Species-Specific Class I Gene Expansions Formed the Telomeric 1 Mb of the Mouse Major Histocompatibility Complex

Toyoyuki Takada,[1,4] Attila Kumánovics,[2] Claire Amadou,[1] Masayasu Yoshino,[1] Elsy P. Jones,[2] Maria Athanasiou,[3,5] Glen A. Evans,[3,6] and Kirsten Fischer Lindahl[1,2,7]

[1]Howard Hughes Medical Institute, [2]Center for Immunology, and [3]McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA

We have determined the complete sequence of 951,695 bp from the class I region of *H2*, the mouse major histocompatibility complex (*Mhc*) from strain 129/Sv (haplotype *bc*). The sequence contains 26 genes. The sequence spans from the last 50 kb of the *H2–T* region, including 2 class I genes and 3 class I pesudogenes, and includes the *H2–M* region up to *Gabbr1*. A 500-kb stretch of the *H2–M* region contains 9 class I genes and 4 pseudogenes, which fall into two subfamilies, *M1* and *M10*, distinct from other mouse class I genes. This *M1/M10* class I gene-cluster is separated from the centromeric *H2–T* and the telomeric *H2–M4*, *–5* and *–6* class I genes by "nonclass I genes". Comparison with the corresponding 853-kb region of the human *Mhc*, which includes the *HLA–A* region, shows a mosaic of conserved regions of orthologous nonclass I genes separated by regions of species-specific expansion of paralogous *Mhc* class I genes. The analysis of this mosaic structure illuminates the dynamic evolution of the *Mhc* class I region among mammals and provides evidence for the framework hypothesis.

[Supplemental material is available online at www.genome.org. The sequence data from this study have been submitted to GenBank under accession nos. AC005413, AC005665, AF532111–AF532117. A preliminary draft sequence was earlier submitted as AC002615 and replaced this year by NT002615.]

H2, the major histocompatibility complex (*Mhc*) on mouse Chr 17, was originally defined by transplantation experiments as the key to tumor and tissue graft compatibility and growth (Snell and Higgins 1951). Since then, an *Mhc* has been identified in all jawed vertebrates studied (Trowsdale 1995), and it is the most gene-dense region in the human genome (MHC Sequencing Consortium 1999). The human (MHC Sequencing Consortium 1999), chicken (Kaufman et al. 1999) and quail (Shiina et al. 1999a) *Mhc*s have all been sequenced, and efforts are underway to clone and sequence the *Mhc* of mouse (Kumánovics et al. 2003), cat (Beck et al. 2001), pig (Chardon et al. 2001; Renard et al. 2001), rabbit (Rogel-Gaillard et al. 2001), rat (Günther and Walter, 2001), cow (Di Palma et al. 2002), medaka (Matsuo et al. 2002), zebrafish (Kuroda et al. 2002), Japanese pufferfish (Clark et al. 2001) and many other *Mhc*s.

The prototypical mammalian *Mhc* encodes highly polymorphic transmembrane proteins, the class I and class II MHC molecules, that bind peptide fragments derived from intra- and extracellular proteins, respectively, and present them on the cell surface for surveillance by T lymphocytes (Yewdell and Bennink 2001; Lennon-Duménil et al. 2002). Several other proteins involved in antigen processing and peptide loading and transport are also encoded in the *Mhc*, as are components of the complement system and several lymphokines. However, about half the genes in this ~3.6-Mb (megabase) complex have no obvious immune function (Beck and Trowsdale, 2000).

Comparative physical mapping revealed extensive conservation of synteny within the *Mhc*, but in a patchwork pattern: so-called framework regions of near complete conservation of content and orientation of genes, mainly of nonimmune function, alternating with regions of species-specific expansion of rapidly evolving MHC class I and, to a lesser extent, class II genes (Amadou 1999; Amadou et al. 1999). All mammals have one or more polymorphic MHC class Ia genes that function in classical antigen presentation. The investigated species have another half dozen or so nonclassical, or class Ib, genes; rats and mice, however, have on the order of 30 or more intact class Ib genes (Amadou et al. 1999; Günther and Walter, 2001). The class Ib genes are generally distinguished by their low polymorphism and often have a limited tissue expression; the proteins they encode usually have shorter cytoplasmic tails, and some lack consensus residues associated with peptide binding (Stroynowski and Fischer Lindahl 1994). While several class Ib molecules are now known to bind peptides, their function is fully understood in only a few cases. HLA-E and its mouse homolog, Qa1 (encoded by *H2-T23*), display leader peptides from class Ia molecules and

Present addresses: [4]Laboratory Animal Science, The Tokyo Metropolitan Institute of Medical Science, 3-18-22, Honkoagome, Bunkyo-ku, Tokyo 113-8613, Japan; [5]Genaissance Pharmaceuticals, Five Science Park, New Haven, Connecticut 06511, USA; [6]Egea Biosciences, Inc., 6759 Mesa Ridge Road, Suite 100, San Diego, California 92121, USA
[7]Corresponding author.
E-MAIL kfl@chop.swmed.edu; FAX (214) 648-5453.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.975303.

