Am. J. Hum. Genet. 64:1709–1718, 1999

Tracing the Origin of HLA-DRB1 Alleles by Microsatellite Polymorphism

Tomas F. Bergström,[1] Hans Engkvist,[1] Rikard Erlandsson,[1] Agnetha Josefsson,[1] Steven J. Mack,[2,3] Henry A. Erlich,[2,3] and Ulf Gyllensten[1]

[1]Department of Genetics and Pathology, Unit of Medical Genetics, Beijer Laboratory, University of Uppsala, Uppsala, Sweden; [2]Department of Human Genetics, Roche Molecular Systems, Alameda, and [3]Children's Hospital, Oakland Research Institute, Oakland, California

Summary

We analyzed the origin of allelic diversity at the class II *HLA-DRB1* locus, using a complex microsatellite located in intron 2, close to the polymorphic second exon. A phylogenetic analysis of human, gorilla, and chimpanzee *DRB1* sequences indicated that the structure of the microsatellite has evolved, primarily by point mutations, from a putative ancestral $(GT)_x(GA)_y$-complex–dinucleotide repeat. In all contemporary *DRB1* allelic lineages, with the exception of the human **04* and the gorilla **08* lineages, the $(GA)_y$ repeat is interrupted, often by a G→C substitution. In general, the length of the 3′ $(GA)_y$ repeat correlates with the allelic lineage and thus evolves more slowly than a middle $(GA)_z$ repeat, whose length correlates with specific alleles within the lineage. Comparison of the microsatellite sequence from 30 human *DRB1* alleles showed the longer 5′ $(GT)_x$ to be more variable than the shorter middle $(GA)_z$ and 3′ $(GA)_y$ repeats. Analysis of multiple samples with the same exon sequence, derived from different continents, showed that the 5′ $(GT)_x$ repeat evolves more rapidly than the middle $(GA)_z$ and the 3′ $(GA)_y$ repeats, which is consistent with findings of a higher mutation rate for longer tracts. The microsatellite-repeat–length variation was used to trace the origin of new *DRB1* alleles, such as the new **08* alleles found in the Cayapa people of Ecuador and the Ticuna people of Brazil.

Introduction

HLA class II loci are characterized by remarkably high levels of polymorphism (Klein and Figueroa 1986; Kappes and Strominger 1988; Trowsdale 1995). For instance, *HLA-DRB1,* with >180 known alleles, and *DPB1,* with >70 alleles, are among the most polymorphic protein-coding regions of the human genome (Bodmer et al. 1997; Marsh 1997; American Society for Histocompatibility and Immunogenetics database). It has been proposed that much of the allelic diversity at the class II *DRB1* locus is ancient, with >50% of the alleles having arisen prior to the separation of human and chimpanzee, 4–7 million years ago (Klein 1987; Mayer et al. 1992; Ayala 1995; Ayala and Escalante 1996). The *DRB1* alleles can be divided into ~13 allelic lineages, on the basis of sequence similarities, and an analysis of intron sequence variability has indicated that, although most of the lineages diverged millions of years ago, the alleles within lineages are more recent (Bergström et al. 1998). However, the very limited amount of intron sequence variability present among alleles within a lineage provides little resolution of the relationships of individual alleles. To trace the origin of individual *DRB1* alleles, therefore, we have analyzed the sequence of a complex microsatellite located ~50 bp 3′ to the polymorphic second exon of the *DRB1* locus. The length polymorphism at microsatellite loci is generally believed to have been generated through slippage of the DNA polymerase during replication, followed by a misalignment of the displaced strands (Levinson and Gutman 1987). Variation in microsatellite–repeat length has generally been analyzed by use of family studies or, more recently, by analysis of sperm (Zhang et al. 1994). In the present study, the evolution of the structure of a complex microsatellite at the *DRB1* locus is analyzed by superimposing the microsatellite sequences onto a phylogenetic tree constructed from the surrounding intron sequences. The coalescence time of the allelic lineages at *DRB1* has been estimated at ~40 million years (Bergström et al. 1998). Thus, analysis of the microsatellite sequences on *DRB1* lineages provides an opportunity to study the evolution of a complex microsatellite sequence over these time intervals. Moreover, the pattern of length variability in samples from different populations provides insight into the evolution of modern human populations.

Received April 15, 1998; accepted for publication March 30, 1999; electronically published May 5, 1999.

Address for correspondence and reprints: Dr. Ulf Gyllensten, Department of Genetics and Pathology, Unit of Medical Genetics, Beijer Laboratory, University of Uppsala Biomedical Center, Box 589, 751 23 Uppsala, Sweden. E-mail: ulf.gyllensten@medgen.uu.se

0002-9297/99/6406-0025$02.00

Material and Methods

Samples and DNA Preparation

DNA was prepared from peripheral blood leukocytes, by standard phenol/chloroform extraction, and was collected by ethanol precipitation. In addition, chimpanzee genomic DNA, derived from two different individuals (Wodka and Brigitte), was generously provided by Dr. Ronald Bontrop (Slierendregt et al. 1993); additional primate DNA samples were taken from the collection of one the authors (U.G.).

Amplification System

An ~900-bp segment, containing 300 bp of intron 1 (the second exon) and ~300 bp of intron 2, was amplified with primers UG355 (GCG GTG CTG GAC GGA TCC TCC TC) and UG357 (TTC CCT TCC TTG CAT CTC TAA) (fig. 1). One μl from the first amplification was used as a template in a second PCR, with primers made to allele-specific motifs in the 5′ part of the second exon, together with the appropriate intron primers, to obtain an allele-specific PCR product for sequencing (fig. 1), and 5′ primers were designed for the specific amplification of exon 2 and intron 2 of the allelic lineages. These primers included an 18-bp segment, identical to the M13 sequencing primers (underlined in the sequences below): **01* (DR01REV [5′-CAG GAA ACA GCT ATG ACC CGT TTC TTG TGG CAG CTT AAG TT-3′]), **15/*16* (DR02REV [5′-CAG GAA ACA GCT ATG ACC CAC GTT TCC TGT GGC AGC CTA AGA GG]-3′); **03, *11, *13,* and **14* (DR3.5.6REV [5′-CAG GAA ACA GCT ATG ACC CAC GTT TCT TGG AGT ACT CTA CGT C-3′]); **04* (DR04REV [5′-CAG GAA ACA GCT ATG ACC CAC GTT TCT TGG AGC AGG TTA AAC A-3′])' **07* (DR07REV [5′-CAG GAA ACA GCT ATG ACC CAC GTT TCC TGT GGC AGG G-3′]); and **08* (DR08REV [5′-CAG GAA ACA GCT ATG ACC CAC GTT TCT TGG AGT ACT CTA CGG G-3′]); and were used with the 3′ primers UG357FRW (5′-TGT AAA ACG ACG GCC AGT TTC CCT TCC TTG CAT CTC TAA-3′) and UG358FRW (5′-TGT AAA ACG ACG GCC AGT AGG ATT CTA AAT GCT CAC AGA T-3′).

