European Journal of Human Genetics (2001) 9, 590–598

www.nature.com/ejhg

ARTICLE

Short tandem repeat (STR) haplotypes in *HLA*: an integrated 50-kb STR/linkage disequilibrium/gene map between the *RING3* and *HLA-B* genes and identification of STR haplotype diversification in the class III region

Igor Vorechovsky*[,1,2], Jana Kralovicova[1], Michael D Laycock[1,2], A David B Webster[2], Steven GE Marsh[3], Alejandro Madrigal[3] and Lennart Hammarström[1]

[1]*Department of Biosciences at NOVUM, Karolinska Institute, S-14157 Huddinge, Sweden;* [2]*Department of Clinical Immunology, University College London, London NW3 2PF, UK;* [3]*Anthony Nolan Research Institute, Royal Free Hospital, London NW3 2QG, UK*

We present a dense STR/linkage disequilibrium(LD)/gene map between the *RING3* and *HLA-B* loci, reference allelic sizes on the most prevalent *HLA* haplotypes and their allelic frequencies in pedigree founders. This resource will facilitate LD, evolution and gene mapping studies, including comparisons of *HLA* and STR haplotypes and identification of *HLA* recombinants. The map was constructed by testing novel and previously reported STRs using a panel of 885 individuals in 211 families and 60 DNA samples from cell lines and bone marrow donors homozygous in the *HLA-A*, *-B* and *-DR* loci selected from over 15 000 entries into the registry of Swedish bone marrow donors. We have also analysed the variability of STR alleles/haplotypes on the most prevalent *HLA* haplotypes to identify STRs useful for fine mapping of disease genes in the region previously implicated in susceptibility to many disorders. The analysis of 40 *HLA-A*01, B*0801, DRB1*03011, DQB1*0201* haplotypes in homozygous donors showed a surprising stability in 23 STRs between the class II recombination hot spot and *HLA-B*, with the average of 1.9% (16/838) variant alleles. However, 40% variant alleles were found at the *D6S2670* locus in intron 19 of the tenascin-X gene both in the families and homozygous donors. The nucleotide sequence analysis of this STR showed a complex polymorphism consisting of tetra- $(CTTT)_{8-18}$ and penta-nucleotide $(CTTTT)_{1-2}$ repeats, separated by an intervening non-polymorphic sequence of 42 bp. The *HLA-A1, B*0801, DRB1*03011, DQB1*0201* haplotypes had five $(CTTT)_{14-18}/(CTTTT)_2$ variants with a predominant $(CTTT)_{16}$ allele, implicating the tetranucleotide component as the source of this ancestral haplotype diversification, which may be due to the location of *D6S2670* in the region of the highest GC content in the human MHC. *European Journal of Human Genetics* (2001) **9**, 590–598.

Keywords: MHC; recombination; linkage disequilibrium; short tandem repeat; GC content; *HLA*

Introduction

Extensive length polymorphism of abundant short tandem repeats (STR)[1] has been successfully used in genetic studies for more than a decade. A recent completion of the nucleotide sequence of a human MHC[2] has enabled the localisation of *HLA* STRs on the physical map.[3–5] Although valid allelic association, linkage, disease gene fine-mapping, population and evolution studies using these genetic markers require knowledge of STR properties such as inheritance and stability, only a limited number of *HLA* STRs has been evaluated in this respect.

In the present study, we have tested a number of novel and previously reported STRs located in the region of over 1.5 Mbp in the centromeric half of *HLA* using a large family

*Correspondence: Igor Vorechovsky, Department of Biosciences at NOVUM, Karolinska Institute, S-14157 Huddinge, Sweden.
Tel/Fax: +468 6089269; E-mail: igvo@cbt.ki.se
Received 2 January 2001; revised 2 April 2001; accepted 23 May 2001

material and a panel of homozygous bone marrow donors and cell lines. We have characterised STR haplotypes and analysed their stability on the most prevalent *HLA* haplotypes in a Caucasian population, with a particular emphasis on the *HLA-A1*, *B8*, *DR3* haplotype which has been associated with a number of autoimmune diseases. We also provide a detailed composite STR/gene/linkage disequilibrium (LD) map with the average STR density of about 50 kb. Finally, we have identified a recent ancestral diversification at an STR located in the region of the highest GC content in the human MHC and show the molecular structure of this complex locus.

Materials and methods

Subjects

Fifty-two samples homozygous at *HLA-A*, *-B* and *-DRB1* loci, including 20 *HLA-A1*, *B8*, *DRB1*03* homozygotes, were obtained from the Tobias Registry of Swedish bone marrow donors (the Huddinge University Hospital, Sweden). The samples were ascertained by a search for *HLA-A, -B* and *-DRB1* homozygotes in over 15 000 registry entries. About half of the selected donors lived in the greater Stockholm area, whereas the remaining cases came from both southern and northern parts of Sweden. Homozygous donors carrying the *HLA-A1, B8, DRB1*03* haplotypes came from all parts of Sweden, with only four out of 20 coming from the Stockholm region. In addition to patients' samples, eight cell lines were analysed. The donor samples have been previously typed using serology (*HLA-A* and *-B*) and low-resolution PCR-SSP (*HLA-DRB1*). Sequence-based typing was carried out to define the *HLA* specificities at the DNA level (see below). The designation, ethnic origin and *HLA* specificities of the analysed cell lines and donors are shown at the Karolinska Institute's Web site (http://www.cbt.ki.se/fam/res/tab1.htm).

STR haplotypes were constructed in 885 family members from 211 pedigrees ascertained through probands with IgA deficiency or common variable immunodeficiency as described previously.[3,6,7] The family material consisted of both single- (n=110) and multiple-case (n=101) pedigrees. Multiple-case families are shown at the Karolinska Institute's web page at http://www.cbt.ki.se/fam/set3-7/scan_set.html.

DNA extraction and PCR amplifications

DNA was extracted from peripheral blood as previously described.[8] All PCR reactions were carried out in a 96-well format in a volume of 20 μl, containing 50 ng of DNA, 0.25 μM of each primer, 200 μM of each dNTP, 10 mM Tris-HCl (pH 8.3), 50 mM KCl, variable concentrations of Mg^{2+} (http://www.cbt.ki.se/fam/res/str.htm), 0.01% gelatine and 0.5 units of Taq polymerase (Pharmacia Biotech, Uppsala, Sweden).

