporting information is available for this article:

Figure S1. The amplification products of the five *POALIN*-PCR assays run on 2% agarose gels and stained with ethidium bromide. The lanes labelled 22–42, COX and PGF in the electrophorograms correspond exactly to the amplified PCR products of the cell-line DNA samples that are listed in Table S1, *Supporting Information*, with the Lab Numbers 22–42, and the cell-line names COX (#66) and PGF (#44), respectively. The lanes labelled 'M' represent the molecular markers.

Table S1. The *HLA* class I and class II alleles and *POALIN* insertion/deletion alleles in 67 reference cell-line DNA samples derived from different ethnic groups.

Table S2. *HLA-DRB1* and *HLA-DQB1* four-digit genotypes and *POALIN* two-digit genotypes in Japanese and Caucasians.

Table S3. Population differentiation at seven *MHC* class II loci.

Table S4. The frequency, percentage and descriptive level of association of *POALIN*s at five loci with *HLA-DRB1* two-digit alleles.

Table S5. The frequency, percentage and descriptive level of association of *POALIN*s at five loci with *DQB1* two-digit alleles.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Immunogenetics (2010) 62:211–230
DOI 10.1007/s00251-010-0427-2

ORIGINAL PAPER

Polymorphic SVA retrotransposons at four loci and their association with classical HLA class I alleles in Japanese, Caucasians and African Americans

Jerzy K. Kulski · Atsuko Shigenari · Hidetoshi Inoko

Received: 29 October 2009 / Accepted: 1 February 2010 / Published online: 20 February 2010

Abstract Polymorphic insertion frequencies of the retrotransposons known as the "SVA" elements were investigated at four loci in the *MHC* class I genomic region to determine their allele and haplotype frequencies and associations with the *HLA-A*, *-B* or *-C* genes for 100 Japanese, 100 African Americans, 174 Australian Caucasians and 66 reference cell lines obtained from different ethnic groups. The SVA insertions representing different subfamily members varied in frequency between none for *SVA-HF* in Japanese and 65% for *SVA-HB* in Caucasians or African Americans with significant differences in frequencies between the three populations at least at three loci. The SVA loci were in Hardy–Weinberg equilibrium except for the *SVA-HA* locus which deviated significantly in African Americans and Caucasians possibly because of a genomic deletion of this locus in individuals with the *HLA-A*24* allele. Strong linkage disequilibria and high percentage associations between the human leucocyte antigen (*HLA*) class I gene alleles and some of the SVA insertions were detected in all three populations in spite of significant frequency differences for the SVA and *HLA* class I alleles between the three populations. The highest percentage associations (>86%) were between *SVA-HB* and *HLA-B*08*, *-B*27*, *-B*37* to *-B*41*, *-B*52* and *-B*53*; *SVA-HC* and *HLA-B*07*; *SVA-HA* and *HLA-A*03*, *-A*11* and *-A*30*; and *SVA-HF* and *HLA-A*03* and *HLA-B*47*. From pairwise associations in the three populations and the homozygous cell line results, it was possible to deduce the SVA and *HLA* class I allelic combinations (haplotypes), population differences and the identity by descent of several common *HLA-A* allelic lineages.

Keywords SVA · HLA class I alleles · Dimorphism · Haplotype · Major histocompatibility complex · Retrotransposon · Retroelement

Introduction

The human *major histocompatibility complex* (*MHC*) class I region is located on chromosome 6 (*6p21.3*) and encodes at least 130 protein-coding genes including the classical and non-classical *human leucocyte antigen* (*HLA*) class I genes that are important in the regulation of the immune response system (Shiina et al. 2009). The identification of gene polymorphisms and diversity in the *MHC* region is a focus of attention for transplant donor and recipient genotyping and many different population, evolution and disease studies (Marsh et al. 2000). A large number of immune diseases have been associated with *HLA* alleles, although most association studies of the *MHC* in autoimmune and inflammatory disease have been limited to a subset of ~20 genes and performed only in small cohorts (Fernando et al. 2008). Many disease associations may result from

Electronic supplementary material The online version of this article (doi:10.1007/s00251-010-0427-2) contains supplementary material, which is available to authorized users.

J. K. Kulski
Centre for Forensic Science, The University of Western Australia, Nedlands, Western Australia 6008, Australia

J. K. Kulski · A. Shigenari · H. Inoko
Division of Molecular Life Science,
Department of Genetic Information, School of Medicine,
Tokai University,
Isehara, Kanagawa, Japan

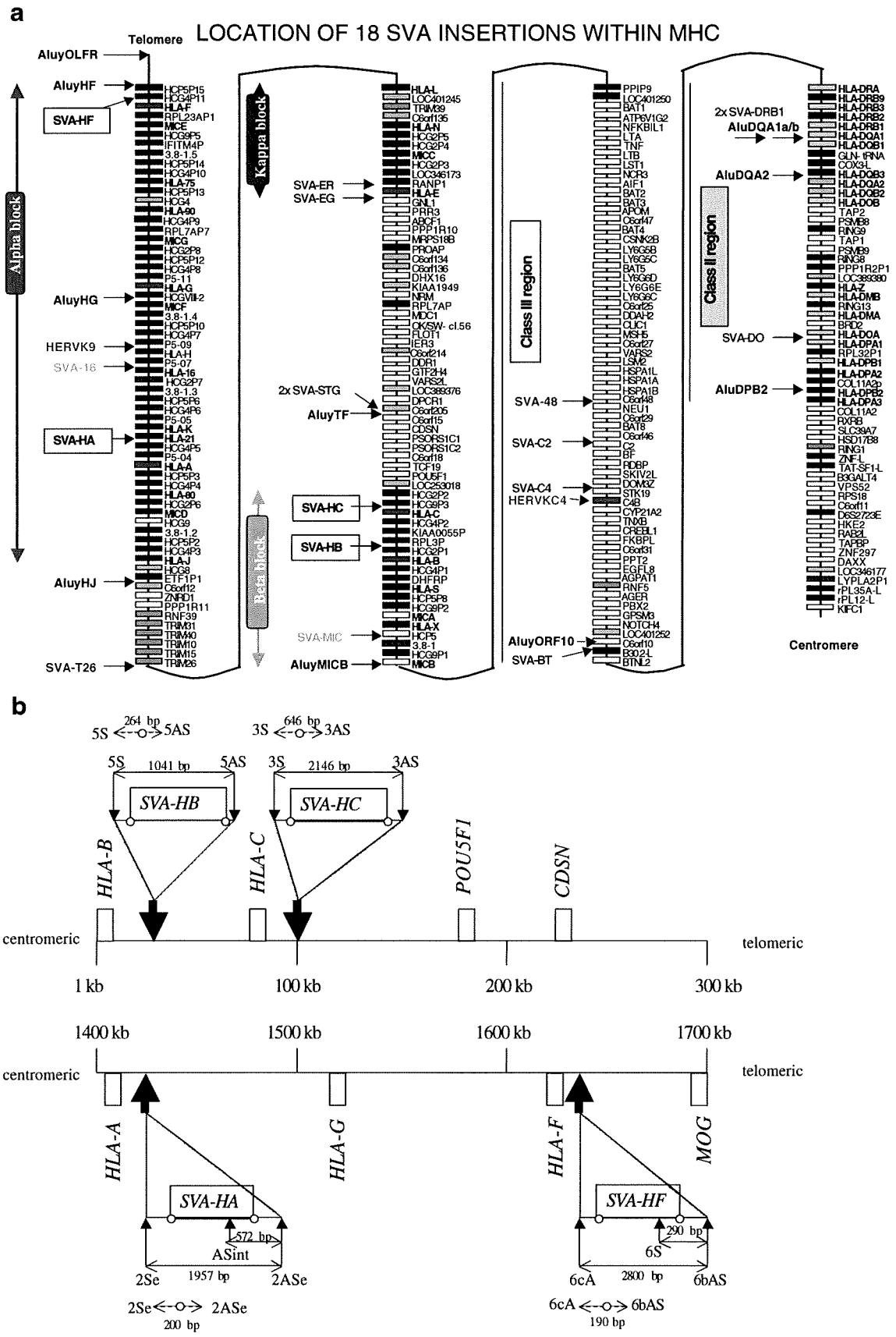
J. K. Kulski (✉)
Centre for Forensic Science, The University of Western Australia,
Mailbag M420. 35 Stirling Highway,
Crawley, Western Australia 6009, Australia
e-mail: kulski@me.com

haplotypes (allele combinations at multiple loci) and the evolutionarily neutral 'hitchhiking effect' of a deleterious allele linked to a beneficial allele (Shiina et al. 2006). Some haplotypes that had been defined solely on the basis of matched *HLA* alleles were later shown to be different when they were also typed for non-HLA alleles including polymorphic microsatellites and retrotransposons (Dunn et al. 2003, 2005; Witt et al. 2000). In this regard, an examination of the relationship between extragenic sequence markers and *HLA* allelic combinations might lead to a better understanding of the genetic diversity and the strength of conserved blocks within the *MHC* class I haplotypes and help identify potential crossing-over (recombination) sites between the *HLA* class I loci. For example, multiple forms and differences of the Japanese *HLA-B*48* haplotype (Dunn et al. 2003) and the European *HLA-B*07* and *-B*57* haplotypes were identified simply on the basis of a few polymorphic Alu insertions (Dunn et al. 2005). The analysis of the presence and absence of polymorphic retrotransposon structures, such as the different members of the Alu Y subfamily, in combination with *HLA* genotyping appear to better resolve class I haplotypic blocks when characterising ethnic and geographic population lineages (Dunn et al. 2007; Tian et al. 2008; Yao et al. 2009) and mapping the genetic association between the *MHC* and disease (Dunn et al. 2006). Of the various retrotransposon families and family members such as the long interspersed nuclear elements (LINEs with the members L1 and L2), the short interspersed nuclear elements (SINEs with the Alu member so called because it was recognised by the restriction enzyme *AluI*), the "SVA" element, so called because of its SINE, variable nucleotide tandem repeat region (VNTR) and Alu components, and the human endogenous retroviruses (HERVs), only five polymorphic Alu loci (Kulski and Dunn 2005) and a polymorphic HERVK9 locus (Kulski et al. 2008) and its long terminal repeat (LTR) sequence MER9 (Kulski et al. 2009) have been investigated in any detail with respect to their genetic diversity within the *HLA* class I region of different populations.

Recent comparative genomic analysis of different *MHC* haplotypes has confirmed that the retrotransposon, SVA, is widely distributed in the *MHC* genomic region and often structurally polymorphic (absent or present) at different loci (Stewart et al. 2004). The human SVA DNA element is a composite retrotransposon composed of two previously identified elements, SINE-R (Ono et al. 1987), a VNTR, and an Alu (Shen et al. 1994). The SVA, first identified within the *RP* gene of the human *MHC* genomic region on chromosome 6 (Shen et al. 1994), was recognised to be a retrotransposon because its nucleotide sequence ended in a polyA tail and was flanked by target site duplication (TSD) sequences which are the duplication products of the genomic target site (TS) sequence involved in the insertion event (Ostertag et al. 2003). The original SINE-R study (Ono et al. 1987) and subsequent studies of the SVA sequences (Bennett et al. 2004; Wang et al. 2005) estimated 2,762–5,000 SVA copies within a human haplotype genome with 27% to 38% of these copies structurally polymorphic (absent or present at the TS insertion locus) and actively mobile. The SVA are non-autonomous retrotransposons, similar to processed pseudogenes (retropseudogenes) and Alu, in that they are dependent on the L1 reverse transcriptase activity and retrotransposon molecular machinery for their continued genomic proliferation in the human population (Garcia-Perez et al. 2007; Strichman-Almashanu et al. 2003). The ancestral SVA first originated in great apes 14 Mya and subsequently proliferated into six subgroups (A–F) based on their sequence divergence and evolutionary time estimates (Wang et al. 2005) with 80% of the SVA members possibly human-specific in origin (Bennett et al. 2004). Some of the structurally polymorphic SVA have been related as causative agents in disease (Ostertag et al. 2003), such as hereditary elliptocytosis (Hassoun et al. 1994) and dystonia (Makino et al. 2007), and in genomic deletion events such as the deletion of the *HLA-A* gene (Takasu et al. 2007).