For amplification of intron 1, 3′ primers for alleles **01* (UG359 [TCC CAT TCA AGA AAT GAC ATT CAA A]); **15/*16* (UG360 [CCC ATT GAA GAA ATG ACA CTC CCT]); **04* (UG361 [CCC GTT GAA GAA ATG ACA CTC ATG]); **07* (UG362 [CGT CCC GTT GAA GAA ATG ACA CTT]), **08* (UG364 [CCC ATT GAA GAA ATA ACA CTC ACC]); and **03, *08, *11, *12, *13,* and **14* (UG365 [AAT GAC ACT CAG ACG TAG AG]) were used, in combination with the 5′ primer UG355 (GCG GTG CTG GAC GGA TCC TCC TC), which amplifies alleles from all *DRB1* allelic lineages.

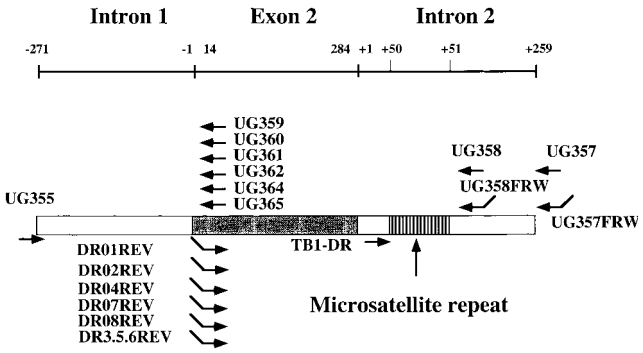

Figure 1 The experimental strategy employed to study the intron-sequence variation. The arrows represent primers used in amplification or sequencing reactions. The exon-2 primers are located in the Hyper Variable Region I, to allow allele-specific amplifications. The bent arrows (tails) in the primers are identical to the M13 sequencing primers.

We performed amplifications using 0.25–0.75 μg genomic DNA, in a 100-μl reaction containing 0.5 mM each primer, 50 mM KCl, 10 mM Tris-HCl (pH 8.3), 1.5 mM $MgCl_2$, 0.2 mM each dNTP, and 2.5 U *Taq* polymerase (Perkin-Elmer). A total of 35 cycles were run in the first PCR, each with denaturation at 94°C for 40 sec, annealing at 60°C for 40 sec, and primer extension at 72°C for 1 min. The second PCR, with 1 μl of the first PCR as template, was run with the same conditions, except that the primer extension done at 72°C was for 40 sec.

Sequencing

The allele-specific PCR products from the second PCR were gel-purified (Genomed) and sequenced with fluorescent cycle sequencing (AmpliTaq FS, Dye Terminator Cycle Sequencing Kit, Perkin-Elmer). For sequencing, the primers UG355, UG358-365, UGTB1-DR (ACA GTG CAG CGG CGA GGT GAG), and the Dye-labeled m13 sequencing primers, identical to the 5′ primer tails introduced in the second PCR, were used.

Phylogenetic and Statistical Analysis

We constructed unweighted pair-group method-of-analysis (Sneath and Sokal 1973) trees, from the intron sequences, using MEGA, which we obtained from pairwise genetic distance estimates. Estimates were corrected for multiple hits with the Jukes-Cantor method (Jukes and Cantor 1969).

Results

Mode and Rate of Microsatellite Structure Evolution

All of exon 2 and ~300 bp of introns 1 and 2, including the microsatellite, were examined for 30 different human *DRB1* alleles representing the allelic lineages **01, *03, *04, *07, *08, *11, *12, *13, *14, *15,* and **16* (for a total of 63 chromosomes). Almost no intron-sequence variation was detected among alleles within allelic lineages, outside the microsatellite region, consistent with findings from other analyses of a subset of these alleles (Bergström et al. 1998). The exceptions were the *DRB1*0408, *08032, *0806, *1301,* and **1303* alleles. One of the two **0408* alleles (from Scandinavia) differed, by a single point mutation in intron 2, from the other **04* alleles. The two *DRB1*08032* alleles (one each from Scandinavia and Asia) differed from other **08* alleles, by a single point mutation in intron 1. *DRB1*0806* appears to have been generated by a recombination event between a **11* and a **08* allele, which explains the intron differences reported by Bergström et al. (1998). Two **1301* alleles— one from Scandinavia and one from the Choco, a South-American population of African origin—were sequenced. The Choco **1301* and a Scandinavian **1302* were found to be identical in their intron sequences. The intron 2 sequence of the Scandinavian **1301* was identical to that of alleles from the **11* lineage, but differed from the Choco **1301* allele, as well as from the Scandinavian **1302* allele, by one substitution in intron 2. The *DRB1*1303* intron sequences were identical to those of all **11* alleles, indicating that the **1303* allele belongs to the **11* lineage. This relationship is also supported by the linkage of **1303* with *HLA-DQA1* and *-DQB1* alleles. *HLA-DRB1*1303* is linked with *DQA1*0501-DQB1*0301,* as are most *DRB1*11* haplotypes. In addition, the *DRB1*1303* sequence of the motif, at codons 25–35 in exon 2, is shared with *DRB1*11* alleles but not with most *DRB1*13* alleles.

In addition, the microsatellite sequence was determined from two gorilla alleles (*OR287-GogoB1*03* and *OR759-GogoB1*08/*12*), and from four chimpanzee alleles (*Wodka-PatrB1*02, 662-PatrB1*03, Debbie-PatrB1*0701,* and *Brigitte-PatrB1*0701*). Gorilla and chimpanzee alleles were classified on the basis of exon sequences (Gyllensten et al. 1991; Slierendregt et al. 1993).