Genotyping

The *HLA-B, HLA-DRB1* and *HLA-DQB1* alleles were determined using sequence-based typing as described previously,[9–11] except for PCR conditions for the *HLA-B* locus. For the PCR amplification of exons 2 and 3 of the *HLA-B* gene, the annealing temperature of 69°C was used with the Mg^{2+} concentration of 1.5 mM.

Allelic sizes at STR loci were determined using an ABI 377 Sequencer (Applied Biosystems, USA) with a 96-well loading option, coupled with GeneScan (v. 3.1) and Genotyper (v. 2.0) software packages (Applied Biosystems, USA). The internal size standard GeneScan-500 TAMRA (Applied Biosystems, USA, part No. 401733) was used with each PCR pool. Allelic sizes were converted to allele numbers using the ALSIZE module of the Genetic Analysis System (GAS), v. 2.0 (A. Young, University of Oxford, UK, http://users.ox.ac.uk/~ayoung/gas.html).

Of over 50 STRs tested, family and homozygous samples were typed using 35 primer pairs in the above-defined region, amplifying 33 distinct loci (Figure 1). The oligonucleotide primers, allelic sizes and their frequencies, heterozygosity, annealing temperatures and concentrations of Mg^{2+} for each STR, including their physical distances on a mosaic of sequenced haplotypes[2] are shown at the Karolinska Institute web page (http://www.cbt.ki.se/fam/res/str.htm). The original references for STR loci (Figure 1 and Table 1) are as follows (from centromere to telomere): *RING3CA*;[12] *D6S2445*;[13] *D6S2659*;[4] *TAP1CA*;[14,15] *D6S2660*;[4] *D6S2661*;[4] *7-8601*[3] or *D6S2444*;[13] *6-38576*[3] or *M6S232*[4] or *D6S2443*;[13] *2E3*;[13] *G51152*[12] or *D6S2663*;[4] *DQCAR (D6S2447)*;[4,16] *DQCARII*[17,18] or *D6S1666*[19] or *D6S2446*;[13] *D6S2664*;[4] *D6F374S1*;[13] *9-9943*[3] or *M6S118*;[4] *8-105224*[3] or *MSS119*;[4] *LH1*;[20] *13-36829* (this study); *10-13898* (this study) or *D6S2667*;[4] *D6S2668*;[4] *8-39925*;[3] *D6S1014*;[4] *11-36252*;[3] *D3A* or *15-47549*[3,20] or *D6S2669*[4]; *D6S2670*;[4] *19-162236*;[3] *9N2* or *21-8631*;[3,20] *9N1*;[20] *D6S273*;[19] *22-38114*;[3] *K11*;[21] *24-140297*;[3] *82-3*;[20] *TNFb*;[4] *62 (=TNFa,b)*;[20] *TNFa-e*;[22] *AB59840*[3] or *C1_2_A*;[23] *AB121571*[3] or *C1_2_C*;[23] *MICA*;[24] *MIB*;[18] *C1_4_1*.[23]

Nucleotide sequence analysis of D6S2670

The STR was amplified using PCR primers 5′-GTGAATTGTGACTGTGCCAGTACAC and 5′-CCACCCACTTCCTCCACTAGAATC. PCR products were subcloned into the pGEM-T vector (Promega). Several clones have been sequenced for each allele using vector primers and the ABIPrism BigDye Terminator Reaction Kit (Applied Biosystems) as described.[8] The size of the sequenced product was compared to that determined by the fragment analysis using the Genescan/Genotyper programmes.

Haplotype construction and linkage disequilibrium analysis

Haplotypes were constructed using the Genehunter (v. 2.0, http://waldo.wi.mit.edu/ftp/distribution/software/genehunter/gh2/)[25] and Simwalk2 (v. 2.60, http://watson.hgen.pitt.edu/register)[26,27] algorithms. Haplotypes were inspected and corrected manually against the Genotyper plots. LD measures between STR loci were computed with