There are at least 18 SVA retrotransposons in the human *MHC* genomic region (Fig. 1). Although some of these are known to be polymorphic structures (Bennett et al. 2004; Stewart et al. 2004; Horton et al. 2008), there have been no detailed studies of their population allele and/or haplotype frequencies and LD or associations with *HLA* alleles. Therefore, in this study, we developed PCR assays to

Fig. 1 **a** Location map for 18 SVA insertions within the *MHC* ▶ genomic region. The location of different SVA, polymorphic Alu and HERV sequences (labelled) are indicated by the *arrows*. The *MHC* gene map and location of the alpha, beta and kappa blocks are based on the information provided by Shiina et al. (2006). The polymorphic Alu locations are taken from Stewart et al. (2004) and Kulski and Dunn (2005). Information on the location of HERVK9 and HERKC4 was taken from Kulski et al. (2008) and Shen et al. (1994), respectively. RepeatMasker (http://www.repeatmasker.org/) was used to identify the locations of the 18 SVA using the HLA genomic sequences of the cell lines COX and PGF downloaded from The *MHC* Haplotype Project at http://www.sanger.ac.uk/HGP/Chr6/MHC/. **b** Summary map of the location of the four SVA insertions investigated in this study and the positions of the PCR primers (Table 2) relative to the SVA insertion target site (*TS*). The genomic distances (kb) start at the *HLA-B* gene and end at the *MOG* gene. The many genes located in the 1,100-kb region between the *CDSN* and *HLA-A* genes (see Fig. 1a) are not shown in this simplified map. The *unlabelled thick vertical arrows* represent the sites of the SVA insertions. The *labelled thin vertical arrows* indicate the relative locations of the PCR primers (listed in Table 2) inside and outside the SVA insertions (*labelled rectangles*). The *solid horizontal double arrows* indicate the size of the PCR products with the SVA insertion, and the *dashed horizontal arrows* indicate the size of the PCR products without the SVA insertion. The *shaded circles* represent the target sites for SVA insertions or the target site duplications that flank the SVA insertions as part of the mechanism for the insertion event


a
LOCATION OF 18 SVA INSERTIONS WITHIN MHC
Telomere
AluyOLFR
AluyHF
SVA-HF
Alpha block
AluyHG
HERVK9
SVA-16
SVA-HA
AluyHJ
SVA-T26
Kappa block
SVA-ER
SVA-EG
2x SVA-STG
AluyTF
SVA-HC
SVA-HB
Beta block
SVA-MIC
AluyMICB
Class III region
SVA-48
SVA-C2
SVA-C4
HERVKC4
AluyORF10
SVA-BT
2x SVA-DRB1
AluDQA1a/b
AluDQA2
Class II region
SVA-DO
AluDPB2
Centromere
b
264 bp
5S
5AS
646 bp
3S
3AS
1041 bp
2146 bp
SVA-HB
SVA-HC
HLA-B
HLA-C
POU5F1
CDSN
centromeric
telomeric
1 kb
100 kb
200 kb
300 kb
1400 kb
1500 kb
1600 kb
1700 kb
HLA-A
HLA-G
HLA-F
MOG
SVA-HA
SVA-HF
572 bp
ASint
2Se
1957 bp
2ASe
200 bp
290 bp
6S
6cA
2800 bp
6bAS
190 bp

detect four SVA polymorphic structures within the *MHC* class I region and examined their allele and haplotype frequencies, LD and percentage association with *HLA-A*, *HLA-B* and *HLA-C* alleles in Japanese, African Americans and Australian Caucasians.

Materials and methods

DNA samples

A reference set of 100 Japanese (J) DNA samples genotyped at the *HLA-A*, *-B* and *-C* loci by DNA sequencing was obtained from the Department of Legal Medicine, Shinshu University School of Medicine, Matsumoto, Nagano, Japan. This reference set of DNA samples represents a Japanese population of registered donors from the Nagano region in the Japanese unrelated bone marrow donor registry (Moriyama et al. 2006). A reference set of 174 Australian-Caucasian (AC or C) DNA samples genotyped for *HLA* alleles at the *HLA-A*, *-B* and *-C* class I gene loci by DNA sequencing was obtained from the Department of Clinical Immunology and Biochemical Genetics, Royal Perth Hospital, Perth, Western Australia (Kulski et al. 2008). This reference set of samples represents a Caucasian population from the seaside town of Busselton in Western Australia (http://www.busseltonhealthstudy.com/). A panel of 100 African-American (AA or A) DNA samples was purchased from Coriell Cell Repositories as Human Variation panel HD100AA (http://ccr.coriell.org/nigms/nigms_cgi/panel.cgi?id=2&query=HD100AA). The African-American DNA samples represent a multiracial population with mostly African and European ancestors and a more complex genetic structure than the Europeans (Caucasians) or Japanese. Another 66 DNA samples extracted from B-lymphoblastoid cell lines of different ethnic origins and genotyped and/or serotyped for *HLA* alleles at the *HLA-A*, *-B* and *-DR* loci (Table 1) were purchased from the European Collection of Cell Cultures (http://www.ecacc.org.uk/). Information about these cell lines can be obtained at http://www.ebi.ac.uk/imgt/hla/help/cell_help.html.

HLA genotyping and nomenclature

The Japanese and Australian-Caucasian DNA samples were previously genotyped for *HLA-A*, *-B* and *-C* alleles to two or four digits by direct sequencing (Kulski et al. 2008; Moriyama et al. 2006). The African-American DNA samples were genotyped for *HLA-A* and *-B* alleles to two or four digits by the PCR–SSOP–Luminex method as previously described (Itoh et al. 2006).

The *HLA* alleles are reported and analysed here statistically as two- or four-digit alleles where the first two digits of an allele such as *02* in *HLA-A*0201* (Table 1) represent the ancestral group or type of highly related alleles and often correspond to the serological antigen carried by an allotype. The third and fourth digits such as the *01* in *A*0201* describe the subtype that has been assigned in the order of the determined DNA sequences and represent differences in one or more nucleotide substitutions that have changed the amino acid sequence of the encoded protein (Marsh 2000).

Location of SVA within the *HLA* genomic region

The map positions of the four SVA, *SVA-HB*, *SVA-HC*, *SVA-HA* and *SVA-HF* distributed across the *HLA* class I genomic region and investigated in this study are shown in Fig. 1. The locations of the SVA within the accession numbers previously reported by Stewart et al. (2004) are listed in Table 2. RepeatMasker (http://www.repeatmasker.org/) was used to identify the locations of the SVA, Alu and HERVs in Fig. 1a using the *HLA* genomic sequences of the cell lines COX and PGF downloaded from The *MHC* Haplotype Project at http://www.sanger.ac.uk/HGP/Chr6/MHC/.

Figure 1b shows a summary map of the location of the four SVA insertions and the positions of the PCR primers (Table 2) relative to the SVA insertion TS. All the PCR primers are placed outside the site of integration for each SVA, except for the primer 6S which is inside the *SVA-HF* sequence and the primer ASint which is inside the *SVA-HA* sequence. The DNA nucleotide sequences of the PCR products for the four SVA insertions with the relative positions of the PCR primers inside and outside of the SVA sequence, the SVA sequence and the TSD flanking the SVA insertion are shown in Electronic supplementary materials (ESM), Fig. S1. These nucleotide sequences are each a single example of the SVA insertion loci from the few that were identified within the NCBI GenBank database using the primer sequences in a BLAST search.

SVA PCR analysis

Table 2 shows the PCR primer nucleotide sequences, amplicon product sizes and the annealing temperatures and cycle times that were used for the amplification of the presence or absence of the *SVA-HF*, *SVA-HA*, *SVA-HC* and *SVA-HB* insertions. PCR assays were designed mostly to detect the presence and absence of the SVA insertion in a single assay by employing sense and antisense primers that flanked the insertion site. Because the insertion product size of the SVA sequence may be beyond the amplification efficiency of normal PCR protocols, we also developed a single PCR assay to detect only the presence of the SVA insertion by amplifying a fragment of the SVA sequence

Table 1 SVA insertions at four loci in homozygous and heterozygous cell lines

Lab no.	Cell line name	IHW no.	Ethnic origin	*HLA-A**	*HLA-B**	*HLA-DRB1**	*SVA-HA*	*SVA-HF*	*SVA-HC*	*SVA-HB*
1	WAL, FD	9129	Caucasoid	*03*	*07*	*1501*	+	+	+	−
2	HO104	9082	French	*03*	*07*	*?*	+	+	+	−
3	SCHU	9013	French	*0301*	*0702*	*1501*	+	+	+	−
4	EA	9081	Scandinavian	*0301*	*0702*	*1501*	+	+	+	−
5	WT100BIS	9006	Italian	*1101*	*3501*	*0101*	+	−	−	+
6	LBF (LBUF)	9048	Caucasoid	*3001*	*1302*	*070101*	−	−	−	−
7	SPL SPACH	9101	Sth American Indian	*3101*	*1501*	*8021*	−	−	−	−
8	LWAGS	9079	Ashkenasi Jewish	*3301*	*1402*	*0102*	−	−	−	−
9	WON, PY	9156	Oriental	*33*	*58*	*0301*	−	−	−	+
10	HAU, ML	9157	Oriental	*33*	*58*	*0301*	−	−	−	+
11	YAR	9026	Ashkenasi Jewish	*2601*	*3801*	*0402*	−	−	−	+
12	IBW9	9049	Sardinian	*3301*	*1402*	*0701*	−	−	−	−
13	WATANABE	9126	Oriental	*02*	*46*	*8032*	−	−	−	−
14	SPO010 SPO	9036	Italian	*0201*	*4402*	*1101*	−	−	−	−
15	AWELLS_WEL	9090	Australian Caucasoid	*0201*	*4402*	*0401*	−	−	−	−
16	EK	9054	Scandinavian	*0201*	*4402*	*1401*	−	−	−	−
17	BM16	9038	Italian	*0201*	*1801*	*1201*	−	−	−	−
18	EJ32B	9085	Australian Caucasoid	*3002*	*1801*	*03*	+	−	−	−
19	BSM	9032	Dutch	*0201*	*1501*	*04*	−	−	−	+
20	BOLETH BO	9031	Swedish	*0201*	*1501*	*0401*	−	−	−	+
21	WT9 (31227ABO)	9061	Italian	*0201*	*1801*	*1401*	−	−	−	−
22	KOSE	9056	German	*0201*	*3503*	*1302 1401*	−	−	−	+
23	BER	9093	German	*0201*	*1302*	*0701*	−	−	−	+
24	E4181324	9011	Australian Caucasoid	*0101*	*52011*	*15021*	−	−	−	+
25	TAB089	9066	Japanese	*0207*	*4601*	*8031*	−	−	−	−
26	J0528239	9041	Italian	*0101*	*3502*	*1104*	−	−	−	+
27	HAM 013	9178	Sth African Caucasoid	*01*	*08*	*1201 0301*	−	−	−	+
28	VAVY	9023	French	*0101*	*0801*	*0301*	−	−	−	+
29	LO541265	9086	Australian Caucasoid	*0101*	*0801*	*0301*	−	−	Hetero	+
30	PF04015	9088	French	*0101*	*0801*	*0301*	−	−	−	+
31	TISI-PMA	9042	–	*0101*	*57*	*07*	−	−	−	+
32	SA	9001	Japanese	*2402*	*0702*	*01*	No PCR prod	−	+	−
33	HOSONUM	9130	Oriental	*24*	*07*	*0101*	No PCR prod	−	+	−
34	KUROIWA	9131	Oriental	*24*	*07*	*0101*	No PCR prod	−	Hetero	−
35	HAY, KJ	9196	Australian Aborigine	*02*	*15*	*1301 14*	−	−	−	Hetero