The polymorphism in the intron 2 microsatellite displays a complex pattern, with both sequence and length variability between alleles (table 1). The simplest structure found is the $(GT)_x(GA)_y$, present in alleles from the **04* lineage. This structure is also found in a gorilla *DRB1*08* (*OR759-GogoB1*08/12*) microsatellite sequence (table 1). Because all other microsatellite structures can be derived from this pair of dinucleotide repeats, with a few substitutions, we assume that $(GT)_x(GA)_y$ is the ancestral microsatellite structure. The microsatellite structure in all other allelic lineages is more complex and can be subdivided into three parts: a 5′ $(GT)_x$ repeat, a 3′ $(GA)_y$ repeat, and a central region containing at least one dinucleotide repeat (i.e., the denoted middle $[GA]_z$ repeat, below). All alleles from an allelic lineage showed the same microsatellite structure (table 1). However, the microsatellite structure differed between each of the individual allelic lineages; thus, the general structure of the microsatellite was strongly associated with individual lineages (table 1).

To study the evolution of the microsatellite, we superimposed its structure, using parsimony considerations, onto the topology of a phylogenetic tree made on the basis of the intron 1 and intron 2 human-allele sequences, excluding the microsatellite repeat (Bergström et al. 1998; fig. 2). A minimum of 11 point mutations are required to generate the eight different human microsatellite structures present among the allelic lineages analyzed. Individual lineages have undergone 0–3 substitutions. In this model, all substitutions distinguishing allelic lineages, except in the **07* lineage, have occurred in the 3′ $(GA)_y$ repeat (fig. 2). The microsatellite sequence of the **0701* allele has been interrupted by a total of three point mutations, located in both the $(GT)_x$ and the $(GA)_y$ repeats. The topology of the phylogenetic tree indicates that the *DRB1*07* allelic lineage is the most ancient (fig. 2). On the basis of this topology, it may be argued that the microsatellite structure of the **07* lineage is the most ancestral. However, the **07* microsatellite structure is considerably more complex than that of most other lineages. Since, in other parts of the phylogenetic tree, the direction of change of the microsatellite structure is toward more complex patterns rather than more simplified, we find it more likely that the ancestral structure was a simple structure, of the type found in the contemporary **04* lineage, than of the type found in the **07* lineage.

On the basis of the microsatellite sequence, it appears that, of the two DR2 lineages (**15* and **16*), **16* was ancestral to **15,* since the **15* microsatellite structure can be derived from the **16* by two additional GA→CA changes (table 1 and fig. 2). We cannot formally exclude the possibility that the **16* lineage was derived from **15* by deletion of a $CA(GA)_nCA$ tract; however, a specific deletion seems less likely than the two-mutation model. Also, the tree indicates that the microsatellite structure of the **03, *11,* and **13* allelic lineages is derived from the **14* lineage. Among the allelic lineages, the **11* and **13* lineages appear to have diverged most recently. The length of the middle GA repeat also indicates that several *DRB1* alleles have been assigned a name that is inconsistent with their evolutionary origins.