13:589–600 ©2003 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/03 $5.00; www.genome.org

binds to inhibitory natural killer cell receptors (Yeager et al. 1997; Braud et al. 1998). The human *MICA* and *MICB* are induced by stress and are ligands for the activating immune receptor NKG2D (Spies 2002). Rodents have no MIC genes in the *Mhc* (Bahram et al. 1994), but they have other class Ib molecules not present in human, such as *M3*, which presents N-formylated peptides to cytotoxic T lymphocytes. M3 provides early protection against bacterial infection, and it can also serve as a histocompatibility antigen by presenting N-formylated polymorphic peptides from mitochondria (Fischer Lindahl et al. 1997).

The classical part of the mouse *Mhc,* including the *H2-K* and *H2-D* class I regions, were cloned and sequenced from strain 129 (haplotype *bc*; Kumánovics et al. 2002). The class I region of *H2* contains more than 60 *Mhc* class I genes and pseudogenes, in about 2 Mb. It is traditionally divided into the *H2-Q, H2-T*, and *H2-M* regions, which were originally defined by recombination. The centromeric 1 Mb of the *H2-M* region is rich in class I genes, whereas the telomeric 1 Mb is rich in olfactory receptor genes and contains just two functional class I genes, *H2-M3* and *H2-M2* (Amadou et al. 1999). Here we report the finished sequence of the class I-rich part of the *H2-M* region.

RESULTS

Genomic Sequencing of the Centromeric *H2–M* Region

We previously constructed a physical map of the *H2-M* region on Chr 17, based on a contig of bacterial artificial chromosome (BAC) clones from strain 129/SvJ (haplotype *bc*; Jones et al. 1999). Eight clones that should span the centromeric *H2-M* region were selected from this BAC contig. In the course of sequencing, it became clear that clones 585c7 and 10i1 did not overlap. The bridging 26,962 bp *Xho*I fragment from clone 255d16 (Jones et al. 1999) was sequenced to fill that gap.

The final 951,695 bp sequence assembly was verified in several ways, and no inconsistency was found. First, a sequence-based, virtual restriction enzyme digest was checked against restriction digests of the individual clones (data not shown); second, the position and sequence of previously mapped genes, pseudogenes, and STS markers were checked (Fig. 1). These markers included eight *H2-M* class I genes (*M1* and *M4-M10*) and six nonclass I genes (*Zfp173* or *Trim26*, *Trim10* or *Herf1*, *Mog*, and *Tctex4*, *-5*, and *-6*), two microsatellite markers (*D17Mit47* and *D17Mit125*), and all the suitable, nonrepetitive, BAC end sequences determined in the previous mapping study (Jones et al. 1999).

Sequence Composition

The sequenced region contains three isochores (Fig. 1). Isochores are long (>100 kb) stretches of DNA with uniform guanine and cytosine (GC) composition (Bernardi 2000; Eyre-Walker and Hurst 2001). Isochores are made visible by plotting the GC content (%) determined within a long (40 kb) sliding window. The central 510-kb stretch, which has 42.0% GC and belongs to the mouse L2 isochore, is flanked by two H1 isochores (>45% GC). The centromeric 190 kb, which includes the last five class I genes of the *H2-T* region, averages 45.6% GC, and the telomeric, 250-kb stretch has 45.3% GC. Peaks of high GC-content, visible in the short 8 kb window shown in grey, correspond to CpG islands (Kundu and Rao 1999; Fig. 1).