Table 1 (continued)

Lab no.	Cell line name	IHW no.	Ethnic origin	*HLA-A**	*HLA-B**	*HLA-DRB1**	*SVA-HA*	*SVA-HF*	*SVA-HC*	*SVA-HB*
36	COX	9022	Sth African Caucasoid	*0101*	*0801*	*0301*	−	−	−	+
37	WBD001816	9154	–	*01*	*17*	*07*	−	−	−	−
38	DBB	9052	Amish	*0201*	*5701*	*0701*	−	−	−	−
39	MOU MANN M	9050	Danish	*2902*	*44031*	*0701*	−	−	−	−
40	PLH	9047	Scandinavian	*0301*	*4701*	*0701*	+	+	−	−
41	PGF	9318	English	*0301*	*0702*	*1501*	+	+	+	−
42	APD	9291	–	*01*	*60*	*0402*	−	−	−	+
43	SSTO-PMA	9302	–	*31*	*15*	*08*	−	−	−	−
44	QBL	9020	Dutch	*2601*	*1801*	*0301*	−	−	−	−
45	AKIBA	9286	Japanese	*2402*	*5201*	*1502*	No PCR prod	−	−	+
46	LKT3	9107	Japanese	*2402*	*5401*	*0405*	No PCR prod	−	−	−
47	HID	9074	Japanese	*0201*	*4001 4006*	*09*	−	−	−	Hetero
48	WON, C	9195	Australian Aborigine	*34*	*40 15*	*1201 0803*	−	−	−	−
49	WON, I	9194	Australian Aborigine	*34*	*40 5601*	*0803 1405*	−	−	−	−
50	DRI, SM	9128	–	*03*	*07 35*	*0101 1501*	+	+	Hetero	Hetero
51	BUR,E	9118	Australian Aborigine	*02 24*	*5601*	*0412 1409*	−	−	−	−
52	REE, GD		Caucasoid	*01 24*	*08*	*0301*	−	−	−	+
53	L0081785	9018	Australian Caucasoid	*0301 2402*	*1801*	*0301*	Hetero	Hetero	−	−
54	EK-TOK	9354	Japanese	*2602 11*	*35 46*	*0405 1101*	+	−	−	Hetero
55	HS67	9277	Japanese	*24 11*	*4801 6701*	*16 08*	Hetero	−	Hetero	Hetero
56	KOZ	9310	Japanese	*24 26*	*40 54*	*09*	−	−	−	−
57	LKT12	9073	Japanese	*2402 3101*	*3501 5201*		−	−	−	+
58	LKT14	9103	Japanese	*2402 2602*	*5101 4006*	*09*	−	−	−	−
59	LKT17	9024	Japanese	*0206 1101*	*1501 3501*	*0403 0406*	Hetero	−	−	+
60	COL, S	9197	Australian Aborigine/Caucasoid	*03 24*	*07 40*	*1501 1405*	Hetero	Hetero	Hetero	−
61	IHL, AD031	9117	Australian Aborigine	*02 31*	*27 40*	*0401 08032*	+	−	−	Hetero
62	IHL, AD036	9124	Australian Aborigine	*02 34*	*56 61*	*0803 1414*	−	−	−	−
63	NON, L	9192	Australian Aborigine	*01 10*	*51 55*	*0406 0803*	−	−	−	Hetero
64	BEA, PL	9138	–	*02 29*	*44 62*	*0401 07*	−	−	−	+
65	MAD, MF	9133	Caucasoid	*01 03*	*08 57*	*0301 07*	Hetero	Hetero	−	+
66	591	9230	Japanese	*1101 3101*	*39013 6701*	*12 15*	Hetero	−	Hetero	+

No PCR prod no PCR product, *Hetero* heterozygous for the presence and absence of the SVA insertion, +: homozygous presence of the SVA insertion, −: homozygous absence of the SVA insertion

Table 2 Primer sequences, amplicon product sizes, PCR conditions, subfamily type and estimated evolutionary age for four MHC class I polymorphic SVA structures

SVA name	Lab. code no.	Primer name	Primer type[a]	Primer sequence 5′-3′	Primer length	Sense+Antisense		PCR Conditions			SVA Subfamily[b]	Evolutionary Age[b] (Mya)	SVA genomic location[c]			SVA target site duplication (TSD) sequence[d]
						Fragment	Size (bp)	Anneal. temp (°C)	Taq Enzyme	Time (min)			Cell line	GeneBank accession no.	Location	
						Presence	Absence	35 cycles								
SVA-HF	6c	6S	int. sense	5′-ACTCCCTAATCTTA AGTACCCAG-3′	23	290		60	Taq Gold[e]	1	SVA_E	3.46	PGF	AL645939	14825-17414	5′-GAAAGACCCA AGCC-3′
		6cA	ex. sense	5′-GTCATTTGGTTTTA AGAGGTAAGAGG-3′	26		190									
		6bAS	ex. antisense	5′-GAATGACCAAGGT TACACGTCTATC-3′	25											
SVA-HA	2a	2Se	ex.sense	5′-CTGTGATAACCC AGAGTATCAGT-3′	23	1957	200	60	LA Taq[f]	3	SVA_D	9.55	PGF	AL671277	9466-11209	5′-GAATTGAGGA GC-3′
		2ASe	ex. antisense	5′-TAGGGATATGGATA CTTCCTCAG-3′	23											
	2c	ASint	int. sense	5′-GTACCCAACAG CTCATTGAGAAC-3′	23	572			Taq Gold							
		2ASe	ex. antisense	5′-TAGGGATATGGATA CTTCCTCAG-3′	23											
SVA-HC	3a	3S	ex. sense	5′-CAATGTTGCAGTCT TCAGTCTA-3′	22	2146	646	65	LA Taq	5	SVA_F	3.18	PGF	AL662844	183342-184827	5′-TTCTGCCTC-3′
		3AS	ex. antisense	5′-ATGTACCCATAC A[C/T]GTACTAAGC-3′	23											
SVA-HB	5	5S	ex. sense	5′-TGCTAGTATCATGT CTAGTGCTG-3′	23	1041	264	60	LA Taq	5	SVA_B	11.56	COX	AL845556	20980-21762	5′-AATTAAT-3′
		5AS	ex. antisense	5′-CTACTCAGGAGAG TCACTTCAAC-3′	23											

[a] int.sense is the sense primer sequence which is internal or within the SVA sequence, whereas ex. sense or ex. antisense is the sense or antisense primer sequences that is exterior to or outside the SVA sequence

[b] Data taken from Wang et al. (2005)

[c] Data taken from Stewart et al. (2004)

[d] The TSD is the SVA target site for the SVA integration and this sequence flanks the SVA sequence after the integration event (Ostertag et al. 2003). The TSD and SVA sequence within the PCR amplification product is shown in ESM Fig. S1

[e] The Taq Gold PCR performed using an initial denaturation step of 94 C for 5 min, then 35 cycles with the above annealing temperatures for 1 min after a denaturation step at 94°C for 1 min and a final extension step at 72°C for 1 min

[f] The LA PCR performed using an initial denaturation step of 94°C for 5 min, then 35 cycles with the above annealing temperatures after a denaturation step at 94°C for 30 s

using an SVA internal and external primer set for *SVA-HA* and *SVA-HF*.

Each SVA PCR assay was performed in 10 μl aliquots using 2 pmol of each primer (200 nmol/l), 1 ng of genomic DNA, 0.25 U of LA *Taq* polymerase (TaKaRa, Shiga, Japan) or Amp Taq gold (Applied Biosystem), 0.8 μl of dNTP mixture (2.5 mM each) and 5 μl of 2× GC reaction buffer 1 with 5 mM $MgCl_2$ purchased from TaKaRa. The PCR was performed in eight strips of 0.2-ml thin-walled PCR tubes (QSP) using a GeneAmp PCR System 9700 Thermal cycler programmed for 35 cycles with a denaturation (96°C, 30 s), annealing and extension step at each cycle (see Table 2 for annealing temperatures and times). The reaction products were stained with ethidium bromide and the sizes compared with molecular size markers by horizontal gel electrophoresis in 2% agarose using Tris-borate-EDTA running buffer.

Figure 2 shows an example of the results of the electrophoresis of the PCR products amplified by the SVA PCR assays of the DNAs extracted from the typing cell line Lab. no. 1 to 21 (Table 1) including the COX (Lab. no. 36) and PGF (Lab. no. 41) control DNA. The amplification bands were easily visualised, genotyped and scored as the presence (+) or absence (−) of an insertion both as either homozygotes (Fig. 2) or heterozygotes (not shown). The result in lane 6 of Fig. 2 for the *SVA-HA* assay (a) was equivocal, but it was later resolved to be a negative after triplicate repeats (data not shown) and by the *SVA-HA* assay (b) using the primer set 2c. Lane 6 in the *SVA-HB* assay (e) was confirmed to be a negative result (no SVA insertion) after triplicate repeats. The 6c primer set (Table 2) permitted the detection of the *SVA-HF* insertion (290 bp) and its absence (190 bp) in a single PCR. The *SVA-HA* insertion was detected using two separate PCR assays. One PCR assay used a set of primers (2a) that detected the presence (1,957 bp) and the absence (200 bp) of the SVA insertion in a single assay, whereas the other PCR assay (primer set 2c) was used only to confirm the presence of the SVA insertion (572 bp). There was 95% agreement between the two PCR assays for the insertion of *SVA-HA*, but where there was disagreement, the positive sample was assumed to be

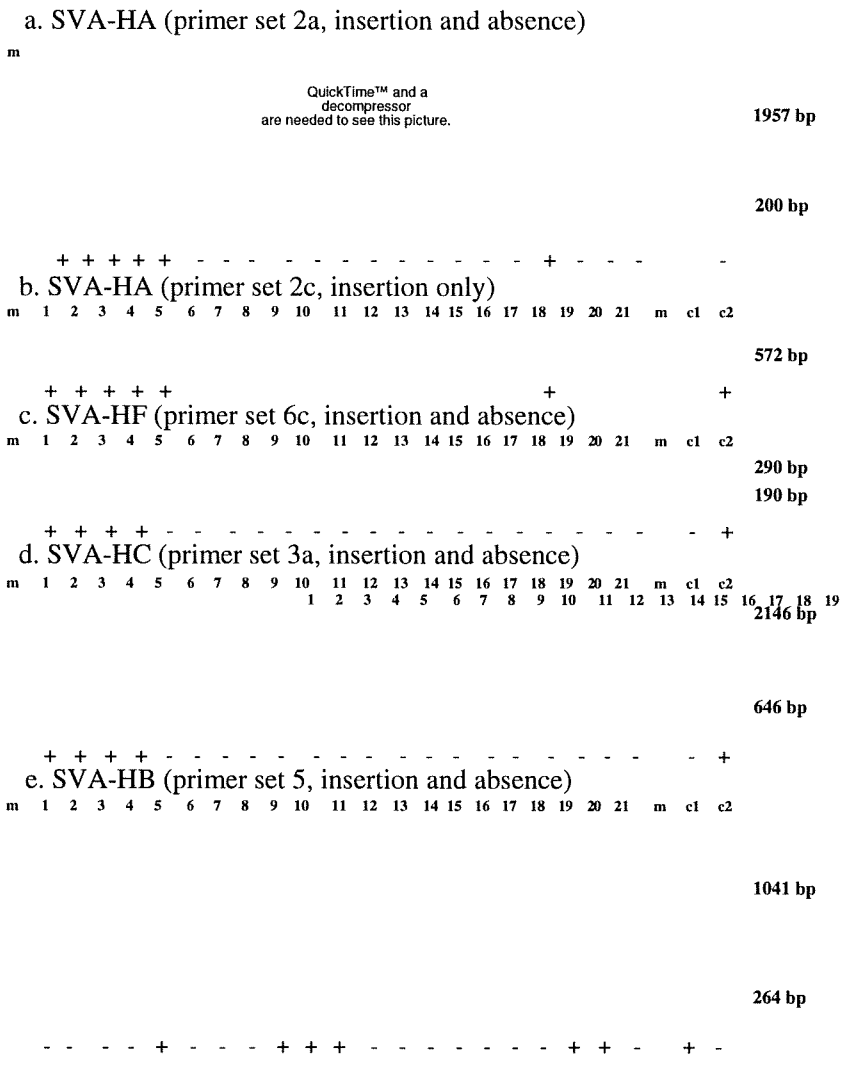


Fig. 2 Electropherograms of the amplification products of the SVA PCR assays (**a**–**e**). The molecular sizes on the right-hand side of the gels are the sizes of the stained PCR products. *Lanes m* molecular weight markers, *lanes c1* and *c2* amplification products from the cells PGF and COX, respectively, in the *SVA-HA* assay (**a**). c1 and c2 are amplification products from the cells COX and PGF, respectively, in the *SVA-HA* assay (**b**) to the *SVA-HB* assay (**e**). *Lanes 1–21* PCR products from the Lab nos. 1 to 21 in Table 1