Table 1

Microsatellite Variation among *DRB1* Alleles

Allele	Origin	No of Alleles Examined	5′-GT Repeat	Middle Repeat	3′-GA Repeat
0101	Eur	1	$(GT)_{16}$	AA GA AA	$(GA)_4$
0101	Eur	1	$(GT)_{18}$	AA GA AA	$(GA)_4$
0103	Eur	1	$(GT)_{16}$	AA GA AA	$(GA)_4$
03011	Eur	1	$(GT)_{17}$	$(GA)_6$ CA $(GA)_3$ CA	$(GA)_5$
GogoB1*03OR287	Gorilla	1	$(GT)_{14}$	$(GA)_5$ GG $(GA)_9$ CA	$(GA)_6$
PatrB1*03662	Chimpanzee	1	$(GT)_{17}$	$(GA)_4$ AA	$(GA)_7$
0401	Eur	2	$(GT)_{20}$	…	$(GA)_{16}$
0402	Eur	1	$(GT)_{22}$	…	$(GA)_{20}$
0404	Eur	1	$(GT)_{22}$	…	$(GA)_{19}$
0404	Sai	1	$(GT)_{20}$	…	$(GA)_{19}$
0407	Eur	1	$(GT)_{22}$	…	$(GA)_{16}$
0407	Sai	1	$(GT)_{22}$	…	$(GA)_{21}$
0408	Eur	2	$(GT)_{20}$	…	$(GA)_{16}$
0411	Sai	1	$(GT)_{21}$	…	$(GA)_{20}$
0411	Asi	1	$(GT)_{22}$	…	$(GA)_{19}$
0701	Eur	1	$(GT)_2$	GG TT $(GT)_7$ $(GA)_8$ GC	$(GA)_2$
PatrB1*0701/Debbie	Chimpanzee	1	$(GT)_2$	GG TT $(GT)_{14}$ $(GA)_4$ GC	$(GA)_2$
PatrB1*0701/Brigitte	Chimpanzee	1	$(GT)_2$	GG TT $(GT)_{10}$ $(GA)_7$ GC	$(GA)_2$
0801	Eur	5	$(GT)_{18}$	$(GA)_7$ CA	$(GA)_5$
0801	Sai	1	$(GT)_{18}$	$(GA)_7$ CA	$(GA)_5$
0801	Eur	1	$(GT)_{19}$	$(GA)_7$ CA	$(GA)_5$
08021*	Sai	3	$(GT)_{17}$	$(GA)_9$ CA	$(GA)_5$
08021	Nai	1	$(GT)_{15}$	$(GA)_9$ CA	$(GA)_5$
08021*	Eur	2	$(GT)_{17}$	$(GA)_9$ CA	$(GA)_5$
08032*	Eur	1	$(GT)_{21}$	$(GA)_7$ CA	$(GA)_5$
08032	Eur	1	$(GT)_{20}$	$(GA)_7$ CA	$(GA)_5$
08032*	Asi	2	$(GT)_{17}$	$(GA)_8$ CA	$(GA)_5$
0803	Aus[a]	1	$(GT)_{16}$	$(GA)_8$ CA	$(GA)_5$
08041	Afr	1	$(GT)_{15}$	$(GA)_8$ CA	$(GA)_5$
08041	Afr	2	$(GT)_{16}$	$(GA)_8$ CA	$(GA)_5$
08041	Afr	1	$(GT)_{17}$	$(GA)_8$ CA	$(GA)_5$
08042	Sai	1	$(GT)_{17}$	$(GA)_9$ CA	$(GA)_5$
0806	Eur[a]	1	$(GT)_{21}$	$(GA)_6$ CA	$(GA)_5$
0806	Afr	4	$(GT)_{21}$	$(GA)_6$ CA	$(GA)_5$
0807	Sai	1	$(GT)_{17}$	$(GA)_9$ CA	$(GA)_5$
0811	Nai[b]	1	$(GT)_{17}$	$(GA)_9$ CA	$(GA)_5$
GogoB1*0812/OR759	Gorilla	1	$(GT)_{14}$	…	$(GA)_{21}$
11011	Eur	1	$(GT)_{18}$	$(GA)_5$ CA $(GA)_3$ CA	$(GA)_6$
11011	Asi	1	$(GT)_{21}$	$(GA)_5$ CA $(GA)_3$ CA	$(GA)_6$
1102	Sai	1	$(GT)_{22}$	$(GA)_3$ CA $(GA)_3$ CA	$(GA)_6$
11041	Eur	1	$(GT)_{23}$	$(GA)_5$ CA $(GA)_3$ CA	$(GA)_6$
11041	Eur	1	$(GT)_{25}$	$(GA)_5$ CA $(GA)_3$ CA	$(GA)_6$
1103	Eur	1	$(GT)_{21}$	$(GA)_{10}$ CA $(GA)_3$ CA	$(GA)_6$
1201	Eur	1	$(GT)_{27}$	$(GA)_{10}$ $(CA)_2$	$(GA)_{10}$
1201	Eur	1	$(GT)_{27}$	$(GA)_{11}$ $(CA)_2$	$(GA)_{10}$
1301	Afr	1	$(GT)_{23}$	$(GA)_{10}$ CA $(GA)_3$ CA	$(GA)_6$
1301	Eur	1	$(GT)_{21}$	$(GA)_{10}$ CA $(GA)_3$ CA	$(GA)_6$
1302	Eur	1	$(GT)_{18}$	$(GA)_{11}$ CA $(GA)_3$ CA	$(GA)_6$
1303	Eur	1	$(GT)_{25}$	$(GA)_7$ CA $(GA)_3$ CA	$(GA)_6$
1401	Eur	1	$(GT)_{26}$	$(GA)_{11}$ CA	$(GA)_6$
1401	Asi	1	$(GT)_{24}$	$(GA)_{12}$ CA	$(GA)_6$
15011	Eur	1	$(GT)_{18}$	$(GA)_5$ CA $(GA)_4$ CA $(GA)_3$ GG AA	$(GA)_6$
15011	Eur	1	$(GT)_{19}$	$(GA)_5$ CA $(GA)_4$ CA $(GA)_3$ GG AA	$(GA)_6$
15021	Eur	1	$(GT)_{27}$	$(GA)_2$ CA $(GA)_4$ CA $(GA)_3$ GG AA	$(GA)_6$
15021	Asi	1	$(GT)_{18}$	$(GA)_5$ CA $(GA)_4$ CA $(GA)_3$ GG AA	$(GA)_6$
PatrB1*02/Wodka	Chimpanzee	1	$(GT)_{18}$	$(GA)_{14}$ CA $(GA)_4$ CA $(GA)_3$ GG AA	$(GA)_6$
1602	Eur	1	$(GT)_{17}$	$(GA)_8$ GG AA	$(GA)_6$
1602	Sai	4	$(GT)_{18}$	$(GA)_8$ GG AA	$(GA)_6$

NOTE—.Abbreviations: Eur = European, Nai = North American Indian, Sai = South American Indian, Afri = African, Aus = Australian, and Asi = Asian.

[a] No intron sequence available.

[b] Until recently, the *DRB1*0811* allele has been found only in North American Indians. This sample was collected from an African American of American Indian ancestry, and is indicative of recent admixture.

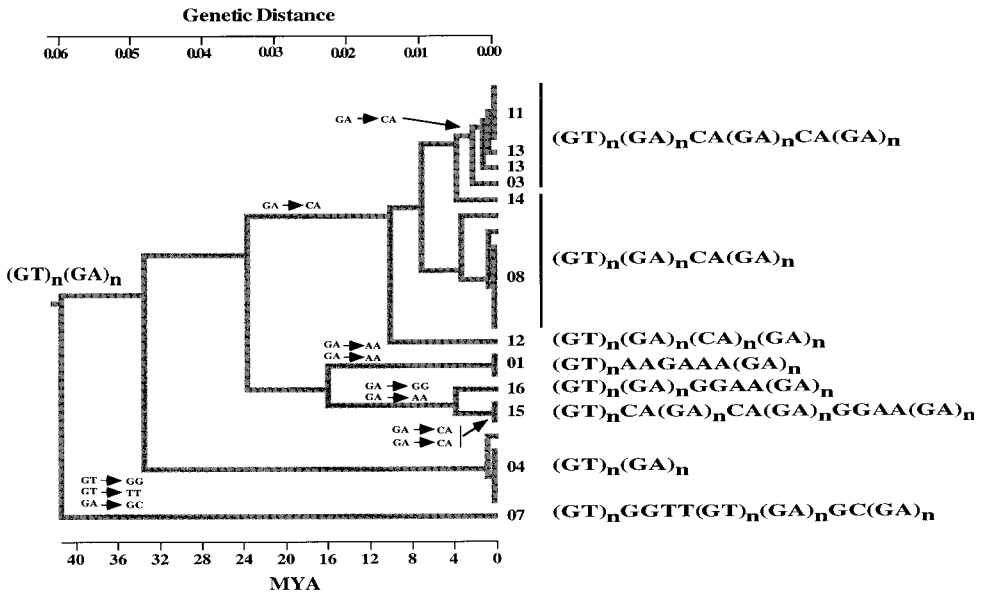

Figure 2 Evolution of the microsatellite structure. The phylogenetic tree is based on the combined sequences for intron 1 and 2, excluding the microsatellite, of 30 human *DRB1* alleles (Bergström et al. 1998). The tree was constructed, with use of the unweighted pair-group method-of-analysis method (Sneath and Sokal 1973), from the pairwise genetic distances calculated with the Jukes-Cantor correction for multiple hits (Jukes and Cantor 1969) and rooted by the midpoint method. The divergence times were estimated with the substitution rate for introns of $1.4 \times {}^{10-9}$/site/year (Li et al. 1996). The microsatellite sequences were superimposed onto this topology, and the inferred mutational events, done on the basis of parsimony considerations, are given in the tree. The contemporary microsatellite structure of allelic lineages (*right*) and the putative ancestral structure (*left*) are shown.