The distribution of LINE (long interspersed nuclear element), SINE (short interspersed nuclear element), and LTR

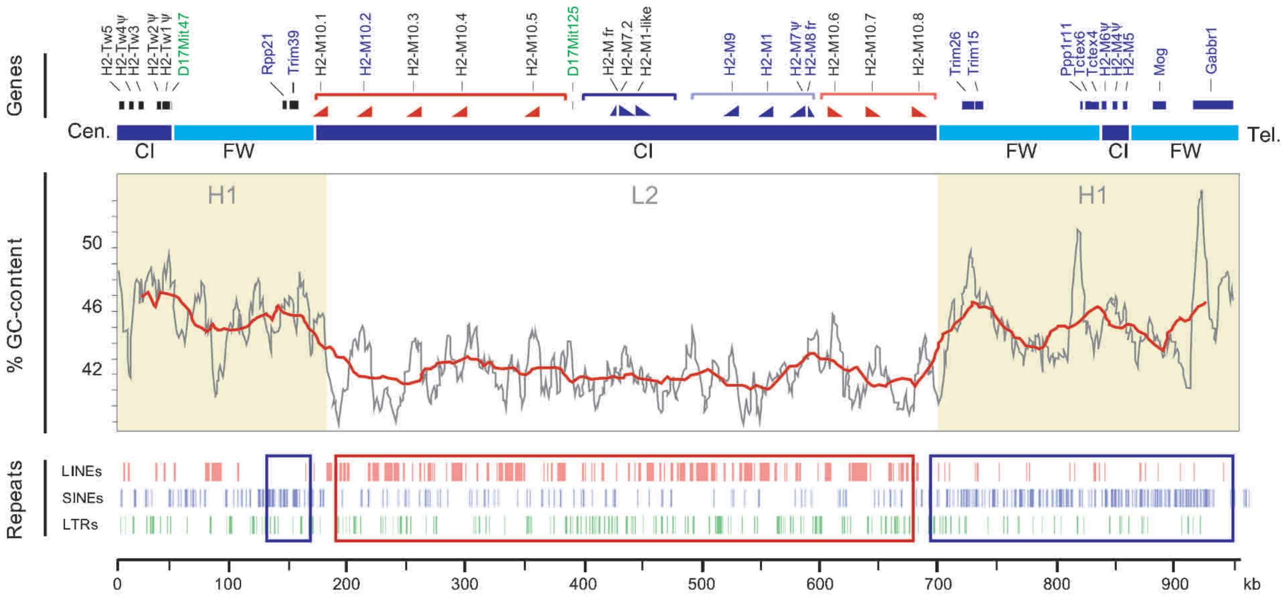

Figure 1 Gene, isochore, and repeat content in the centromeric *H2-M* region. Genes and markers are indicated on *top* by name, with already known genes in blue, newly identified class I genes in black, and microsatellites in green. The bar marks framework (FW) regions and class I (CI) expansion zones in light and dark blue. The GC-content in percentage was calculated using an 8-kb window/1-kb shift (grey line) and a 40-kb window/10-kb shift (red line) sliding window analysis. The class I gene-rich region from *H2-M10.1* to *H2-M10.8* forms an L2 isochore with an average of 42% GC, and it is surrounded by H1 isochores with an average of 45% GC. The distribution of repetitive elements is indicated in the *bottom* part of the figure. Red and blue boxes mark regions particularly rich in LINEs and SINEs, respectively. The class I gene orientation is indicated to help to understand the duplications and inversions that may have given rise to the present day *H2-M1* (light and dark blue) and *H2-M10* (light and dark orange) families.

(long terminal repeat) genome-wide repeat-elements, as defined by the RepeatMasker program, is shown in the bottom part of Figure 1. The sequence contains 18.9% LINE repeats (18.68% LINE1, 0.17% LINE2, and 0.01% L3/CR1), 8.1% SINE repeats (3.59% B1, 4.18% B2-B4, 0.15% ID, and 0.14% MIR), 9.2% LTR retroposons (3.53% MaLR, 0.14% ERVL, 0.71% ERV class I, and 2.31% ERV class II), and 0.5% DNA transposons. Supplemental Research Data shows the repeat distribution in greater detail, including the position of simple sequence repeats.

Two distinct regions, boxed in blue on Figure 1, have a high density of SINEs and low density of LINEs. The 50 kb in the centromeric H1 isochore contains 17.2% SINEs and 1.8% LINEs; the 250-kb telomeric H1 isochore contains 16.3% SINEs and 3.4% LINEs. The central L2 isochore of 510 kb, boxed in red, has only 4.0% SINEs but 29.7% LINEs. The LTR retroposons and DNA transposons are uniformly distributed.