correct. Only a single PCR assay was used for the detection of the *SVA-HC* and *SVA-HB* structural alleles, respectively, after initial testing for a number of different primer sets using typing cell lines, including COX and PGF which were previously known to have one or other of the SVA insertions.

SVA allele, genotype and diplotype designations

For data analysis, the allele with no SVA insertion (absent) at a locus was designated with the number 1, and the allele with the SVA insertion (present) at a locus was designated with the number 2. The number 0 was given to a reproducible failed PCR reaction to indicate that the failure was possibly due to a deletion or mutation at one or both primer sites. Genotypes at a particular locus were designated the allele numbers 11 (homozygous absence of the SVA insertion on both chromosomes), 22 (homozygous presence of the SVA insertion on both chromosomes) and 12 (heterozygous locus, absence or presence of the SVA insertion on one of the chromosomal pairs). Diplotypes were designated with a series of letters A, I or H where A is the homozygote genotype 11 with SVA absent at the same locus of each chromosome, I is the homozygote genotype 22 with SVA present at the same locus of each chromosome, and H is the heterozygous genotype 12 with one SVA present and one SVA absent at the same locus of each chromosome.

As with Alu deletions, a SVA deletion is expected to involve either the entire retrotransposon sequence together with one or both of the flanking genomic regions (Kulski et al. 1999) or a portion of the retrotransposon sequence whereby a fragment or signature of the original insertion event is retained (Edwards and Gibbs 1992). On the basis of our confirmatory sequencing analysis (data not shown) and the expected and observed PCR product sizes, the SVA allele designation of "absent" is unlikely to reflect a full-length deletion of a previously inserted SVA sequence located between the target site duplication sequences (ESM Table S1).

Statistical analyses

Gene and allele frequencies, heterozygosity, Hardy–Weinberg equilibrium (HWE) and LD pairwise tests (D' and r^2) of the association between the *HLA* class I gene two-digit alleles and the SVA alleles were performed using the online software programmes GENEPOP (Raymond and Rousset 1995) at http://genepop.curtin.edu.au/, LINKDOS (Garnier-Gere and Dillmann 1992) at http://genepop.curtin.edu.au/linkdos.html, DE FINETTI at http://ihg.gsf.de/cgi-bin/hw/hwa1.pl (Sasieni 1997) and also the downloaded programmes MIDAS (Gaunt et al. 2006) from http://www.oege.org/software/midas/index.shtml, Arlequin v3.1 (Excoffier et al. 2005) from http://cmpg.unibe.ch/software/arlequin3 and FSTAT v 2.9.3.2 (Goudet 1995) from http://www2.unil.ch/popgen/softwares/fstat.htm. Heterozygosity (H) was estimated as $2pq$, where p and q are the allele frequencies. A 2×2 contingency two-sided test with 2 df and Fisher's exact test in the DE FINETTI programme were used to detect the significant difference for SVA frequencies between populations.

The LD tests and LD values D' and r^2 and haplotype frequencies were calculated in a multi-allele pairwise analysis of the association between the *HLA* class I gene two-digit alleles and the SVA alleles using the default settings for MIDAS (Gaunt et al. 2006). The Arlequin haplotype frequencies were calculated for genotypic data with unknown gametic phase using expectation–maximisation (EM) algorithm and parameter settings of 50 for the number of starting values for EM, 100–500 for the initial conditions for bootstraps, 100–5,000 for the number of iterations and 1,000 for the number of bootstrap replicates. No statistical corrections were applied for multiple testing to maintain sensitivity (Perneger 1998). Unpaired t tests between the means were performed with GraphPad Software online at http://www.graphpad.com/quickcalcs/ttest1.cfm?Format=SD. The percentage association between an SVA insertion and *HLA* allele was calculated as the percentage of the total *HLA* allele frequency that was associated with the presence of the SVA insertion at an inferred haplotype using the observed haplotype frequency data generated by the Midas software.

Phylogeny of *HLA-A* alleles

HLA-A protein sequences representing different alleles (cDNA sequences) were obtained from the Anthony Nolan Bone Marrow Trust HLA database (http://www.anthonynolan.org.uk/HIG/seq/hla.html). The multiple protein sequences were aligned using the CLUSTAL W 1.8 programme, and the phylogenetic analysis was undertaken using the neighbour joining (NJ) method (Saitou and Nei 1986) at DDBJ (http://www.ddbj.nig.ac.jp/search/clustalw-e.html) with the default settings for "protein" type, p-correction, a phylip distance (output tree) tree and bootstrap analysis (1,000 counts, 111 seed). The phylip format (phb) of the phylogenetic tree was displayed using the NJplot programme (Perrière and Gouy 1996).

Results and discussion

HLA class I and SVA allelic and haplotype associations in reference cell lines

To test the reliability of the SVA PCR assays and examine the haplotype linkage between the *HLA-A*, *HLA-B* and

HLA-DRB1 and SVA alleles, we analysed a DNA reference set of 66 typing cell lines of various ethnic origins with 46 cells that were homozygous for *HLA-A* and *HLA-B* alleles and 26 that were heterozygous either for the *HLA-A* or *HLA-B* alleles. Table 1 lists the individual cell lines by Lab. number, name, IHW ID, ethnicity, *HLA-A*, *HLA-B* and *HLA-DRB1* alleles and the SVA insertion results at the four loci, *SVA-HF*, *SVA-HA*, *SVA-HC* and *SVA-HB*.

Of the two SVA insertion loci near the *HLA-A* gene locus, *SVA-HF* was linked to all ten samples with the homozygous or heterozygous *HLA-A*03* allele and in particular the five homozygous samples with the *HLA-A*03/HLA-B*07/DRB1*1501* haplotype. None of the other nine *HLA-A* alleles were linked to *SVA-HF*. The *SVA-HA* insertion was linked to all ten samples with the *HLA-A*03* allele, but also to the *HLA-A*11* and *HLA-A*3002* alleles. Interestingly, the *SVA-HA* PCR yielded no amplification products in the five samples with the homozygous *HLA-A*24* allele, suggesting that the *SVA-HA* locus is probably deleted in most *HLA-A*24* haplotypes. In the Caucasian population analysis, the *SVA-HA* PCR also failed to amplify the DNA sample from a single case of the *HLA-A*24/24* homozygote (data not shown). This finding is consistent with a previous report of a 50-kb genomic deletion upstream of the *HLA-A* gene with the *HLA-A*24* allele (Watanabe et al. 1997) and the deletion of the MER9-LTR in the *HLA-A*24*, which is only 4.2 kb from the *SVA-HA* locus in the other *HLA-A* alleles (Kulski et al. 2009).

Of the two SVA insertion loci near the *HLA-B* gene locus, the *SVA-HC* insertion was found in all eight homozygous samples with the *HLA-B*07* allele and the two heterozygous *HLA-B*67* samples, but was not linked to the other *HLA-B* alleles. In contrast, the *SVA-HB* insertion was present in 24 of the 46 homozygous samples and linked to a variety of *HLA-B* alleles including *HLA-B*0801*, *-B*1302*, *-B*1402*, *-B*1501*, *-B*17*, *-B*3501*, *-B*3502*, *-B*3508*, *-B*3801*, *-B*4001*, *-B*4403*, *-B*4701*, *-B*5201*, *-B*5701*, *-B*58*, and *-B*60*. The *SVA-HB* insertion was absent from samples with *HLA-B*07*, *-B*1402*, *-B*1801*, *-B*4006*, *-B*4402* and *-B*46*.

HWE, allele and diplotype frequencies of SVA insertions in three populations

PCR for the detection of the presence and/or absence of the SVA insertion at four loci was performed on three distinct populations. In the 174 Australian Caucasians, the *SVA-HB* PCR failed to amplify in a HLA-*B*46/48* heterozygote, whereas in the 100 African Americans, the *SVA-HB* PCR had failed to amplify in four cases, a *HLA-B*39/53*, *-B*35/68*, *-B*40/40* and one case with an unknown genotype. Similarly, the *SVA-HC* PCR failed to amplify using the DNA from 5 of 100 African Americans, *HLA-B*02/11*, *-B*15/51* and three cases with unknown *HLA-B* genotypes. These PCR assays may have failed in part because nucleotide mutations (SNPs and/or indels) at the primer binding sites within the template DNA prevented hybridisation between primer(s) and target DNA templates. The *SVA-HB* and *SVA-HC* genomic regions within the *MHC* have extensive nucleotide variability and are possibly susceptible to a rapid mutation rate (Gaudieri et al. 2000). These negative samples were excluded from the calculation of HWE and the frequency of the SVA insertions, but future assays may benefit from the inclusion of sequencing analysis and/or internal PCR controls to better assess the reasons for the failed PCR reactions.

The frequencies and the results of the HWE test for the SVA genotypes in the *MHC* class I region of 100 Japanese, 100 African-American, and 174 Australian-Caucasian DNA samples are shown in Table 3. The SVA genotypes were in HWE at three of the four loci. The main exception was for *SVA-HA* when not corrected for its deletion on the *HLA-A*24* haplotype. If the *SVA-HA* on the *HLA-A*24* haplotype was counted as absent rather than deleted, then the number of observed heterozygotes was increased and the number of observed homozygotes for the *SVA-HA* insertion was decreased, resulting in HWE at the *P* value of 0.5244.

The statistical differences and *P* values between the three populations for the allele and the genotypic insertion frequency, heterozygosity and homozygosity for the four SVA loci are shown in Table 4. The highest SVA insertion frequency (range, 0.25–0.65) was for *SVA-HB* in all three populations (Table 3) with a significant difference ($P<0.05$) between the Japanese and African Americans or Australian Caucasians, but not between the African Americans and Australian Caucasians ($P>0.05$, Table 4). However, there was a significant difference ($P=0.026$) in the heterozygous frequency between African Americans and Australian Caucasians. The lowest SVA insertion frequencies (<0.15) were for *SVA-HC* and *SVA-HF* (Table 3) with significant differences ($P<0.05$) between the Australian Caucasians and the African Americans or Japanese, but not between the African Americans and Japanese for the frequency of the *SVA-HC* insertion (Table 4). The Japanese had no detectable *SVA-HF* insertions. The highest *SVA-HA* frequency was 0.23 in Australian Caucasians when the deletion of the *SVA-HA* locus was taken into account for the HWE calculations ($P=0.5244$) in individuals with the *HLA-A*24*. The Armitage's test for trend also shows some similarity ($P>0.05$) between the *SVA-HA* or *SVA-HC* insertions of Caucasians and African Americans and between the *SVA-HC* insertions of African Americans and Japanese (Table 4).