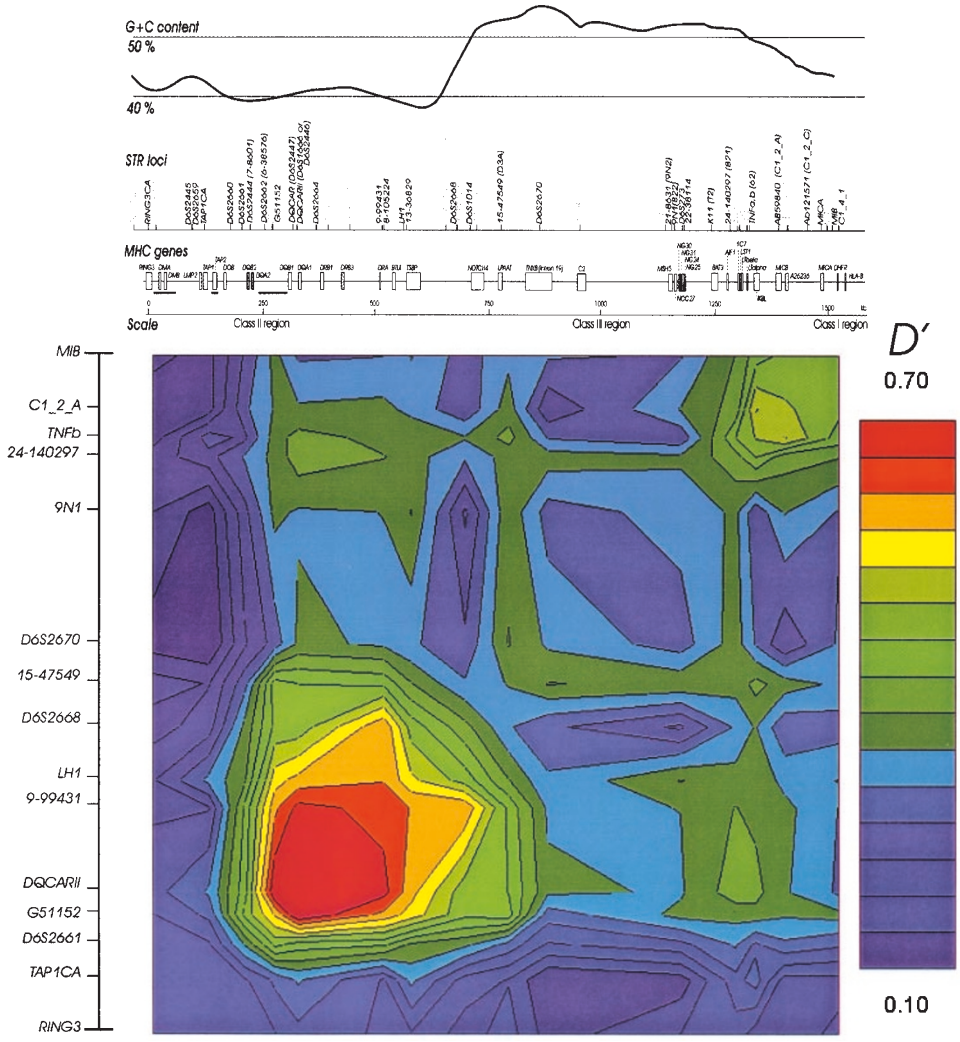

Figure 1 Integrated STR/LD/gene map of the centromeric half of *HLA*. The map is based on the nucleotide sequence of a human MHC as described.[2] Previously reported recombination hot spots in the class II region are shown as bars.[30,46] Only a subset of MHC genes relevant to the position of each STR is provided. Pseudogenes or genes that exist only on some *HLA* haplotypes are shown as grey boxes. STR loci with limited heterozygosity or those not used in this study are indicated in grey. Map distances are drawn to scale and represent a particular haplotype sequenced. They may vary on different haplotypes due to insertion/deletion polymorphism.

the GOLD programme[28] using the Simwalk2-generated haplotypes as an input. LD measures were defined as previously described.[28,29] The graphical representation of the LD pattern across the tested region was constructed with the GOLD utility.

Results

Integrated STR/LD/gene map in the centromeric half of HLA

Novel and previously reported STR loci have been placed on the physical map (Figure 1) and STR haplotypes were constructed using 885 members of 211 families. The STR haplotypes containing a subset of STR loci exhibiting unequivocal allele clusters were used to produce the LD/STR/gene map (Figure 1, lower panel). The map is accompanied by exact numerical representations of pair-wise Lewontin's D′ coefficients and several other LD measures, such as Cramer's coefficients (http://www.cbt.ki.se/fam/res/LD.htm).

This analysis showed a maximum LD at the *HLA-DQ/DR* region and a steep LD fall in the proximal area, starting just centromeric of *G51152*, followed by a less precipitous decline between the first and second class II recombination hot spots[30] and reaching low levels at *TAP1/TAP2* (Figure 1). Low LD measures were found for STRs centromeric of the *DOB* gene, with the lowest values in the region between *RING3* and *TAP1CA*.

Table 1 Predominant alleles observed in STR loci between the *TAP1* and *HLA-B* genes on major *HLA* haplotypes in homozygous bone marrow donors

Number of haplotypes analysed	HLA haplotypes HLA-DRB1	HLA-B	HLA-A	STR haplotypes TAP1CA D6S2660	7-8601 D6S2661	6-38576 (D6S2444)	G51152 (D6S2662)	DQCAR (D6S2663)	DQCARII (D6S2447)	9-99431 D6S2664	8-105224 (D6S2446)	(M65118)	(M65119)	LH1	13-36829 D6S2668	D6S1014	15-47549 (D3A)	D6S2669	21-86319N1 (9N2)	D6S2670	D6S273 (8Z-2)	K11 22-38114	24-140297 (8Z-1)	62 TNFb	(TNFa,b)	AB59840	111C1_2_A	C1_2_C (AB121571)	MIB MICA	C1_4_1
8	*0302*04011*1501	2,3,24	192 262	137	150	229	226	110	192	282	237	254	91	178	156	147	264	122	226	263	102	136	228	x	224	100 146	305	250	236	182	256	217
4	*0302*15011*1501	2	190 270	133	154	223	222	102	198	null	222	229	89	178	163	147	266	124	205	x	x	x	x	x	214	100 146	305	250	236	182	256	217
6	*0301*04011*4402	2	192 270	141	146	196	214	116	192	282	237	254	91	178	156	147	270	128	214	263	98	134	230	x	x	108 x	291	236	250	183	286	217
4	*0202*0701	*4403123	190 270	129	160	204	214	112	216	276	223	229	83	176	159	144	276	134	x	x	x	x	224	152	x	110 159	299	244	249	185	266	217
12	*0602*15011*070211	2,3,24	190 270	133	154	223	222	102	198	null	222	229	89	178	163	147	266	124	205	267	96	136	226	154	x	118 167	297	242	250	183	266	225
6	*0501*0101	*3501 3	190 270	137	146	194	244	102	200	x	239	250	101	192	159	147	264	122	217	257	98	134	228	152	216	106 156	307	252	249	194	282	229
4	ND	*04011*270522,24	x 262	133	146	194	214	110	192	282	237	x	91	178	156	147	264	x	213	257	102	134	226	156	x	108 158	291	236	243	179	270	217
42	*0201*03011*0801	1,2	192 262	133	154	196	216	98	202	283	235	248	77	176	163	153	270	128	226	253	98	140	220	154	212	100 148	x	256	252	183	280	221
			6/36 7/40	1/40	6/42	5/34	1/40	0/40	2/38	0/40	1/38	0/32	0/421/42		0/42	0/40	3/42	3/42	16/42	0/36	0/323/42		0/42	0/401/34		2/422/42	x	2/34	0/34	0/420/380/32		

*Proportion of variant alleles on the HLA-A1/2, B*0801, DRB1*03011, DQB1*0201 haplotypes*

STR loci are ordered from centromere to telomere. Their location relative to *HLA* genes is shown on the physical map drawn to scale (Figure 1 and at http://www.cbt.ki.se/fam/res/tab1.htm). Allelic sizes are in bp; 'x', no predominant allele on a particular haplotype. Non-amplified alleles at the *D6S2664* locus are designated 'null'. Full genotypes are available at http://www.cbt.ki.se/fam/res/tab1.htm for each STR and sample, including samples from the Centre d'Etude du Polymorphisme Humain (CEPH) family as a size reference. PCR conditions, oligonucleotide primers, the observed heterozygosity and polymorphism information content at each STR are shown at http://www.cbt.ki.se/fam/res/str.htm. The *C1_2_A/AB59840* and *15-47549(D3A)/D6S2669* primers amplify the same repeats.