For instance, for both intron 1 and intron 2, the **1103* allele appears to belong to the **13* lineage rather than to the **11* lineage, whereas the **1303* allele belongs to the **11* lineage. As noted previously, the exon 2 sequence of **1303* is more similar to *DRB1*11* alleles than to most *DRB1*13* alleles.

The microsatellite sequences from the chimpanzee and gorilla support this evolutionary scenario. The gorilla **08* sequence has the simplest structure, $(GT)_x(GA)_y$, which differs from the human **08* microsatellite structure by a single point mutation that splits the 3′ $(GA)_y$ dinucleotide repeat (table 1). The chimpanzee *PatrB1*02* microsatellite allele has a structure identical to that of the *HLA-DRB1*15* alleles, which indicates that this structure is at least as old as the divergence between the two species, which occurred 4–7 million years ago. The structure of the *PatrB1*03* allele is quite similar to that of *HLA-DRB1*03* alleles, and the difference could be explained by two point mutations. The gorilla *GogoB1*03* structure is more similar to that of *HLA-DRB1*03* alleles and differs at only one dinucleotide repeat. Finally, the two *PatrB1*07* alleles have a microsatellite structure that is virtually identical to that of *HLA-DRB1*07* and is different only in the length of some of the dinucleotide repeats. This structure appears to predate the separation of the two species; the phylogenetic tree indicates that it is present on one of the most ancient allelic lineages. This analysis of nonhuman primate microsatellite sequences thus has confirmed the antiquity of some of the microsatellite structures and has provided support for the notion that the ancestral sequence may have had the structure $(GT)_x(GA)_y$.

The extent of length variability among alleles differed between the 5′ $(GT)_x$ and the 3′ $(GA)_y$ repeats. The 3′ $(GA)_y$-dinucleotide–repeat number is associated with specific allelic lineages, whereas the repeat number of the middle $(GA)_z$ tract correlates with specific alleles within an allelic lineage (table 1). This correlation may be explained by several different hypotheses, such as inherent differences in mutation rates among the different parts of the microsatellite or the functional involvement of the microsatellite in generating the exon 2 sequence polymorphism (see below). The 5′ $(GT)_x$ repeat shows extensive length variability among alleles, within an allelic lineage, as well as limited variability among samples with the same *DRB1* allele (see below), indicating that the 5′ repeat is evolving faster than the middle and 3′ repeats.

A possible explanation for the difference in amount of variability between the 3′ and 5′ parts of the microsatellite is that the mutation rate increases with the length of the uninterrupted tract. To evaluate this possibility, we compared multiple copies of the same exon allele from individuals of different geographic origin. We

examined multiple copies of each of 11 *DRB1* alleles (**0404, *0407, *0411, *0801, *0802, *08032, *11011, *1301, *1401, *15021,* and **1602*; table 1). Of these, all alleles except **0407,* showed variation in the 5′ $(GT)_x$ repeat; three alleles (**08032, *1401,* and **15021*) varied in the middle $(GA)_z$ repeat; and two alleles (**0407* and **0411*) were found to differ in the length of the 3′ $(GA)_y$ repeat (table 1). In general, the length of the 5′ $(GT)_x$ repeat is the longest, followed by the middle $(GA)_z$ repeat, and, finally, the 3′ $(GA)_y$ repeat. Thus, the pattern of mutational differences among multiple copies of the same allele is consistent with the notion that long, uninterrupted repeats, such as the 5′ $(GT)_x$ tract, are less stable than the shorter $(GA)_y$ tract.

Further support for decreased stability in long, uninterrupted repeats was seen in the **04* alleles, which are associated with the putative ancestral microsatellite structure $(GT)_x(GA)_y$, where x = 20–22 and y = 16–21. The **04* alleles represent an exception to all other *DRB1* alleles, in that the 3′ $(GA)_y$ region of the microsatellite is uninterrupted. For instance, if the variability is related mainly to the length of the dinucleotide–repeat tract then *both* the 5′ $(GT)_x$ and the 3′ $(GA)_y$ repeat of the **04* alleles would be expected to vary, as observed. Our data suggest that the middle and 3′ $(GA)_y$ regions, in the other *DRB1* alleles, are more stable, as a result of single bp substitutions interrupting the tracts. However, the difference in variability between the middle $(GA)_z$ and the 3′ $(GA)_y$ repeat cannot be attributed entirely to the repeat length, because these tracts have comparable lengths.

Tracing the Origin of DRB1*08 *Alleles*

The variability in the dinucleotide–repeat lengths can be used to infer phylogenetic relationships among alleles, including those within the same lineage. For example, all **08* alleles have the 3′ repeat $(GA)_5$. Among all other alleles tested, only *DRB1*0301* has this repeat length. However, the overall repeat structure for *DRB1*0301* is distinct from the **08* structure, in that the **03* microsatellite has a second G→C mutation interrupting the middle $(GA)_z$ repeat. Thus, the 3′ $(GA)_5$ appears to be a lineage-specific repeat and, presumably, has not changed in length since the divergence of the **08* lineage from others. However, the middle $(GA)_z$ repeat varies among different **08* alleles, and the length of this repeat constitutes an allele-specific marker. The **0801* alleles tested ($n = 7$) have a middle $(GA)_7$ repeat, **08021* ($n = 6$) has a $(GA)_9$ repeat, **08041* ($n = 5$) has a $(GA)_8$ repeat, and **0806* ($n = 5$) has a $(GA)_6$ repeat. The **08032* allele is the only exception to this allele-specific pattern; an **08032* allele of European origin has a middle $(GA)_7$ repeat, whereas a **08032* allele of Asian origin has a $(GA)_8$ repeat. Thus, in general, the middle $(GA)_z$ repeat does not appear to have been altered since the time when these *DRB1* alleles diverged from each other. The 5′ $(GT)_x$ repeat, as noted previously, evolves faster than the middle and 3′ $(GA)_y$ repeats, and may be useful in tracing the origin of populations, rather than of alleles or of lineages.