The frequency of the absence and/or presence of the SVA insertions at the four loci can be estimated as diplotype (a set of haplotype pairs or genotypes at multiple

Table 3 SVA insertion frequencies and HWE test results for Australian Caucasians, Japanese and African Americans

Ethnic group and SVA genotypes	*SVA loci*					
	SVA-HB	*SVA-HC*	*SVA-HF*	*SVA-HA*[a]	*SVA-HA*[b]	*SVA-HA*[c]
Australian Caucasians, sample no.	173	174	174	174	130	174
No. of observed homozygote insertion	70	3	3	20	11	11
No. of observed heterozygote	86	29	44	50	49	59
No. of observed homozygote deletion	17	142	127	104	80	104
No. of expected homozygote insertion	73.81	1.76	3.59	11.64	18.58	9.43
No. of expected heterozygote	78.38	31.48	42.82	66.72	64.84	62.15
No. of expected homozygote deletion	20.81	140.76	127.59	95.64	56.58	102.43
Insertion frequency	0.65	0.1	0.14	0.26	0.36	0.23
SD of insertion frequency	0.024	0.017	0.019	0.026	0.025	0.023
HWE test P exact	0.24	0.39	1	0.001366	0.00653	0.5244
Japanese, sample number	100	100	100	100		
No. of observed homozygote insertion	7	0	0	0		
No. of observed heterozygote	36	6	0	11		
No. of observed homozygote deletion	57	94	100	89		
No. of expected homozygote insertion	6.25	0.09	0	0.3		
No. of expected heterozygote	37.5	5.82	0	10.39		
No. of expected homozygote deletion	56.25	94.09	0	89.3		
Insertion frequency	0.25	0.03	0	0.06		
SD of insertion frequency	0.031	0.012	0	0.016		
HWE test P exact	0.79	1	0	1		
African American, sample no.	96	95	100	100	92	100
No. of observed homozygote insertion	43	0	0	7	6	6
No. of observed heterozygote	36	1	8	22	21	23
No. of observed homozygote deletion	17	94	92	71	65	71
No. of expected homozygote insertion	38.76	0	0.16	3.24	2.96	3.06
No. of expected heterozygote	44.48	0.99	7.68	29.52	27.08	28.88
No. of expected homozygote deletion	12.76	94	92.16	67.24	61.96	68.06
Insertion frequency	0.64	0.01	0.04	0.18	0.18	0.17
SD of insertion frequency	0.038	0.005	0.014	0.03	0.031	0.029
HWE test P exact	0.077	1	1	0.015529	0.036646	0.0725

http://ihg2.helmholtz-muenchen.de/cgi-bin/hw/hwa1.pl

[a] Includes samples with *HLA-A*24* allele

[b] *HLA-A*24* samples removed from analysis

[c] Includes samples with *HLA-A*24* allele, but all SVA positive samples associated with *HLA-A*24* allele were called heterozygote and not homozygotes

loci) frequencies by counting the number of individuals with homozygous insertions (I), heterozygotes (H) or the absence of SVA insertions as homozygous genotypes (A) at each locus. ESM Table S1 shows the diplotype percentage frequencies at the four SVA loci in the three populations with ten different diploids in Japanese, 17 in African Americans and 33 in Caucasians. The absence of SVA insertions as homozygous genotypes (AAAA) at the four loci was 50% in the Japanese in comparison to 9% in the African Americans and 2.3% in Australian Caucasians. The percentage frequency of the diplotype, IAAA (no. 25), was significantly higher ($P<0.05$) in the Australian Caucasians (23%) and African Americans (22.4%) than in the Japanese (5%). The African Americans (12%) had a significantly higher ($P<0.05$) percentage frequency of the IAHA diplotype (no. 27), which differentiated them from the Australia Caucasian (5.2%) and Japanese (1%). On the other hand, all three populations had 22.4% to 29% of the HAAA diplotype (no. 12). Some multiple SVA homologous insertions (diploid nos. 11, 18, 24, 29, 30, 31) were found in Caucasians and to a lesser extent in African Americans, but not in the Japanese.

Table 4 Test for association and SVA loci differentiation between populations

Population comparisons	Significance *P* level at four SVA insertion loci			
	SVA-HB	*SVA-HC*	*SVA-HA*	*SVA-HF*
Tests for association (TA) by Sasieni analysis[a]				
Allele frequency difference, AC v J	1.10E−19	0.0025	2.73E−09	1.87E−08
Allele frequency difference, AC v AA	0.67963	0.00002	0.0323	0.00015
Allele frequency difference, AA v J	1.51E−14	0.12327	0.0001	0.0077
Heterozygous, AC v J	7E−4/1E−10	0.43/0.009	0.04/8E−5	1/3E−8
Heterozygous, AC v AA	0.166/0.026	0.748/9E−5	0.65/0.13	0.46/0.0004
Heterozygous, AA v J	4E−5/7E−4	1/0.6	0.07/0.02	1/0.004
Homozygous, AC v J	3.04E−17	0.1605	0.00006	0.12624
Homozygous, AC v AA	0.21399	0.1605	0.13991	0.14237
Homozygous, AA v J	5.61E−12	1	0.00389	1
Armitage's trend test: AC v J	6.03E−18	0.00358	1.16E−07	3.09E−08
Armitage's trend test: AC v AA	0.68073	0.00006	0.05623	0.00014
Armitage's trend test: AA v J	6.68E−12	0.06342	0.0004	0.00389
Population differentiation (PD) by GenePop analysis[b]				
Genic differentiation, all populations[c]	>0.00000	>0.00000	>0.00000	>0.00000
Genic differentiation, AC v J	>0.00000	0.00135	>0.00000	>0.00000
Genic differentiation, AC v AA	7.79E−01	>0.00000	0.02757	0.00001
Genic differentiation, AA v J	>0.00000	0.12219	0.00004	0.00779
Genotypic differentiation, all[d]	>0.00000	>0.00000	>0.00000	>0.00000
Genotypic differentiation, AC v J	>0.00000	0.00251	>0.00000	>0.00000
Genotypic differentiation, AC v AA	7.82E−01	>0.00000	6.10E−02	>0.00000
Genotypic differentiation, AA v J	>0.00000	0.11869	0.00026	0.0061

An unbiased estimate of the P-value of a log-likelihood ratio (G) based exact test was performed.

AC Australian Caucasians, *J* Japanese, *AA* is African American

[a] TA used the software programme at http://ihg2.helmholtz-muenchen.de/cgi-bin/hw/hwa1.pl

[b] PD by GenePop Analysis using software at http://genepop.curtin.edu.au/

[c] Genic differentiation is the distribution of alleles in the various populations

[d] Genotypic differentiation is the distribution of diploid genotypes in the various populations

Four-locus SVA haplotype frequencies and population differences

ESM Table S1 shows the maximum-likelihood estimations of the four-locus SVA haplotype frequencies using the EM algorithm in the Arlequin software package and the *P* values of the population differences analysed by the unpaired Student's *t* test. There were 15 four-locus SVA haplotypes for the three populations with 7 in the Japanese, 9 in the African Americans and 14 in the Caucasians. There were another seven haplotypes with missing data at one or more haplotypes in the three populations. The SVA null haplotype (no SVA insertions) was the most common in Japanese at 70.9%, which was significantly different ($P<0.0001$) to the 24.6% in African Americans and 19.2% in Caucasians. The most frequent SVA haplotypes in Caucasians and African Americans was the single SVA-HB insertion at a frequency of 47.1% for both populations. The most frequent multiple SVA insertions were the SVA-HA/SVA-HB haplotype at 11.2% in African Americans, 9% in Caucasians and 2.8% in Japanese.

The percentage association, LD and haplotype frequencies of SVA alleles paired with *HLA* class I alleles in three populations

The LD as a D' or r^2 measure, the observed haplotype frequencies of SVA associations and the percentage association (percentage of the *HLA* class I allele total frequency associated with an SVA insertion) between SVA insertions and *HLA-A*, *-B* or *-C* alleles at paired loci in African Americans, Caucasians and Japanese were calculated using the Midas software (Gaunt et al. 2006) and are shown in the ESM Table S3 to Table S5, respectively. The LD pairwise allele analysis by LinkDos, GenePop (option2, suboption1) and Midas each revealed significant LD

($P<0.05$) between *SVA-HF* and *SVA-HA*, *SVA-HF* and *SVA-HC*, *SVA-HA* and *SVA-HC* and between *SVA-HC* and *SVA-HB* in Caucasians, but not in African Americans or Japanese. On the other hand, there was significant LD and a high percentage association between some of the SVA insertions and *HLA* class I alleles in the three populations.

The percentage frequency (>50%) of the *HLA* class I alleles associated with one of the four SVA insertions in the *SVA/HLA* class I haplotypes of each population is shown in Table 5. The moderate frequency SVA insertions *SVA-HA* and *SVA-HB* (Table 3) are in strong association (>50%) with both *HLA-A* and *HLA-B* alleles probably due to linkage disequilibria between *HLA-A* and *-B* and the formation of different haplotypes. For example, some of the strong associations between the SVA insertion and the *HLA-A* and *-B* allelic pairs in Table 5 can be seen linked together in the haplotypes of some of the cell lines listed in Table 1, such as the *SVA-HA* insertion in the haplotypes *HLA-A*1101/HLA-B*3501* (cell line 5) and *HLA-A*0301/HLA-B*4701* (cell line 40) and the *SVA-HB* insertion in the haplotypes *HLA-A*1101/HLA-B*3501* (cell line 5), *HLA-A*33/HLA-B*5801* (cell lines 9 and 10), *HLA-A*2601/HLA-B*3801* (cell line 11), *HLA-A*0201/HLA-B*1501* (cell lines 19 and 20), *HLA-A*0201/HLA-B*3503* (cell line 22), *HLA-A*0201/HLA-B*1302* (cell line 23), *HLA-A*0101/HLA-B*5201* (cell line 24) and *HLA-A*0101/HLA-B*0801* (cell lines 28 to 30). The *SVA-HA* insertion in the haplotype *HLA-A*0301/HLA-B*0701* (Table 5) had a 49% association with *HLA-B*07* and 95% association with *HLA-A*03* in Caucasians (ESM Table S4).

The *SVA-HF* insertion was most frequently (90.4%) associated with the *HLA-A*03* allele (Table 5) and the Caucasian *HLA* haplotype *A*0301/C*0702/B*0702/DRB1*1501/DQB*0602* (data not shown). Forty-seven (27%) of the 174 Caucasian individuals and none of the 100 Japanese individuals were positive for *SVA-HF*. Of the non-*HLA-A*03* alleles, ten *SVA-HF*-positive samples were associated with a variety of *HLA-A* alleles including one sample homozygous for *HLA-A*24*. This implies a crossing-over (recombination) rate of about 21% between the *HLA-A* and *SVA-HF* loci (220 kb, Fig. 1b) in the 41 Caucasians with the *HLA-A*03*. In African Americans, only 8 of 100 (8%) individuals were *SVA-HF*-positive, with four of these having the *HLA-A*03* allele. However, 7 out of the 11 individuals with the *HLA-A*03* allele were *SVA-HF*-negative. *SVA-HF* was also associated strongly (>60%) with the low-frequency *HLA-B* alleles, HLA-*B*39*, *-B*47* and *-B*52* (Table 5) as also seen in the cell line PLH (lab no. 40) with the *HLA-A*0301/HLA-B*4701* haplotype (Table 1).