In contrast, the LD decline was much slower in the opposite direction with another increase in the region containing the tumour necrosis factor (*TNF*) gene cluster and the *HLA-B* locus (Figure 1). The map, which represents so far the most detailed representation of LD progression in this region, also reveals examples of 'long-range' LD, such as between *24-140297/TNF* and *DQ/DR* region or *15-DRA* and *TNF* (Figure 1).

STR stability on the HLA-A1, B*0801, DRB1*03011, DQB1*0201 haplotype

Predominant STR alleles on this haplotype are shown in the lower panel of Table 1, together with the observed proportion of variant alleles at each STR, except *RING3*, *D6S2445* and *D6S2659*. No clear allele predominance was found for these proximal STRs, which are located in the recombination hot-spot-containing region of low LD centromeric of *TAP2* (Figure 1).[30–32] STRs in the region flanked by the *HLA-DQB1* and *-DRA* genes, which contains the most polymorphic genes in the analysed area, showed no variant alleles on the *HLA-A1, B*0801, DRB1*03011, DQB1*0201* haplotypes, except for another homozygous genotype at *DQCARII* in one sample (full genotypes for each donor and locus are shown at http://www.cbt.ki.se/fam/res/tab1.htm). In contrast, in the region proximal to *G51152*, a marker located just centromeric to the *DQB1* gene (Figure 1), we found a total of 74/292 (25.3%) variant alleles. This proportion was only 3.6% (32/878) for *G51152* and all distal STRs (Table 1, Figure 1 and http://www.cbt.ki.se/fam/res.tab1.htm), with a prominent exception of the *D6S2670* locus. Excluding this locus, only 16/838 allelic variants (1.9%) or 15/419 variant genotypes (3.6%) were identified in the distal region, indicating a surprising stability of STR alleles on the *HLA-A1, B*0801, DRB1*03011, DQB1*0201* haplotypes.

Complex structure of the D6S2670 locus

In contrast to the majority of STRs, only one half of the *HLA-A1, B*0801, DRB1*03011, DQB1*0201* donors were homozygous at the *D6S2670* locus in the class III region, with just 7/20 samples homozygous for the predominant 226-bp allele (Table 1). This allele was observed on 24/40 (60%) *HLA-A1, B*0801, DRB1*03011, DQB1*0201* haplotypes. A total of 24 alleles were found in 885 family members and 52 homozygous donors (the observed heterozygosity was 0.903 and the polymorphism information content was 0.895). The proportion of variant alleles in families on the corresponding STR haplotype (Table 1) was not significantly different from donors (data not shown). No null alleles were observed in families, but we found a single case of non-inheritance, manifested as a 4-bp expansion of the 230-bp allele passed to the son from his heterozygous father (Figure 2), suggesting an STR mutation.

Because *D6S2670* was identified as the most variable STR on this haplotype from over 50 STRs tested in this region

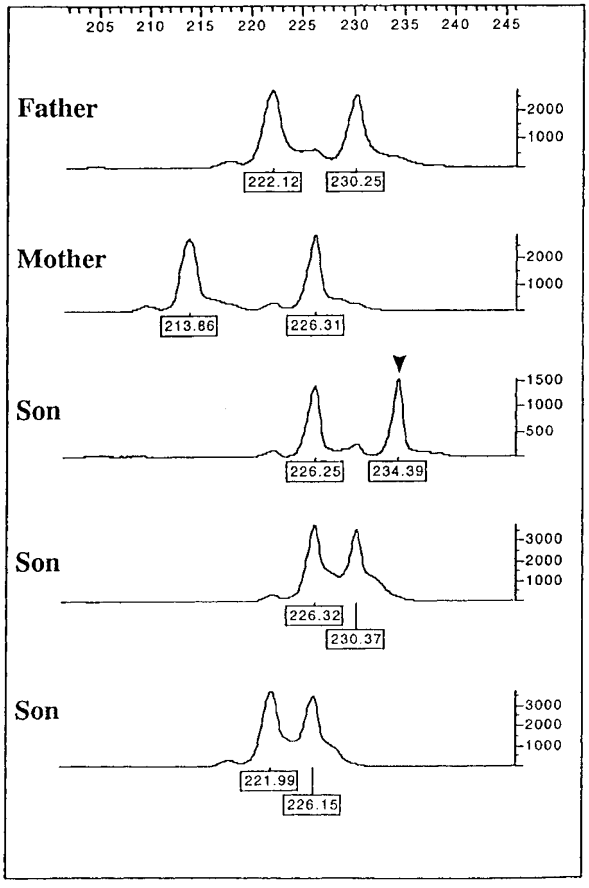

Figure 2 STR mutation at the *D6S2670* locus. Expanded 234-bp allele in the first son (Swedish family cv37) is indicated by an arrow. This family showed a correct Mendelian inheritance for over 250 STRs distributed on all chromosomes (Vorechovsky *et al.*, manuscript in preparation).

(data not shown), we determined its structural diversity by nucleotide sequencing (Table 2). These results indicated a complex locus, consisting of two closely linked polymorphic tetra- $(CTTT)_{8-18}$ and penta-nucleotide $(CTTTT)_{1-2}$ repeats, with an intervening non-polymorphic sequence of 42 bp. An additional polymorphism was identified as a 2-bp deletion in the $(CTTT)_{11-14}$ alleles in absolute LD with the $(CTTTT)_2$ allele (Table 2).The physical distance between the deletion and the intervening sequence was identical on each allele and was equal to 10 bp, indicating that the tetranucleotide variability was restricted to its distal part. Whereas the pentanucleotide component was not polymorphic on the *HLA-A1, B*0801, DRB1*03011, DQB1*0201* haplotypes and always consisted of two tandem repeats, the tetranucleotide repeat had five $(CTTT)_{14-18}$ alleles, with the predominant $(CTTT)_{16}$ allele present on 60% haplotypes, indicating the source of the observed haplotype diversification (Table 3). The STR haplotypes in families showed a similar distribution of allelic sizes at this locus, but no haplotype breakdown in the region telomeric of *D6S2670* (data not shown), indicating that meiotic recombination cannot account for the *D6S2670* instability. Since distinct *D6S2670* alleles were associated with identical *HLA* specificities on the *HLA-A1, B*0801, DRB1*03011, DQB1*0201* haplotypes, the diversification observed at *D6S2670* could not be explained by an ancestral diversification in the analysed *HLA* genes flanking the unstable locus.