The association of the middle $(GA)_z$–repeat length with specific **08* alleles can be used to test hypotheses for the origin of recently discovered **08* alleles in Native American populations. These **08* alleles (1) could have been generated recently, in situ, from a putative parental allele (Titus-Trachtenberg et al. 1994; Mack and Erlich 1998); (2) could represent an allele from the ancestral population; or (3) could represent a more recent admixture. For example, the **08042* allele is found at a high frequency (*f*) of .05 only among the Cayapa of Ecuador. This allele has been postulated to have arisen from a point mutation in codon 86, from the pan–American Indian **08021* allele, subsequent to the colonization of South America (Titus-Trachtenberg et al. 1994). This inference was made on the basis of both the distribution of *DRB1* alleles in American Indian and other populations and on the DR-DQ linkage-disequilibrium pattern. The microsatellite of the **08042* allele—like that of the **08021* allele, a middle $(GA)_9$ repeat—has been consistent with this hypothesis. Similarly, the newly discovered **0807* allele, which differs from **08021* by a single point substitution in codon 57, and has been observed at very high frequency ($f = .23$) among the Ticuna of Brazil (Mack and Erlich 1998), as well as in other native South American groups, has been postulated to have been generated by a point mutation at **08021*. The microsatellite of the **0807* allele also has a middle $(GA)_9$ repeat. In addition, the **0811* allele, which also differs from the **08021* by a single substitution at codon 57, has a middle $(GA)_9$ repeat. This allele has been found recently in the Na-Denè–speaking Native American populations, the Navajo in the Southwest United States (Williams et al. 1994) and the Tlingit of Southeast Alaska (Smith et al. 1996). This pattern is consistent with the hypothesis that the *DRB1*08021* allele represents the ancestral American Indian **08* allele. The **0807* and **0811* alleles have been derived more recently (i.e., since humans colonized the Americas), from the **08021* allele, by point mutations or gene-conversion events involving codon 57, whereas the **08042* allele has been generated by a point mutation in codon 86.

The middle $(GA)_z$ repeat can also be used to distinguish between in situ origin of alleles and other explanations for rare alleles, in American Indian groups, such as recent admixture. The microsatellite of the single *DRB1*0801* allele, found in the Ticuna, has a middle $(GA)_7$ repeat, which is identical to Caucasian and African **0801* alleles but distinct from the repeat length of the **08021* allele. The origin of this allele was ten-

tatively attributed to admixture, because of its absence from other American Indian groups, rather than to generation from **08021* (Mack and Erlich 1998). The microsatellite sequence supports admixture rather than local generation of the **0801* allele from **08021*. A possible scenario for the microsatellite and exon changes that have taken place in the **08* lineage in the Americas and elsewhere is shown in figure 3. The length of the 5′ $(GT)_x$ repeat in the **08* lineage can also differ among samples with the same allele and may serve as a population marker. For instance, all American Indian **08* alleles (**08042*, **0807*, and **0811*), of putatively recent origin, have the 5′ $(GT)_{17}$-repeat length, as do the proposed parental Cayapa and Ticuna **08021* alleles. However, the Havasupai (Arizona) **08021* allele has the 5′ $(GT)_{15}$–repeat length, suggesting that the length of this repeat has changed since the Havasupai population became isolated from other American Indian groups.

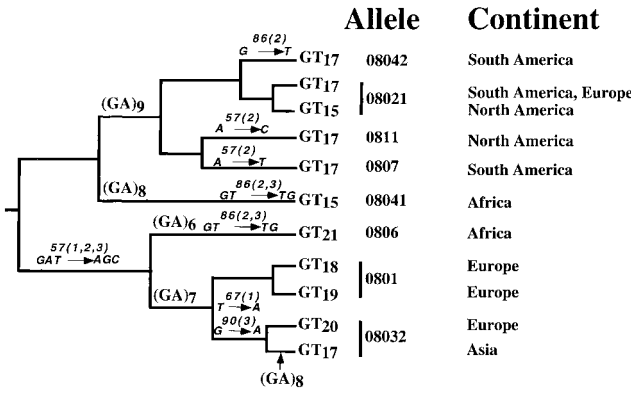

Figure 3 Evolutionary scenario for the **08* alleles, based on the exon-2 sequences and the microsatellite repeats. The topology is from the exon tree of Bergström et al. (1998). The parts in italics indicate inferred changes in exon 2. The repeat length is shown in boldface. The positions of the nucleotide substitutions are indicated by an arrow. The codon is shown above the arrow, and the numbers in parentheses indicate the position in the codon that has changed. The **0802* allele ($[GA]_9$) and the **0801* allele ($[GA]_7$) are equally likely to be the root in this model. This figure assumes that the **0802* allele is ancestral.

Discussion

Mechanism for the Generation of New Alleles

We have used a phylogenetic analysis to study the evolution of a complex microsatellite in intron 2 of the *HLA-DRB1* locus. The deep evolutionary roots of the *DRB1* allelic lineages provide an opportunity to study the accumulation of point mutations, as well as variation in repeat length, over the estimated coalescence time for lineages of 40 million years (Bergström et al. 1998). The microsatellite has evolved from the putative $(GT)_x(GA)_y$ ancestral structure in three main directions: (1) by interruptions of the $(GA)_y$ repeat, at various locations, by point mutations (such as is the case in the **01, *07, *08, *11, *12, *13, *14,* and **15* alleles); (2) by modification of the $(GT)_x$ repeat (such as in **0701*); and (3) by alterations in $(GT)_x$– or $(GA)_y$–repeat length (**04*). The putative rate of point mutations altering the microsatellite structure was found to be slightly, but not significantly, lower than the rate estimated for surrounding intron sequences. Thus, in contrast to repeat–length variability, point mutations in the microsatellite do not appear to occur at an elevated rate.