The *SVA-HA* insertion, which has low to moderate frequencies in the three populations, was found to be relatively haplospecific with significant LD ($P<0.05$) and strong percentage associations (>79%) for *HLA-A*03*, *HLA-A*11* and *HLA-A*30* in Caucasians and African Americans, but not the Japanese. The percentage association between the *SVA-HA* insertion and *HLA-A*11* was strong in the Caucasians (100%), moderate in African Americans (57.2%) and weak in the Japanese (37.5%) and with significant LD ($P<0.05$) in all three populations. The identity by descent of the *SVA-HA* insertion with a phylogenetic reconstruction of the *HLA-A* alleles (Fig. 3) suggests that the *SVA-HA* insertion in Caucasians and African Americans has been inherited mainly through the *HLA-A*03* and *HLA-A*30* lineages after their separation from the *HLA-A*01* lineage. In Caucasians and African Americans, the *SVA-HA* insertion has strong association with the HERVK9 insertion, which is located between *SVA-HA* and *SVA-HF* (Fig. 1a) and also has a strong association with the *HLA-A*03* and *-A*30* alleles (Kulski et al. 2008).

Strong SVA insertion associations of more than 80% with HLA alleles were observed between *SVA-HB* and *HLA-B*08*, *-B*13*, *-B*15*, *-B*27*, *-B*35*, *-B*37* to *-B*41*, *-B*45* and *-B*50*; *SVA-HC* and *HLA-B*07*; *SVA-HA* and *HLA-A*03*, *-A*11*, *-A*30* and *-A*69*; *SVA-HF* and *HLA-A*03*; and *SVA-HB* and *HLA-A*01*, *-A*23*, *-A*25*, *-A*30* and *-A*69* in Caucasians. Similar trends were observed in Japanese and African Americans. The few *HLA-B* alleles that did not have the *SVA-HB* insertion were *HLA-B*07*, *-B*14*, *-B*4402* and *-B*18*. The *SVA-HB* insertion also had strong associations with a number of different *HLA-C* 2-digit alleles, such as *HLA-C*03*, *-C*04*, *-C*06*, *-C*12*, *-C*16* and *-C*17* in Caucasians and *HLA-C*02*, *-C*06* and *-C*12* in Japanese (Table 5).

The *SVA-HC* insertion is in significant LD ($P<0.05$) and has 86.4% and 31.3% association with *HLA-B*07* and *HLA-C*07*, respectively, in the Caucasian population (Table 5 and ESM Table S4) and was found in all the *HLA-B*07* homozygous cell lines (Table 1). Twenty-nine (90.6%) of 32 *SVA-HC*-positive Caucasian individuals also had the *HLA-B*0702* and *HLA-C*0702* alleles (data not shown). Only four (12%) of the 33 Caucasian individuals with the *HLA-B*0702* and *HLA-C*0702* allelic combination did not have an accompanying *SVA-HC* insertion. The *SVA-HC* insertion has significant ($P<0.05$) LD, but a weak percentage association with *HLA-B*07* in Japanese (21.7%) and African Americans (7.7%). Only 3 of the 13 (23.1%) individuals with the *HLA-B*0702* and *HLA-C*0702* alleles were positive for the *SVA-HC* insertion in the Japanese, and only one of 12 (8.3%) African Americans with the *HLA-B*07* allele also had the *SVA-HC* insertion (data not shown). The frequency of the *SVA-HC* insertion is three to ten times higher in the Caucasians than in Japanese or African Americans (Table 3). This suggests that *SVA-HC* was inserted originally into a member of the Caucasian population and is

Table 5 Percentage (>50%) association of the SVA insertions at four loci with particular *HLA* class I alleles in the SVA/*HLA* class I haplotype pairs of each population

SVA allele insertion	*HLA* class I allele	Population (n)	Frequency of SVA insertion	Frequency of *HLA* allele	% Association of SVA with *HLAcI* allele
SVA-HF	*HLA-B*47*	African American (96)	0.005	0.005	100.0
SVA-HF	*HLA-B*39*	Caucasian (174)	0.019	0.026	74.0
SVA-HF	*HLA-B*52*	Caucasian (174)	0.005	0.009	60.2
SVA-HF	*HLA-A*03*	Caucasian (173)	0.112	0.124	90.4
SVA-HA	*HLA-A*30*	African American (96)	0.124	0.146	85.0
SVA-HA	*HLA-A*03*	African American (96)	0.046	0.057	79.9
SVA-HA	*HLA-A*11*	African American (96)	0.009	0.016	57.2
SVA-HA	*HLA-A*11*	Caucasian (172)	0.084	0.084	100.0
SVA-HA	*HLA-A*30*	Caucasian (172)	0.032	0.032	100.0
SVA-HA	*HLA-A*69*	Caucasian (172)	0.003	0.003	100.0
SVA-HA	*HLA-A*03*	Caucasian (172)	0.119	0.125	94.8
SVA-HA	*HLA-A*01*	Japanese (100)	0.005	0.005	100.0
SVA-HA	*HLA-B*41*	African American (96)	0.010	0.010	100.0
SVA-HA	*HLA-B*47*	African American (96)	0.005	0.005	100.0
SVA-HA	*HLA-B*54*	African American (96)	0.005	0.005	100.0
SVA-HA	*HLA-B*71*	African American (96)	0.010	0.010	100.0
SVA-HA	*HLA-B*42*	African American (96)	0.049	0.049	72.7
SVA-HA	*HLA-B*45*	African American (96)	0.009	0.009	57.2
SVA-HA	*HLA-B*38*	Caucasian (173)	0.009	0.009	100.0
SVA-HA	*HLA-B*41*	Caucasian (173)	0.006	0.006	100.0
SVA-HA	*HLA-B*13*	Caucasian (173)	0.010	0.014	69.7
SVA-HA	*HLA-B*52*	Caucasian (173)	0.006	0.009	66.7
SVA-HA	*HLA-B*35*	Caucasian (173)	0.042	0.072	58.8
SVA-HA	*HLA-B*39*	Caucasian (173)	0.015	0.026	57.9
SVA-HA	*HLA-B*37*	Japanese (99)	0.005	0.005	100.0
SVA-HA	*HLA-C*12*	Caucasian (172)	0.029	0.047	62.7
SVA-HA	*HLA-C*04*	Caucasian (172)	0.052	0.033	56.1
SVA-HA	*HLA-C*06*	Japanese (99)	0.005	0.005	100.0
SVA-HB	*HLA-A*26*	African American (93)	0.011	0.011	100.0
SVA-HB	*HLA-A*36*	African American (93)	0.016	0.016	100.0
SVA-HB	*HLA-A*43*	African American (93)	0.005	0.005	100.0
SVA-HB	*HLA-A*66*	African American (93)	0.022	0.022	100.0
SVA-HB	*HLA-A*80*	African American (93)	0.005	0.005	100.0
SVA-HB	*HLA-A*74*	African American (93)	0.006	0.054	88.1
SVA-HB	*HLA-A*29*	African American (93)	0.009	0.059	84.1
SVA-HB	*HLA-A*68*	African American (93)	0.012	0.059	79.4
SVA-HB	*HLA-A*24*	African American (93)	0.011	0.048	77.3
SVA-HB	*HLA-A*33*	African American (93)	0.023	0.075	70.0
SVA-HB	*HLA-A*34*	African American (93)	0.015	0.048	68.4
SVA-HB	*HLA-A*03*	African American (93)	0.019	0.054	64.9
SVA-HB	*HLA-A*30*	African American (93)	0.055	0.145	62.1
SVA-HB	*HLA-A*02*	African American (93)	0.074	0.183	59.7
SVA-HB	*HLA-A*23*	African American (93)	0.038	0.081	52.7
SVA-HB	*HLA-A*23*	Caucasian (172)	0.005	0.005	100.0
SVA-HB	*HLA-A*25*	Caucasian (172)	0.006	0.006	100.0
SVA-HB	*HLA-A*30*	Caucasian (172)	0.032	0.032	100.0
SVA-HB	*HLA-A*69*	Caucasian (172)	0.003	0.003	100.0

Table 5 (continued)

SVA allele insertion	*HLA* class I allele	Population (n)	Frequency of SVA insertion	Frequency of *HLA* allele	% Association of SVA with *HLAcI* allele
SVA-HB	*HLA-A*01*	Caucasian (172)	0.145	0.177	81.9
SVA-HB	*HLA-A*24*	Caucasian (172)	0.084	0.105	79.8
SVA-HB	*HLA-A*31*	Caucasian (172)	0.032	0.041	78.7
SVA-HB	*HLA-A*26*	Caucasian (172)	0.020	0.026	77.1
SVA-HB	*HLA-A*29*	Caucasian (172)	0.047	0.061	76.6
SVA-HB	*HLA-A*68*	Caucasian (172)	0.032	0.049	64.2
SVA-HB	*HLA-A*02*	Caucasian (172)	0.133	0.247	53.8
SVA-HB	*HLA-A*11*	Caucasian (172)	0.045	0.084	53.8
SVA-HB	*HLA-A*01*	Japanese (100)	0.005	0.005	100.0
SVA-HB	*HLA-A*33*	Japanese (100)	0.017	0.030	57.1
SVA-HB	*HLA-B*08*	African American (93)	0.059	0.059	100.0
SVA-HB	*HLA-B*27*	African American (93)	0.032	0.032	100.0
SVA-HB	*HLA-B*39*	African American (93)	0.011	0.011	100.0
SVA-HB	*HLA-B*40*	African American (93)	0.011	0.011	100.0
SVA-HB	*HLA-B*41*	African American (93)	0.011	0.011	100.0
SVA-HB	*HLA-B*47*	African American (93)	0.005	0.005	100.0
SVA-HB	*HLA-B*50*	African American (93)	0.011	0.011	100.0
SVA-HB	*HLA-B*52*	African American (93)	0.016	0.016	100.0
SVA-HB	*HLA-B*53*	African American (93)	0.086	0.086	100.0
SVA-HB	*HLA-B*71*	African American (93)	0.011	0.011	100.0
SVA-HB	*HLA-B*72*	African American (93)	0.005	0.005	100.0
SVA-HB	*HLA-B*78*	African American (93)	0.011	0.011	100.0
SVA-HB	*HLA-B*35*	African American (93)	0.090	0.108	83.8
SVA-HB	*HLA-B*58*	African American (93)	0.037	0.048	77.3
SVA-HB	*HLA-B*42*	African American (93)	0.050	0.070	71.8
SVA-HB	*HLA-B*13*	African American (93)	0.019	0.027	71.5
SVA-HB	*HLA-B*15*	African American (93)	0.052	0.075	68.8
SVA-HB	*HLA-B*45*	African American (93)	0.011	0.016	66.6
SVA-HB	*HLA-B*44*	African American (93)	0.057	0.086	66.3
SVA-HB	*HLA-B*51*	African American (93)	0.016	0.027	60.0
SVA-HB	*HLA-B*08*	Caucasian (173)	0.142	0.142	100.0
SVA-HB	*HLA-B*27*	Caucasian (173)	0.032	0.032	100.0
SVA-HB	*HLA-B*37*	Caucasian (173)	0.020	0.020	100.0
SVA-HB	*HLA-B*38*	Caucasian (173)	0.009	0.009	100.0
SVA-HB	*HLA-B*39*	Caucasian (173)	0.026	0.026	100.0
SVA-HB	*HLA-B*40*	Caucasian (173)	0.075	0.075	100.0
SVA-HB	*HLA-B*41*	Caucasian (173)	0.006	0.006	100.0
SVA-HB	*HLA-B*45*	Caucasian (173)	0.006	0.006	100.0
SVA-HB	*HLA-B*50*	Caucasian (173)	0.006	0.006	100.0
SVA-HB	*HLA-B*13*	Caucasian (173)	0.013	0.014	86.6
SVA-HB	*HLA-B*35*	Caucasian (173)	0.060	0.072	83.7
SVA-HB	*HLA-B*15*	Caucasian (173)	0.079	0.098	80.8
SVA-HB	*HLA-B*55*	Caucasian (173)	0.022	0.029	76.3
SVA-HB	*HLA-B*52*	Caucasian (173)	0.006	0.009	71.3
SVA-HB	*HLA-B*58*	Caucasian (173)	0.006	0.009	71.3
SVA-HB	*HLA-B*57*	Caucasian (173)	0.020	0.035	57.6
SVA-HB	*HLA-B*13*	Japanese (99)	0.010	0.010	100.0
SVA-HB	*HLA-B*27*	Japanese (99)	0.015	0.015	100.0