STR *vs* HLA haplotypes

With the exception of *D6S2670*, the overall rarity of variant STR alleles on the *HLA-A1, B*0801, DRB1*0301* haplotypes indicated that genotypes at multiple STRs may have a high predictive value for the determination of *HLA* specificities/haplotypes. To assess the STR variability on the most common haplotypes in the population, we have typed a panel of homozygous cell lines and donors (Table 1 and http://www.cbt.ki.se/fam/res/tab1.htm). These results showed that each *HLA* haplotype had a specific set of STR alleles on the chromosome and also suggested their low variability, as was observed for the *HLA-A1, B*0801, DRB1*03011, DQB1*0201* haplotype, although only a small number of homozygous carriers could be tested. At *D6S2670*, we did not observe clearly predominant alleles on some haplotypes (Table 2), suggesting a similar repeat instability at this locus and/or a lower LD on other haplotypes in this region. No other tested haplotype consisted of the same set of STR alleles as the *HLA-A1, B*0801, DRB1*03011, DQB1*0201* haplotype (Table 1 and http://www.cbt.ki.se/fam/res/tab1.htm).

The comparison of *HLA* and STR haplotypes (Table 1) also shows regions of lineage diversification for several *HLA* specificities. While the *HLA-B*15, DRB1*04* and *HLA-B*44, DRB1*04* haplotypes were found to diversify between *15-D3A* and *D6S1014*, the *HLA-B*07, DRB1*15* and *HLA-B*15, DRB1*15* haplotypes show distinct STR haplotypes in a more distal region closer to *D6S2670*. In contrast, the *HLA-B*44, DRB1*04* and *HLA-B*44, DRB1*07* lineages diversified closer to the *HLA-B*44* gene. Similarly, several *HLA-B*27*-containing haplotypes (*HLA-B*27, DRB1*01* and *HLA-B*27, DRB1*04*) evolved to exhibit distinct set of STR alleles on the chromosome very close to the *HLA-B* locus (Table 1 and data not shown).

Whereas single STR genotype could not reliably predict *HLA* specificities or haplotypes, multiple STR alleles on the chromosome were specific for a particular ancestral *HLA* haplotype (Table 1). For example, we observed that the 217-bp allele at *C1_4_1* was present on haplotypes carrying *HLA-B*44, -B*57* and *-B*15* alleles, but allelic sizes in three adjacent centromeric loci were specific for each of these lineages. Similarly, although a 100-bp allele at *TNFb* was predictive of each of the *HLA-B*08, -B*15 (B62)* and *-B*14 (B65)* specificities, the allelic size at the *62* STR marker could distinguish the *HLA-B*08* carriers

Table 2 Complex structure of the *D6S2670* locus

Allele number	*Allele sizes* (bp)	*Allele frequency*	$(CTTT)_n$	$(CTTTT)_n$	*CT deletion in (CTTT) repeat*	*Predominant* HLA *haplotype*
1	189	0.010	8	1	–	
2	190	0.007	7	2	–	
3	198	0.006	9	2	–	
4	201	0.005	11	1	–	
5	202	0.052	10	2	–	
6	204	0.002	11	2	+	
7	205	0.041	12	1	–	*HLA-B*07, DRB1*15*
8	206	0.012	11	2	–	
9	208	0.033	12	2	+	
10	209	0.055	13	1	–	
11	210	0.035	12	2	–	
12	212	0.015	13	2	+	
13	213	0.007	14	1	–	
14	214	0.077	13	2	–	*HLA-B*44, DRB1*04*
15	216	0.002	14	2	+	
16	217	0.040	15	1	–	*HLA-B*35, DRB1*01*
17	218	0.118	14	2	–	
18	221	0.045	16	1	–	
19	222	0.166	15	2	–	
20	225	0.027	17	1	–	
21	226	0.180	16	2	–	*HLA-B*08, DRB1*03, HLA-B*15, DRB1*04*
22	229	0.004	18	1	–	
23	230	0.056	17	2	–	
24	234	0.005	18	2	–	

Allelic frequencies were estimated in founders and married-ins of families with immunoglobulin A deficiency and common variable immunodeficiency using the GAS programme; *n*, number of tandem tetra- and pentanucleotide repeats.

Table 3 Tetranucleotide repeat component accounts for the length polymorphism at *D6S2670* on the *HLA-A1, B*0801, DRB1*03011, DQB1*0201* haplotypes

Allele size (bp)	*Number of observed haplotypes*	$(CTTT)_n$	$(CTTTT)_n$
234	1	18	2
230	1	17	2
226	24	16	2
222	11	15	2
218	3	14	2

from the remaining samples, suggesting a 2-bp expansion at *TNFa* on *HLA-B*08* haplotypes. Four adjacent loci centromeric of the *HLA-B* locus could then distinguish all tested haplotypes (Table 1 and http://www.cbt.ki.se/fam/res/tab1.htm).

Discussion

Stability of STR loci on HLA haplotypes

We have shown that the overall allelic variability at STR loci is very low in Swedish homozygous carriers of the *HLA-A1, B*0801, DRB1*03011, DQB1*0201* haplotypes, except for *D6S2670* (Table 1). This locus was exceptional in having five alleles on this extended haplotype, consistent with the observed high proportion of allelic variants on other haplotypes studied. Our results demonstrate a recent diversification of *HLA-A1, B*0801, DRB1*03011, DQB1*0201* haplotypes, most likely due to mutations in the tetranucleotide component of *D6S2670*, but not as a result of a meiotic recombination. The *D6S2670* STR is in intron 19 of the *TNXB* gene, which is located in the region with the highest GC content in the MHC, exceeding 53% (Figure 1).[2] The observed instability of this STR (Tables 1 – 3 and Figure 2) may thus be compositional and related to the GC content or a transition from a region with lower GC content to a higher GC content in the immediate vicinity of the repeat.[2] Although this region lies in the previously proposed H3 isochore with the GC contents of over 52%,[33] recent sequence analyses of the first draft of the human genome do not support the existence of such distinct homogenous distributions.[34]

There is a remarkable paucity of STRs between *D6S2670* and *9N2*, which is an area characterised by a high GC content[2] (Figure 1). However, the density of microsatellites does not appear to correlate with the GC content since the adjacent telomeric region is rich in microsatellites and has a similar GC content (Figure 1). A recent study described variability at the *D3A* and *9N1* loci on *HLA-B18-DR3* haplotypes, while the STR stability was generally higher in the telomeric region.[35,36] In our study, the *15-D3A STR*, which is also located in the region of high GC content about 90 kb centromeric of *D6S2670* (Figure 1), had a higher than average allelic variability as well, with three out of 40 variants in homozygous donors and 12% variants

in families. This proportion seems to be higher than that observed in Sardinian populations, but lower than in the UK population,[35,36] which may be due to different haplotypes analysed in these studies. It will be interesting to study the role of population- or haplotype-specific factors in the STR variability, but our results and results of others[35,36] suggest that both factors probably account for only smaller differences in the proportion of variant STR alleles and that most of the observed STR variability is inherent to a particular tandem repeat. The *D3A* locus also showed a higher pair-wise LD measures with the *TNF* STRs (Figure 1).