The 5′ $(GT)_x$ repeat showed greater variability among alleles than the 3′ $(GA)_y$ repeat. This difference in microsatellite variability could reflect a mutational bias that is related to repeat length. Comparison of the microsatellite repeat–length variability, among multiple copies of the same exon allele, revealed a three- to fivefold higher variation in 5′ $(GT)_x$–repeat length than in 3′ $(GA)_y$ repeat. The 5′ $(GT)_x$ repeat is generally the longest of the dinucleotide repeats, which suggests that the difference in stability is because of the length of uninterrupted dinucleotide repeats. Similar results have been obtained from sequence analysis, of orthologous microsatellite sequences, between human and chimpanzee (Blanquer-Maumont and Crouau-Roy 1995; Garza et al. 1995; Crouau-Roy et al. 1996). For example, the complex microsatellite *D4S404* was found to be highly conserved between species, in three regions of short (CA) tracts (ranging in size from 2–4 repeat units), whereas a longer (CA) tract (8–12 repeat units) was variable, both between and within species (Crouau-Roy et al. 1996). A microsatellite (*D4S885*), showing higher heterozygosity in humans than in chimpanzee, was found to contain an uninterrupted $(AC)_{22}$ tract in the shortest human allele sequenced. However, in chimpanzee, the same microsatellite region was interrupted, resulting in the structure $(AC)_{10–12}(AT)(AC)_2$. Similarly, microsatellites near the *HLA-DQ* locus (Jin et al. 1996; Macaubas et al. 1997) and tumor necrosis genes (*TNFα* and *TNFβ*; Blanquer-Maumont and Crouau-Roy 1995) were found to show higher stability in interrupted tracts.

In general, microsatellite polymorphism is thought to have been generated by slippage of the DNA polymerase (Levinson and Gutman 1987). An interruption of long $(NN)_n$ tracts may increase the fidelity of DNA polymerase, and thus result in the increased stability seen in mammalian complex microsatellites (Wierdl et al. 1997). In yeast it has been demonstrated that microsatellite instability is dependent on the length of the repeat units. Repeats of 33 and 51 bp were found to be 20–60 times more unstable than repeats of 15 bp (Wierdl et al. 1997). Also, the introduction of a single-variant repeat was found to stabilize a poly-GT run about fivefold, resulting in a lower rate of new mutations (Petes et al. 1997).

Since the same effect was seen in strains with mutations in *pms1, msh2,* and *msh3,* the stability observed after introducing a single-variant repeat cannot be because of the DNA-mismatch–repair system. The length of uninterrupted tracts cannot explain, however, all the differences in variability within the *DRB1* microsatellite. For instance, the increased variability in the middle $(GA)_z$ repeat, relative to the 3′ $(GA)_y$ repeat, cannot be attributed to repeat length, because they are similar in length. Perhaps there is a 5′-to-3′–gradient in variability, as postulated by Jeffreys et al. (1994), for minisatellites, within the *DRB1* microsatellite.

The Generation of DRB1 *Alleles and Human Origins*

The microsatellite was shown to be useful for tracing the origin of exon sequences, as was demonstrated by the identification of a putative ancestral sequence for the *DRB1*08042, *0807,* and **0811* alleles. Surprisingly, different parts of the microsatellite correlated with either specific lineages or specific alleles within a lineage (e.g., **08*); the middle $(GA)_z$ repeat correlated with specific alleles (e.g., $[GA]_7$ with **0801*), whereas the 3′ $(GA)_5$ repeat was restricted to **08* and the related lineage, **03*. The reason for this striking pattern is not clear. Perhaps these allele and lineage associations simply reflect differences in the underlying rate of diversification of the different $(GA)_n$ repeats. According to this "rate-coincidence model," the 3′ $(GA)_5$ repeat changed in length in the period after the **08* lineage diverged from the others (within the last few million years), whereas the middle $(GA)_z$ repeat changed in length during the period in which the alleles diversified (within the last few 100,000 years). An alternative model invokes some unknown recombination mechanism to explain the correlation; perhaps the putative gene-conversion event that generates exon 2 diversity (Bergström et al. 1998) may also somehow alter the length of the middle $(GA)_z$ repeat. This model, however, is hard to reconcile with the observation that gene conversion does not appear to extend into the intron (Bergström et al. 1998).

Microsatellite markers flanking a coding sequence may also be used to distinguish between single or multiple origins of an allele within a particular coding sequence. In principle, samples from different geographic regions, with the same exon 2 sequence, could have resulted from a single origin and migration or, alternatively, from multiple independent origins. To examine the possibility of multiple origins for individual *DRB1* alleles, we examined the intron sequences of multiple copies of alleles with identical exon sequences, sampled from individuals of different continents (table 1). The allelic lineages have highly diverged intron sequences (Bergström et al. 1998), as well as pronounced variation in the microsatellite structure (table 1). If alleles with identical exon sequences have been derived from different allelic lineages, this should be reflected by the flanking intron sequences including the microsatellite structure. Previously, a single allele (**0806*) has been found that appears to have been generated by a recombination event between alleles from different allelic lineages (Bergström et al. 1998). Thus, analysis of intron sequences is useful for detecting alleles that have originated through such recombination events.

All multiple copies of the same allele from different populations had intron sequences consistent with the hypothesis that they were derived from the same lineage. The microsatellite sequences were also consistent with the notion that alleles have been generated from the same allelic lineage, because different copies of alleles with the same exon 2 sequence did not carry different microsatellite structures. Using additional microsatellite sequences (Epplen et al. 1997), we never found the same exon 2 sequences in combination with different microsatellite structures. Thus, our results provide no evidence for the notion that alleles with identical exon sequences have been generated repeatedly by recombination between different allelic lineages.

Since only alleles resulting from an exchange *between* lineages could be unambiguously identified, new alleles, generated either by point mutations or gene-conversion events within the second exon, or by recombination events between alleles from the *same* allelic lineage, could go undetected. Comparisons of alleles within lineages indicate that some of the **04* alleles represent cases of potential multiple origins; a **0404* allele from Europe has the $(GT)_{22}(GA)_{19}$–repeat sequence, whereas the **0404* allele from South America has the $(GT)_{20}(GA)_{19}$ sequence. This microsatellite difference may reflect independent origins of the **0404* allele, which has been suggested by Mack and Erlich (1998), to explain the diversity of **04* alleles in the Americas. Alternatively, the microsatellite sequence could have diverged after the migration from a single **0404* source. We also note that two pairs of different **04* alleles have the same microsatellite sequence: **0401* (Europe) and **0408* (Europe) are both $(GT)_{20}(GA)_{16}$. Perhaps **0408* was derived from *0401, by an exchange at codon 70. Alternatively, the similarity could be fortuitous. Similarly, **0404* (Europe) and **0411* (Asia) are both $(GT)_{22}(GA)_{19}$. Also, the European **08032* and the Asian and Australian **08032* alleles differ in the middle GA repeat—$(GA)_7$ versus $(GA)_8$. This observation could reflect independent origins for the exon 2 sequence or an addition of a middle GA repeat in the microsatellite subsequent to migration or expansion. In the first scenario, the $(GA)_8$ **08032* might have been generated from another **08* allele with $(GA)_8$. *DRB1*08041,* the only other $(GA)_8$ allele, differs, at multiple exon 2 positions, from **08032* and thus does not seem a likely parental allele. In the second, and per-

haps more likely, scenario, the $(GA)_8$ **08032* allele was generated from the $(GA)_7$ **08032*. Thus, the analysis of the polymorphic complex microsatellite, 50 bp 3′ of the second exon of *DRB1,* provides an opportunity to examine the evolution of microsatellite sequences as well as to trace the origin of individual *DRB1* alleles.