Table 5 (continued)

SVA allele insertion	*HLA* class I allele	Population (n)	Frequency of SVA insertion	Frequency of *HLA* allele	% Association of SVA with *HLAcI* allele
SVA-HB	*HLA-B*37*	Japanese (99)	0.005	0.005	100.0
SVA-HB	*HLA-B*55*	Japanese (99)	0.017	0.030	57.2
SVA-HB	*HLA-B*52*	Japanese (99)	0.078	0.141	55.4
SVA-HB	*HLA-C*06*	Caucasian (172)	0.073	0.073	100.0
SVA-HB	*HLA-C*12*	Caucasian (172)	0.047	0.047	100.0
SVA-HB	*HLA-C*16*	Caucasian (172)	0.049	0.049	100.0
SVA-HB	*HLA-C*17*	Caucasian (172)	0.009	0.009	100.0
SVA-HB	*HLA-C*03*	Caucasian (172)	0.171	0.186	92.0
SVA-HB	*HLA-C*04*	Caucasian (172)	0.080	0.093	86.1
SVA-HB	*HLA-C*02*	Caucasian (172)	0.030	0.035	86.0
SVA-HB	*HLA-C*02*	Japanese (99)	0.005	0.005	100.0
SVA-HB	*HLA-C*06*	Japanese (99)	0.005	0.005	100.0
SVA-HB	*HLA-C*12*	Japanese (99)	0.095	0.152	62.7
SVA-HC	*HLA-B*07*	Caucasian (174)	0.092	0.106	86.4

This is a summary table of the more detailed percentage frequency associations and LD analyses between SVA and *HLA* allelic pairs for each population presented in ESM Tables S3 to S5. The haplotype frequency of the SVA absent and present at each loci was associated with a particular *HLA* class I gene allele in the pairwise haplotype LD analysis using the Midas programme of Gaunt et al. (2006)

now being transmitted through to other population groups at a lower frequency.

A six-locus haplotype analysis of the *HLA-A/HLA-B/* four-locus SVA insertions was performed on the population data using the likelihood method and EM algorithm in the Arlequin software package. Table 6 shows the frequency of the six most common six-locus haplotypes in the three populations. The two most common estimated haplotypes in the African Americans were *HLA-A*02/HLA-B*35/SVA-HB* and *HLA A*02/HLA-B*44/SVA-HB* at 2.5% each. The most common estimated haplotypes in Caucasians and Japanese was *HLA-A*01/HLA-B*08/SVA-HB* at 9.4% and *HLA-A*24/HLA-B*52/SVA-HB* at 6.7%, respectively. In Caucasians, an estimated haplotype containing SVA multilocus insertions,

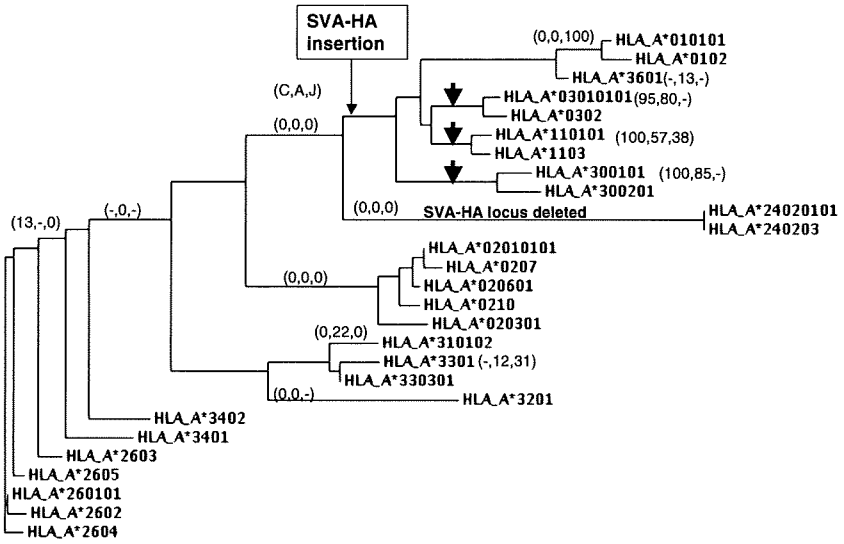


Fig. 3 Phylogenetic tree of *HLA-A* alleles and association with the *SVA-HA* insertion. The phylogenetic tree was constructed using the Jukes–Cantor distance calculations and the neighbour-joining method. Minimum evolution analysis of the complete cDNA sequences of *HLA-A* alleles. The *SVA-HA* insertion association with *HLA-A* alleles is lineage-dependent. *Vertical arrowheads* indicate strong association with *SVA-HA* insertion. The hypothetical point for the *SVA-HA* insertion is indicated. The *numbers in parenthesis* are the percentage association of an *HLA-A* allele with the *SVA-HA* insertion in Caucasians (*C*), African Americans (*A*) and Japanese (*J*) reading from *left to right* (C, A, J). The *dash in place of a number in parenthesis* indicates the *HLA-A* allele is missing from the population. No outgroup was included as the topology is similar to previous publications (McKenzie et al. 1999)

Table 6 Maximum-likelihood (ML) frequency estimation (Arlequin) of the six most common six-locus HLA-A/HLA-B/SVA-four-loci haplotypes in each of 3 populations

ML haplotype		Haplotype
Freq.	SD	
African-American (n=200 chromosomes, 125 listed haplotype frequencies of a possible 460)		
0.0250	0.0134	HLA-A*02/HLA-B*35/SVA-HB/other 3 SVA absent
0.0250	0.0147	HLA-A*02/HLA-B*44/SVA-HB/other 3 SVA absent
0.0200	0.0106	HLA-A*23/HLA-B*49/SVA-HB/other 3 SVA absent
0.0200	0.0107	HLA-A*03/HLA-B*57/SVA-HA/other 3 SVA absent
0.0200	0.0106	HLA-A*33/HLA-B*15/SVA-HB/other 3 SVA absent
0.0200	0.0110	HLA-A*33/HLA-B*42/SVA-HB/other 3 SVA absent
Australian-Caucasian (n=348 chromosomes, 138 listed haplotype frequencies of a possible 852)		
0.0939	0.0158	HLA-A*01/HLA-B*08/SVA-HB/other 3 SVA absent
0.0758	0.0149	HLA-A*02/HLA-B*44/all 4 SVA absent
0.0371	0.0097	HLA-A*29/HLA-B*44/SVA-HB/other 3 SVA absent
0.0360	0.0108	HLA-A*03/HLA-B*07/SVA-HF/SVA-HA/SVA-HC/ other SVA absent
0.0260	0.0099	HLA-A*02/HLA-B*07/SVA-HC/other 3 SVA absent
0.0247	0.0092	HLA-A*02/HLA-B*27/SVA-HB/other 3 SVA absent
Japanese (n=200 chromosomes, 72 listed haplotype frequencies of a possible 217)		
0.0669	0.0206	HLA-A*24/HLA-B*52/SVA-HB/other 3 SVA absent
0.0636	0.0203	HLA-A*02/HLA-B*40/all 4 SVA absent
0.0600	0.0178	HLA-A*24/HLA-B*54/all 4 SVA absent
0.0406	0.0185	HLA-A*24/HLA-B*52/all 4 SVA absent
0.0396	0.0160	HLA-A*24/HLA-B*15/all 4 SVA absent
0.0383	0.0161	HLA-A*02/HLA-B*51/all 4 SVA absent

Arlequin conventional EM algorithm (one pass), bootstrap was 1,000, number of random conditions for EM was 50

*HLA-A*03/HLA-B*07/SVA-HF/SVA-HA/SVA-HC*, was relatively high at 3.6%.

The evolution of classical HLA class I gene/SVA insertion haplotypes

The *SVA-HB* insertion appears to be the oldest of the four SVA insertions in this study based on its high insertion frequency, significant LD and high percentage association with many different *HLA-B* alleles in the cell lines and the three populations. This was confirmed by the Repeat-Masker alignment programme that identified the *SVA-HB* sequence as a member of the SVA_B subfamily (Table 2) that emerged 11.56 Mya, probably before the origin of the great apes, the orangutan, the gorilla, the chimpanzee and the human (Wang et al. 2005). In comparison, Repeat-Masker and phylogenetic analysis (Wang et al. 2005) diagnosed the *SVA-HA* insertion as a SVA-D subtype that emerged with the great apes about 9.55 Mya, the *SVA-HF* insertion as a SVA_E subtype that emerged in the human evolutionary lineage about 3.46 Mya and the *SVA-HC* insertion, which has the lowest overall population frequency (Table 3), as a SVA_F subtype that emerged 3.18 Mya after the separation of the chimpanzee and human lineages. On this basis, it is evident that *SVA-HB* is the oldest and *SVA-HC* is the youngest of the four SVA insertions in this study.

The level of association of a SVA insertion with a particular *HLA* allele in a population might depend largely on the frequency of the *HLA* allele that was originally linked to the SVA insertion and whether or not the original *HLA* allele and SVA combination (haplotype) has changed with the ensuing generations due to either sequence mutation of the original *HLA* allele or the occurrence of allelic exchange due to crossing-over events (as outlined in ESM Fig. S2). Most *HLA*/SVA haplotypes generated due to a mutation in the *HLA* gene are expected to have remained at 100% linkage (identity by descent) if no further changes had occurred in the haplotype structure in subsequent generations due to crossing-over or deletions. On the other hand, *HLA* class I gene and SVA allelic combinations due to crossing over are expected to have generated mostly low frequency haplotypes, unless the crossing over had occurred very early in the *HLA* allele's life history. On the basis of phylogeny, the *SVA-HA* appears to have been first

inserted in the *HLA-A*24* or *HLA-A*30* lineage before being passed on to the *HLA-A*03* and *HLA-A*11* lineages (Fig. 3), assuming that the original *SVA-HA/HLA-A* haplotype has survived. The low percentage association of *SVA-HA* with other *HLA-A* alleles such as *HLA-A*01*, *-A*26*, *-A*31*, *-A*33*, *-A*36* and *-A*66* may have then arisen by crossing over or by other genomic exchange mechanisms. Because the *SVA-HB* is an older insertion than *SVA-HA* and has a much higher frequency across three populations, the original *SVA-HB/HLA-B* haplotype was not identified among the many different *SVA-HB/HLA-B* allele combinations. Most of the SVA associations with *HLA-B* alleles appear, however, to be *HLA-B* allele lineage-dependent (McKenzie et al. 1999) with a small phylogenetic cluster of *HLA-B* alleles, such as *HLA-B*07*, *-B*48* and *-B*81*, largely free of *SVA-HB* insertions, and the neighbouring phylogenetic clusters of *HLA-B*08, -B*27, -B*40* and *-B*47* with 100% *SVA-HB* association in Caucasians and African Americans. Some neighbouring phylogenetic clusters such as those with *HLA-B*14* or *-B*39* that have diverged from *HLA-B*07* or *HLA-B*08* (McKenzie et al. 1999) were found to have a low (<15%) or high (100%) association with *SVA-HB*, respectively. Thus, a phylogenetic tree analysis (data not shown) may reveal which associations between the *SVA-HB* and *HLA-B* alleles have arisen by direct lineage inheritance, *HLA-B* allelic mutations, crossing over or gene conversion events as outlined in the schemes shown in ESM Fig. 2S and for the *SVA-HA* insertion in the *HLA-A* lineages (Fig. 3). In this regard, phylogenetic analyses of the *HLA-B* allele associations with the *SVA-HB* insertion, similar to that shown for *HLA-A* alleles and the *SVA-HA* insertion in Fig. 3, may help reveal interesting *HLA-B* allele and haplotype lineages when more population frequency data have been obtained.