A low-to-high GC content transition was recently reported in the 3′ flanking region of the *NF1* gene, where very high LD was associated with a low GC content of about 37%. In contrast, a flanking region, exhibiting a marked LD decrease, had GC content over 50%.[37] It may be relevant in this respect that the affinity of *E. coli* RecA protein, which is essential for recombination, is influenced by the GC content of its template, suggesting a possible exclusion of GC-rich sequences from recombination *in vivo*.[38]

Microsatellites have generally a higher mutation rate than single nucleotide polymorphisms with an average of 10^{-3} per locus per gamete per generation, but this varies widely for STRs across the genome.[39] Therefore, the haplotype diversity is expected to be higher for STRs than for single nucleotide polymorphisms. The mutation rate increases with allele length, but does not seem to be affected by the size difference between an individual's two alleles.[40] Paternally transmitted mutations are in excess over those transmitted maternally,[39,40] consistent with our observation (Figure 2). The mutation rate of dinucleotide repeats was previously suggested to be lower than that of tetranucleotide repeats,[39,40] although *in vitro* studies[41] and indirect approaches relying on allelic frequencies in populations[42] do not support this. The interpretation of the STR variability in families may thus be facilitated by haplotype data, sex of transmitting parents and the increment and direction of the mutated repeat.

We observed a number of heterozygous genotypes in homozygous stretches in donor samples. This may be explained by genuine heterozygosity generated by recombination and not detected by low resolution *HLA* typing of our material. This was the case for donors previously typed as homozygous using serology as our sequence-based typing showed genuine heterozygosity in the *HLA* class II genes with an additional allele was found in five donor samples, consistent with the observed heterozygous STR genotypes (http://www.cbt.ki.se/fam/res/tab1.htm). Alternatively, heterozygosity in long genomic regions may also result from mutation or gene conversion events. Large autozygous regions detected in members of consanguineous families were found to be interrupted by occasional heterozygosity at STR loci, suggesting a recent diversification by mutation.[43]

Population diversity

Although little variability was observed for STR loci on Swedish haplotypes, population diversity is likely to increase STR diversity. The cell line VAVY from a French donor showed genotypes identical to Swedish *HLA-A1, B*0801, DRB1*03011* homozygous donor samples, except for the *AB59840 (C1_2_C)* and *D6S273* loci (http://www.cbt.ki.se/fam/res/tab1.htm). Similarly, a variant 142-bp allele at *D6S273* on the *HLA-B*08-DRB1*03011* haplotype, the largest of all alleles identified at this STR, was observed in one Northern Italian and one UK family, but never in large numbers of Swedish carriers of this haplotype. In contrast, the cell line AMALA of South American Indian origin carrying *HLA-B*1501* alleles had STR alleles identical to the *HLA-B*1501* positive samples in our study. Although identical *HLA* specificities were found with different STR genotypes, it was a set of STR alleles on the chromosome or a longer STR haplotype in the region sufficiently close to relevant *HLA* genes that gave clearly recognizable, specific patterns.

In a recent report,[44] the variability at *HLA* STRs was found to be higher for some haplotypes in non-Finnish donors as compared to Finnish donors for loci outside the *HLA-DR-DQ* region. However, apart from more than 10% mismatches at the *NOTCH4* locus amplifying a large PCR product, this proportion was close to zero in the flanking loci in unrelated Finnish donors.[44] This study also reported that three of five analysed *HLA-A*01, B*08, DR*03* haplotypes had a variant STR allele. Despite a higher number of the analysed *HLA-A1, B*08, DRB1*03* haplotypes and a higher number of STRs used in our study, the STR variability appeared to be lower for Swedish haplotypes. The higher proportion of the *HLA-A1, B*08, DRB1*03* haplotypes with variant STR alleles found in the Finnish study could be due to larger distances of the tested STRs from relevant *HLA* genes, haplotyping heterozygous family members or a different separation technique.[44] Garcia-Merino *et al*.[45] found only 2.5% variant alleles on Caucasoid *HLA-A1-B8-DR3* haplotypes at three *TNF* STRs, a proportion comparable to the present study.

Pattern of LD

Our study well illustrates marked differences in LD measures in the analysed area of over 1.5 Mb and the extraordinary position of the selection-driven *HLA-DQ/DR* region on the LD map (Figure 1). The course of LD appears to correspond well to published pair-wise data, with a peak LD across the *HLA-DQ-DR* region and marked drop in global LD just centromeric.[35,36] The absence of significant LD, previously observed for several pairs of class III loci such as *Bf-C4* or *C2-C4*,[36] was in a region of low LD measures in our study (Figure 1). A recent report showed little correspondence of D' with crossover frequency in a small area of the *TAP2* recombination hot spot using sperm typing,[46] suggesting that LD measures are a poor predictor of recombination frequencies in regions influenced by selection. Our map shows the

steepest decline of LD between the *HLA-DQB1* gene and *HLA-DQA2/DQB2* pseudogenes, which coincides well with previously identified first recombination hot spot in the class II region (Figure 1), while LD at *TAP2* already declined substantially.

HLA typing without HLA typing?