Acknowledgments

We are grateful to Ray Apple, Elizabeth Trachtenberg, and Dory Bugawan for DNA samples, and to Hans Ellegren for comments on earlier versions of the manuscript. This work was supported by grants from the Beijer Foundation, the Swedish Natural Sciences Research Council, the Marcus Borgström Foundation, the Erik Philip Sörensen Foundation, and the National Institutes of Health.

Electronic-Database Information

The URL for data in this article is as follows:

American Society for Histocompatibility and Immunogenetics database, http://www.swmed.edu/home_pages/ASHI/sequences/drbdna.txt (for a fully updated compilation of class I and class II alleles)

References

Ayala FJ (1995) The myth of Eve: molecular biology and human origins. Science 270:1930–1936

Ayala FJ, Escalante AA (1996) The evolution of human populations: a molecular perspective. Mol Phylogenet Evol 5: 188–201

Bergström TF, Josefsson A, Erlich HA, Gyllensten U (1998) Recent origin of HLA–DRB1 alleles and implications for human evolution. Nat Genet 18:237–242

Blanquer-Maumont A, Crouau-Roy B (1995) Polymorphism, monomorphism, and sequences in conserved microsatellites in primate species. J Mol Evol 41:492–497

Bodmer JG, Marsh SG, Albert ED, Bodmer WF, Bontrop RE, Charron D, Dupont B, et al (1997) Nomenclature for factors of the HLA system, 1996. Tissue Antigens 49:297–321

Crouau-Roy B, Service S, Slatkin M, Freimer N (1996) A fine-scale comparison of the human and chimpanzee genomes: linkage, linkage disequilibrium and sequence analysis. Hum Mol Genet 5:1131–1137

Epplen C, Santos EJM, Guerreiro JF, Helden P, Epplen JT (1997) Coding versus intron variability: extremely polymorphic HLA-DRB1 exons are flanked by specific composite microsatellites even in distant populations. Hum Genet 99: 399–406

Garza JC, Slatkin M, Freimer NB (1995) Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. Mol Biol Evol 12:594–603

Gyllensten U, Sundvall M, Ezcurra I, Erlich HA (1991) Genetic diversity at class II *DRB* loci of the primate MHC. J Immunol 146:4368–4376

Jeffreys AJ, Tamaki K, MacLeod A, Monckton DG, Neil DL, Armour JAL (1994) Complex gene conversion events in germline mutation at human minisatellites. Nature Genetics 6:136–145

Jin L, Macaubas C, Hallmayer J, Kimura A, Mignot E (1996) Mutation rate varies among alleles at a microsatellite locus: phylogenetic evidence. Proc Natl Acad Sci USA 93: 15285–15288

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–32

Kappes D, Strominger JL (1988) Human class II major histocompatibility complex genes and proteins. Ann Rev Biochem 57:991–1028

Klein J (1987) Origin of major histocompatibility complex polymorphism: the trans-species hypothesis. Hum Immunol 19:155–62

Klein J, Figueroa F (1986) Evolution of the major histocompatibility complex. Crit Rev Immunol 6:295–386

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4:203–221

Li WH, Ellsworth DL, Krushkall J, Chang BH, Hewitt-Emmett DH (1996) Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. Mol Phylogenet Evol 5:182–187

Macaubas C, Jin L, Hallmayer J, Kimura A, Mignot E (1997) The complex mutation pattern of a microsatellite. Genome Res 7:635–641

Mack SJ, Erlich HA (1998) HLA class II polymorphism in the Ticuna of Brazil: Evolutionary implications of the DRB1*0807 allele. Tissue Antigens 51:41–50

Marsh SG (1997) Nomenclature for factors of the HLA system. Tissue Antigens 50:207

Mayer WE, O'hUigin C, Zaleska-Rutczynska Z, Klein J (1992) Trans-species origin of MHC-DRB polymorphism in the chimpanzee. Immunogenetics 37:12–23

Petes TD, Greenwell PW, Dominska M (1997) Stabilization of microsatellite sequences by variant repeats in the yeast Saccharomyces cerevisiae. Genetics 146:491–498

Slierendregt BL, Kenter M, Otting N, Anholts J, Jonker M, Bontrop RE (1993) Major histocompatibility complex class II haplotypes in a breeding colony of chimpanzees (*Pan troglodytes*). Tissue Antigens 42:55–61

Smith AG, Nelson JL, Regen L, Guthrie LA, Donadi E, Mickelson EM, Hansen JA (1996) Six new DR52-associated DRB1 alleles, three of DR8, two of DR11, and one of DR6, reflect a variety of mechanisms which generate polymorphism in the MHC. Tissue Antigens 48:118–126

Sneath PHA, Sokal RR (1973) Numerical taxonomy. Freeman, San Francisco

Titus-Trachtenberg EA, Rickards O, De Stefano GF, Erlich HA (1994) Analysis of HLA class II haplotypes in the Cayapa Indians of Ecuador: a novel DRB1 allele reveals evidence for convergent evolution and balancing selection at position 86. Am J Hum Genet 55:160–167

Trowsdale J (1995) "Both man & bird & beast": comparative organization of MHC genes. Immunogenetics 41:1–17

Wierdl M, Dominska M, Petes TD (1997) Microsatellite instability in yeast: dependence on the length of the microsatellite. Genetics 146:769–779

Williams TM, Wu J, Foutz T, McAuley JD, Troup GM (1994) A new DRB1 allele (DRB1*0811) identified in Native Americans. Immunogenetics 40:314

Zhang L, Leeflang EP, Yu J, Arnheim N (1994) Studying human mutations by sperm typing: instability of CAG trinucleotide repeats in the human androgen receptor gene. Nat Genet 7:531–535