In comparison to *SVA-HB and SVA-HA*, the other SVA insertions appear to have originated more recently and consequently have been linked to fewer specific *HLA* class I alleles, such as between *SVA-HC* and *HLA-B*0702* and *HLA-C*0702*. Thus, the frequencies of the *SVA-HC* and *SVA-HF* are much lower (Table 3) and their insertions are associated at a high percentage with only a few *HLA* class I alleles rather than a wider range and a relatively larger number of *HLA* class I alleles as in the case of *SVA-B*. In this regard, the frequencies and *HLA* allelic relationships for the other 14 or more SVA sequences that have been identified within the *MHC* genomic region (Fig. 1a) warrant investigation.

Conclusions

This is the first comparative genetic study of a set of multilocus SVA in the *HLA* class I region of three distinct populations which shows that these SVA alone or together with the *HLA* class I alleles are informative genetic markers for the identification of allele and haplotype lineages and variations within the same or different populations. This study has provided an insight into the relationships between SVA structural polymorphism at four loci and the adjacent classical *HLA* class I loci and has helped in the characterisation of extended *HLA* haplotypes. These issues are of importance for population differentiation studies in anthropology and DNA forensics, the identification of disease associations and for a better definition of donor–recipient compatibility in bone marrow grafts through the typing of haplospecific markers.

Finally, the PCR assays for the SVA insertions in this study were developed from SVA sequences in the common Caucasian haplotypes, *HLA-A*0101/HLA-B*0801/HLA-DRB1*0301* and *HLA-A*0301/HLA-B*0702/HLA-DRB1*1501* (Stewart et al. 2004). This method of SVA PCR primer selection and development is unlikely to have contributed significantly to an ascertainment bias where the frequencies of the SVA insertion alleles and haplotypes in this study were highest in the Caucasians, intermediate for African Americans and lowest in Japanese because only two distinct alleles (absent and present) are the intrinsic measure of this system. Ascertainment bias is more likely to occur in the analysis of SNP data and affect the frequency spectra and summary population statistics such as frequencies of SNP alleles, genotypes and haplotypes when minor SNP alleles are not accounted for in geographically restricted ascertainment population panel of a small number of individuals (Clark et al. 2005). A number of other common Caucasian HLA haplotypes that have been completely sequenced (Horton et al. 2008) were not investigated for additional polymorphic SVA markers in this study. The development of polymorphic SVA insertion markers to better discriminate between Caucasians, Japanese, African Americans and other population groups may require a more detailed analysis of the other common *HLA* haplotype genomic sequences of Caucasian and non-Caucasian populations.

Acknowledgements We thank Paula M Moolhuijzen for her help with the initial bioinformatics genomic analysis for some of the PCR primer sets, Professor M Ota for the Japanese *HLA*-typed DNA samples, Dr. Campbell Witt for the Australian Caucasian *HLA*-typed DNA samples and Dr. Takashi Shiina for helpful discussions.

References

Bennett EA, Coleman LE, Tsui C, Pittard WS, Devine SE (2004) Natural genetic variation caused by transposable elements in humans. Genetics 168:933–951

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Neilsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15:1496–1502

Dunn DS, Ota M, Inoko H, Kulski JK (2003) Association of MHC dimorphic Alu insertions with HLA class I and MIC genes in Japanese *HLA-B*48 haplotypes. Tissue Antigens 62:259–262

Dunn DS, Tait BD, Kulski JK (2005) The distribution of polymorphic Alu insertions within the MHC class I *HLA-B*7 and *HLA-B*57 haplotypes. Immunogenetics 56:765–768

Dunn DS, Inoko H, Kulski JK (2006) The association between non-melanoma skin cancer and a young dimorphic Alu element within the major histocompatibility complex class I genomic region. Tissue Antigens 68:127–134

Dunn DS, Choy MK, Phipps ME, Kulski JK (2007) The distribution of major histocompatibility complex class I polymorphic Alu insertions and their associations with HLA alleles in a Chinese population from Malaysia. Tissue Antigens 70:136–143

Edwards MC, Gibbs RA (1992) A human dimorphism resulting from loss of an *Alu*. Genomics 14:590–597

Excoffier L, Laval G, Schneider S (2005) Arlequin ver. 3.0: an integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1:47–50

Fernando MM, Stevens CR, Walsh EC, De Jager PL, Goyette P, Plenge RM et al (2008) Defining the role of the *MHC* in autoimmunity: a review and pooled analysis. PLoS Genet 4:e1000024

Garcia-Perez JL, Doucet AJ, Bucheton A, Moran JV, Gilbert N (2007) Distinct mechanisms for trans-mediated mobilization of cellular RNAs by the LINE-1 reverse transcriptase. Genome Res 17:602–611

Garnier-Gere P, Dillmann C (1992) A computer program for testing pairwise linkage disequilibria in subdivided populations. J Heredity 83:239

Gaudieri S, Dawkins RL, Habara K, Kulski JK, Gojobori T (2000) SNP profile within the human major histocompatibility complex reveals an extreme and interrupted level of nucleotide diversity. Genome Res 10:1579–1586

Gaunt TR, Rodriguez S, Carlos Zapata C, Day INM (2006) MIDAS: software for analysis and visualisation of interallelic disequilibrium between multiallelic markers. BMC Bioinformatics 7:227–238

Goudet J (1995) FSTAT (version 1.2): a computer program to calculate *F*-statistics. J Heredity 86:485–486

Hassoun H, Coetzer TL, Vassiliadis JN, Sahr KE, Maalouf GJ, Saad ST, Catanzariti L, Palek J (1994) A novel mobile element inserted in the alpha spectrin gene: spectrin dayton. A truncated alpha spectrin associated with hereditary elliptocytosis. J Clin Invest 94:643–648

Horton R, Gibson R, Coggill P, Miretti M, Allcock RJ, Almeida J, Forbes S, Gilbert JG, Halls K, Harrow JL, Hart E, Howe K, Jackson DK, Palmer S, Roberts AN, Sims S, Stewart CA, Traherne JA, Trevanion S, Wilming L, Rogers J, de Jong PJ, Elliott JF, Sawcer S, Todd JA, Trowsdale J, Beck S (2008) Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. Immunogenetics 60:1–18

Itoh Y, Inoko H, Kulski JK, Sasaki S, Meguro A, Takiyama N, Nishida T, Yuasa T, Ohno S, Mizuki N (2006) Four-digit allele genotyping of the *HLA-A* and *HLA-B* genes in Japanese patients with Behcet's disease by a PCR–SSOP–Luminex method. Tissue Antigens 67:390–394

Kulski JK, Dunn DS (2005) Polymorphic Alu insertions within the major histocompatibility complex class I genomic region: a brief review. Cytogenet Genome Res 110:193–202

Kulski JK, Gaudieri S, Martin A, Dawkins RL (1999) Coevolution of PERB11 (MIC) and HLA class I genes with HERV-16 and retroelements by extended genomic duplication. J Mol Evol 49:84–97

Kulski JK, Shigenari A, Shiina T, Ota M, Hosomichi K, James I, Inoko H (2008) Human endogenous retrovirus (HERVK9) structural polymorphism with haplotypic *HLA-A* allelic associations. Genetics 180:445–457

Kulski JK, Shigenari A, Shiina T, Hosomichi K, Yawata M, Inoko H (2009) HLA-A allele associations with viral MER9-LTR nucleotide sequences at two distinct loci within the MHC alpha block. Immunogenetics 61:257–270

Makino S, Kaji R, Ando S, Tomizawa M, Yasuno K, Goto S, Matsumoto S, Tabuena MD, Maranon E, Dantes M, Lee LV, Ogasawara K, Tooyama I, Akatsu H, Nishimura M, Tamiya G (2007) Reduced neuron-specific expression of the TAF1 gene is associated with X-linked dystonia-parkinsonism. Am J Hum Genet 80:393–406

Marsh SG (2000) WHO Nomenclature Committee for Factors of the *HLA* System. Nomenclature for factors of the *HLA* system, update July 2000. Tissue Antigens 56:476–477

Marsh SGE, Parham P, Barber DL (2000) The HLA factsbook. Academic, London

McKenzie LM, Pecon-Slattery J, Carrington M, O'Brien SJ (1999) Taxonomic hierarchy of HLA class I allele sequences. Genes Immun 1:120–129

Moriyama Y, Kato K, Mura T, Juji T (2006) Analysis of *HLA* gene frequencies and *HLA* haplotype frequencies for bone marrow donors in Japan. MHC 12:83–201 (in Japanese)

Ono M, Kawakami M, Takezawa T (1987) A novel human nonviral retroposon derived from an endogenous retrovirus. Nucleic Acids Res 15:8725–8737

Ostertag EM, Goodier JL, Zhang Y, Kazazian HH Jr (2003) SVA elements are nonautonomous retrotransposons that cause disease in humans. Am J Hum Genet 73:1444–1451

Perneger TV (1998) What is wrong with Bonferroni adjustments. Br Med J 136:1236–1238

Perrière G, Gouy M (1996) WWW-Query: an on-line retrieval system for biological sequence banks. Biochimie 78:364–369

Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J Heredity 86:248–249

Saitou N, Nei M (1986) The number of nucleotides required to determine the branching order of three species, with special reference to the human–chimpanzee–gorilla divergence. J Mol Evol 24:189–204

Sasieni PD (1997) From genotypes to genes: doubling the sample size. Biometrics 53:1253–1261

Shen L, Wu LC, Sanlioglu S, Chen R, Mendoza AR, Dangel AW, Carroll MC, Zipf WB, Yu CY (1994) Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon–intron structure, composite retroposon, and breakpoint of gene duplication. J Biol Chem 269:8466–8476

Shiina T, Ota M, Shimizu S, Katsuyama Y, Hashimoto N, Takasu M, Anzai T, Kulski JK, Kikkawa E, Naruse T, Kimura N, Yanagiya K, Watanabe A, Hosomichi K, Kohara S, Iwamoto C, Umehara Y, Meyer A, Wanner V, Sano K, Macquin C, Ikeo K, Tokunaga K, Gojobori T, Inoko H, Bahram S (2006) Rapid evolution of major histocompatibility complex class I genes in primates generates new disease alleles in humans via hitchhiking diversity. Genetics 173:1555–1570

Shiina T, Hosomichi K, Inoko H, Kulski JK (2009) The HLA genomic loci map: expression, interaction, diversity and disease. J Hum Genet 54:15–39

Stewart CA, Horton R, Allcock RJ, Ashurst JL, Atrazhev AM, Coggill P, Dunham I, Forbes S, Halls K, Howson JM, Humphray SJ, Hunt S, Mungall AJ, Osoegawa K, Palmer S, Roberts AN, Rogers J, Sims S, Wang Y, Wilming LG, Elliott JF, de Jong PJ,