We have shown that major *HLA* haplotypes exhibit specific sets of STR haplotypes closely linked to the *HLA* genes if sufficient number of STR loci were typed. *HLA* haplotypes/specificities could be deduced from the STR haplotypes, which may be constructed using a very limited set of PCR reactions pooled and loaded onto a single sequencer lane. This is of interest for low-resolution/high-throughput routine matching and the detection and mapping of *HLA* recombinants. We have focused on the *HLA-A1, B*0801, DRB1*03011, DQB1*0201* haplotype, because its population prevalence is high (about 5.5% in Sweden, Bone Marrow Donors Worldwide at http://www.bmdw.org/) in order to obtain a sufficient number of homozygous cases. A small proportion of variant alleles/genotypes found on this haplotype may be comparable to that for other haplotypes, despite their lower analysed numbers. For example, in 10 samples carrying common *HLA-B*1501* specificities, only four out of 80 variant alleles (5%) were observed in four STR loci located within a 150-kb distance centromeric of the *HLA-B* gene. All the observed variants were limited to a single sample 12989 (http://www.cbt.ki.se/fam/res/tab1.htm), strongly suggesting heterozygosity in the region. However, the *HLA-A1, B*0801, DRB1*03011, DQB1*0201* haplotype may exhibit higher LD than other haplotypes, which could be reflected in a higher stability of STR haplotypes observed in this study. A detailed map of STRs in this region and reference allelic sizes will make it possible to address this issue in different populations using rich world-wide resources of DNA samples already typed for MHC specificities.

The composite population frequency of recognisable *HLA* haplotypes (Table 1) may constitute a large subset of all individuals in transplant matching programmes. For example, the haplotype frequency, as estimated from a sample of over 15 000 Swedish donors, was 3.2% for haplotype *HLA-A2-B62-DR4*, 1.2% for haplotype *HLA-A2, B8, DR3*, 2.6% for haplotype *HLA-A2, B44, DR4*, 1.1% for haplotype *HLA-A1, B57, DR7*, 3.2% for haplotype *HLA-A3, B7, DR15*, 2.4% for haplotype *HLA-A2, B7, DR15*, 2.1% for *HLA-A2, B60(40), DR6* and 2.4% for haplotype *HLA-A3, B35, DR1*, totalling to about a third of all observed haplotypes among Swedish prospective donors. These haplotypes all had a recognisable pattern of STR alleles (http://www.cbt.ki.se/fam/res/tab1.htm and Table 1).

In summary, we show a dense integrated STR/LD/gene map in the centromeric half of the human MHC, together with reference STR alleles on the most prevalent *HLA* haplotypes in a Caucasian population, their allelic frequencies in family founders, observed heterozygosity, PCR primers and PCR conditions. We found a remarkable conservation of STR alleles on the most common *HLA-A1-B*0801-DRB1*03011, DQB1*0201* haplotype in the vast majority of loci distal to the *DQB1* gene and characterised the unstable *D6S2670* locus in the *TNXB* gene, located in the region with the highest GC content in the MHC. The map and reference panel will facilitate population, evolution, case-control, linkage and linkage disequilibrium studies in this region.

Note added in proof

Matsuzaka et al *(*Tissue Antigens **56:** *492–500 and* **57:** *397–404) recently reported 22 novel polymorphic STRs in the class II region and 8 such loci in the class III region. However, of the 22 class II loci, three were described previously (M2_2_36 by Reiss* et al*:* Immunogenetics*;* **32:** *110–116; M2_2_23 in*[12] *and M2_4_25 in*[4]*). The M2_2_48 STR corresponds to the 13_36829 locus in this work. Eight STR primer pairs showed discrepancies between the observed and expected heterozygosity, suggesting the presence of null alleles. Eleven STRs had the observed heterozygosity below 50%. Similarly, of eight novel polymorphic STRs in the class III region, three were reported previously (N3_2_3 is identical to K11 in*[21]*/N3_2_4 corresponds to 22-38114 in*[3]*/N3_2_2 is the same as 9N1 or 82-2 in*[20]*). Thus, as in this study, a systematic search for suitable genetic markers yielded only few novel polymorphic STRs useful for gene mapping studies, suggesting that the map of HLA STRs is close to completion.*

Acknowledgments

This study was supported by the Karolinska Institute, the Primary Immunodeficiency Association of the United Kingdom and the Swedish Foundation for Strategic Research.

References

1 Weber JL, May PE: Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am J Hum Genet* 1989; **44**: 388–396.
2 Consortium: Complete sequence and gene map of a human major histocompatibility complex. *Nature* 1999; **401**: 921–923.
3 Vorechovsky I, Cullen M, Carrington M, Hammarstrom L, Webster AD: Fine mapping of IGAD1 in IgA deficiency and common variable immunodeficiency: identification and characterization of haplotypes shared by affected members of 101 multiple-case families. *J Immunol* 2000; **164**: 4408–4416.
4 Nair RP, Stuart P, Henseler T, *et al*: Localization of psoriasis-susceptibility locus PSORS1 to a 60-kb interval telomeric to HLA-C. *Am J Hum Genet* 2000; **66**: 1833–1844.
5 Foissac A, Salhi M, Cambon-Thomsen A: Microsatellites in the HLA region: 1999 update. *Tissue Antigens* 2000; **55**: 477–509.
6 Vorechovsky I, Zetterquist H, Paganelli R, *et al*: Family and linkage study of selective IgA deficiency and common variable immunodeficiency. *Clin Immunol Immunopathol* 1995; **77**: 185–192.
7 Vorechovsky I, Webster AD, Plebani A, Hammarstrom L: Genetic linkage of IgA deficiency to the major histocompatibility complex: evidence for allele segregation distortion, parent-of-origin penetrance differences, and the role of anti-IgA antibodies in disease predisposition. *Am J Hum Genet* 1999; **64**: 1096–1109.
8 Vorechovsky I, Luo L, Dyer MJ *et al*: Clustering of missense mutations in the ataxia-telangiectasia gene in a sporadic T-cell leukaemia. *Nat Genet* 1997; **17**: 96–99.

9 Pozzi S, Longo A, Ferrara GB: HLA-B locus sequence-based typing. *Tissue Antigens* 1999; **53**: 275 – 281.
10 Kotsch K, Wehling J, Blasczyk R: Sequencing of HLA class II genes based on the conserved diversity of the non-coding regions: sequencing based typing of HLA-DRB genes. *Tissue Antigens* 1999; **53**: 486 – 497.
11 Voorter CEM, Kik MC, van den Berg-Loonen EM: High-resolution HLA typing for the DQB1 gene by sequence-based typing. *Tissue Antigens* 1998; **51**: 80 – 87.
12 Beck S, Abdulla S, Alderton RP *et al*: Evolutionary dynamics of non-coding sequences within the class II region of the human MHC. *J Mol Biol* 1996; **255**: 1 – 13.
13 Ellis MC, Hetisimer AH, Ruddy DA *et al*: HLA class II haplotype and sequence analysis support a role for DQ in narcolepsy. *Immunogenetics* 1997; **46**: 410 – 417.
14 Beck S, Kelly A, Radley E, Khurshid F, Alderton RP, Trowsdale J: DNA sequence analysis of 66 kb of the human MHC class II region encoding a cluster of genes for antigen processing. *J Mol Biol* 1992; **228**: 433 – 441.
15 Carrington M, Dean M. A polymorphic dinucleotide repeat in the third intron of TAP1. *Hum Mol Genet* 1994; **3**: 218.
16 Macaubas C, Hallmayer J, Kalili J *et al*: Extensive polymorphism of a (CA)n microsatellite located in the HLA-DQA1/DQB1 class II region. *Hum Immunol* 1995; **42**: 209 – 220.
17 Lin L, L. J, Kimura A, Carrington M, Mignot E: DQ microsatellite association studies in three ethnic groups. *Tissue Antigens* 1997; **50**: 507 – 520.
18 Martin MP, Harding A, Chadwick, R., Kronick M *et al*: Characterization of 12 microsatellite loci of the human MHC in a panel of reference cell lines. *Immunogenetics* 1998; **47**: 131 – 138.
19 Dib C, Faure S, Fizames C *et al*: A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* 1996; **380**: 152 – 154.
20 Hsieh S-L, March R, Khanna A, Cross SJ, Campbell RD: Mapping of 10 novel microsatellites in the MHC class III region: application to the study of autoimmune disease. *J Rheumatol* 1997; **24**: 220 – 222.
21 Colonna M, Ferrara GB, Strominger J, Spies T: Hypervariable microsatellites in the central MHC class III region; in Tsuji K, Aizawa M, Sasazuki T (eds): *HLA 1991*. Oxford: Oxford University Press, 1991; 179 – 180.
22 Udalova IA, Nedospasov SA, Webb GC, Chaplin DD, Turetskaya RL: Highly informative typing of the human TNF locus using six adjacent polymorphic markers. *Genomics* 1993; **16**: 180 – 186.
23 Tamiya G, Ota M, Katsuyama Y *et al*: Twenty-six new polymorphic microsatellite markers around the HLA-B, -C and -E loci in the human MHC class I region. *Tissue Antigens* 1998; **51**: 337 – 346.
24 Mizuki N, Ota M, Kimura M *et al*: Triplet repeat polymorphism in the transmembrane region of the MICA gene: a strong association of six GCT repetitions with Behcet disease. *Proc Natl Acad Sci USA* 1997; **94**: 1298 – 1303.
25 Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES: Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* 1996; **58**: 1347 – 1363.
26 Sobel E, Lange K: Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet* 1996; **58**: 1323 – 1337.
27 Weeks DE, Sobel E, O'Connell JR, Lange K: Computer programs for multilocus haplotyping of general pedigrees. *Am J Hum Genet* 1995; **56**: 1506 – 1507.
28 Abecasis GR, Cookson WO: GOLD – graphical overview of linkage disequilibrium. *Bioinformatics* 2000; **16**: 182 – 183.
29 Hedrick PW: Gametic disequilibrium measures: proceed with caution. *Genetics* 1987; **117**: 331 – 341.
30 Cullen M, Noble J, Erlich H *et al*: Characterization of recombinants in the HLA class II region. *Am J Hum Genet* 1997; **60**: 397 – 407.
31 Begovich AB, McClure GR, Suraj VC *et al*: Polymorphism, recombination, and linkage disequilibrium within the HLA class II region. *J Immunol* 1992; **148**: 249 – 258.
32 Klitz W, Claiborne Stephens J, Grote M, Carrington M: Discordant patterns of linkage disequilibrium of the peptide-transported loci within the HLA class II region. *Am J Hum Genet* 1995; **57**: 1436 – 1444.
33 Bernardi G: Isochores and the evolutionary genomics of vertebrates. *Gene* 2000; **241**: 3 – 17.
34 Lander ES, Linton LM, Birren B *et al*: Initial sequencing and analysis of the human genome. *Nature* 2001; **409**: 860 – 921.
35 Herr M, Dudbridge F, Zavattari P *et al*: Evaluation of fine mapping strategies for a multifactorial disease locus: systematic linkage and association analysis of IDDM1 in the HLA region on chromosome 6p21. *Hum Mol Genet* 2000; **9**: 1291 – 1301.
36 Sanchez-Mazas A, Djoulah S, Busson M *et al*: A linkage disequilibrium map of the MHC region based on the analysis of 14 loci haplotypes in 50 French families. *Eur J Hum Genet* 2000; **8**: 33 – 41.
37 Eisenbarth I, Vogel G, Krone W, Vogel W, Assum G: An isochore transition in the NF1 gene region coincides with a switch in the extent of linkage disequilibrium. *Am J Hum Genet* 2000; **67**: 873 – 880.
38 Gruss A, Moretto V, Ehrlich SD, Duwat P, Dabert P: GC-rich DNA sequences block homologous recombination in vitro. *J Biol Chem* 1991; **266**: 6667 – 6669.
39 Weber JL, Wong C: Mutation of human short tandem repeats. *Hum Mol Genet* 1993; **2**:1123 – 1128.
40 Ellegren H: Heterogeneous mutation processes in human microsatellite DNA sequences. *Nat Genet* 2000; **24**: 400 – 402.
41 Schlotterer C, Tautz D: Slippage synthesis of simple sequence DNA. *Nucl Acids Res* 1992; **20**: 211 – 215.
42 Chakraborty R, Kimmel M, Stivers DN, Davison LJ, Deka R: Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proc Natl Acad Sci USA* 1997; **94**: 1041 – 1046.
43 Broman KW, Weber JL: Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. *Am J Hum Genet* 1999; **65**: 1493 – 1500.
44 Karell K, Klinger N, Holopainen P, Levo A, Partanen J: Major histocompatibility complex (MHC)-linked microsatellite markers in a founder population. *Tissue Antigens* 2000; **56**: 45 – 51.
45 Garcia-Merino A, Alper CA, Usuku K *et al*: Tumor necrosis factor (TNF) microsatellite haplotypes in relation to extended haplotypes, susceptibility to diseases associated with the major histocompatibility complex and TNF secretion. *Hum Immunol* 1996; **50**: 11 – 21.
46 Jeffreys AJ, Ritchie A, Neumann R: High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. *Hum Mol Genet* 2000; **9**: 725 – 733.