Gene Organization

Table 1 summarizes the gene and pseudogene content and the Supplemental Research Data shows the features of the sequenced region in detail. Exons of known genes were identified from mRNA data, whereas exons of predicted genes were identified by the prediction programs and from EST data, if available. We defined the subregions based on the gene content: The first 52 kb represents the telomeric part of the *H2-T* class I region. The transition from the *H2-T* to the *H2-M* region spans from 53 kb to 163 kb of the entire sequence (Supplemental Research Data). It contains several repeated fragments and presumed pseudogenes with similarity to genes encoding various nuclear proteins: Arl1, Doc-1, LSm5, Histone H3, and Hmg14. An Hmg14 (high-mobility-group chromatin protein) fragment was previously mapped to this region by hybridization (Johnson et al. 1992). We found a homolog of the human *RPP21* subunit of the RNase P holoenzyme in this region, next to *Trim39* (or *Rnf23*; Jarrous et al. 2001).

The *H2-M* region itself consists of class I-rich regions separated by clusters of nonclass I genes, termed framework genes (Amadou 1999; Amadou et al. 1999). In the 790 kb of the *H2-M* region sequenced here, 18 previously characterized genes and pseudogenes were identified by similarity searches. We newly identified seven intact *Mhc* class I genes and two nonclass I RING finger genes (Table 1).

Table 1. Genes and Pseudogenes in the Mouse Telomeric Class I Region

Name					
Class I	**Non class I**	**Position**	**Direction**	**ORF**	**Features**
H2-Tw5[a]		4,316–8,242	–	+	H2-T, similar to X16219 and U21906
H2-Tw4ψ[a]		17,389–21,407	–	–	H2-T, similar to X16211
H2-Tw3[a]		25,163–28,708	–	+	H2-T, similar to X16209 and NM_008208
H2-Tw2ψ[a]		39,715–44,582	–	–	H2-T, similar to X16208
H2-Tw1ψ[a]		46,137–51,153	–	–	H2-T, similar to X15147
	Rpp21	145,924–148,093	–	+	Homologous to human ribonuclease P subunit, RPP21 (AF212152)
	Trim39 (Rnf23)	149,120–162,262	–	+	RING finger gene, Rnf23 (AB046382)
H2-M10.1[a]		174,553–176,694	–	+	H2-M10 family
H2-M10.2		214,020–216,155	–	+	H2-M10 (AF016309)
H2-M10.3[a]		256,236–258,381	–	+	H2-M10 family
H2-M10.4ψ[a]		296,108–299,108	–	–	H2-M10 family
H2-M10.5[a]		356,884–359,049	–	+	H2-M10 family
H2-M7.2[a]		429,850–432,445	+	+	H2-M1 family
H2-M1 likeψ[a]		442,912–446,564	–	–	H2-M1 family
H2-M9		525,306–527,525	–	+	H2-M9 (NM_008205)
H2-M1		554,882–557,072	–	+	H2-M1 (M20985)
H2-M7ψ		585,974–588,651	–	–	H2-M7 (L14276)
H2-M8		588,936–589,638	+	–	H2-M8 (L14277), gene fragment
H2-M10.6[a]		613,059–616,059	+	–	H2-M10 family
H2-M10.7[a]		643,734–645,899	+	+	H-M10 family
H2-M10.8[a]		683,235–685,369	+	+	H2-M10 family
	Trim26 (Zfp173)	721,721–730,407	+	+	RING finger gene, homologous to human ZNF173 (XM_004549)
	Trim15 (ZfpB7)	731,722–738,217	–	+	RING finger gene, homologous to human ZNFB7 (U34249)
	Trim10 (Herf1)	740,628–748,861	+	+	Hematopoiesis-specific RING finger gene (AF134811)
	ZfpU[a]	754,088–760,157	–	+	Predicted RING-Finger gene
	Trim31 (ZfpRingR)	769,287–780,722	+	+	RING finger gene, homologous to human TRIM31 (NM_007028)
	ZfpW[a] (Zfp412)	814,205–819,102	+	+	RING finger gene, homologous to human HZFw (AF238315-17)a
	Ppp1r11 (Tctex5)	820,505–822,902	–	+	Homologous to human PPP1R11 gene (NM_021959)
	Tctex6	825,575–829,647	–	+	Homologous to human ZNRD1 gene (AF024617)
	Tctex4	829,819–836,835	+	+	Homologous to human HTEX4 gene (AF032110)
H2-M6ψ		839,538–842,840	–	–	H2-M6 (L14280)
H2-M4ψ		849,199–853,199	–	–	H2-M4 (L14278)
H2-M5		858,131–860,729	–	+	H2-M5 (L14279)
	Zfp57	874,239–881,843	+	+	KRAB box containing C2H2 type zinc finger gene (NM_009559)
	Mog	881,951–894,435	–	+	Myelin oligodendrocyte glycoprotein (Mog; AH006537)
	Gabbr1	917,439–944,422	+	+	γ-aminobutyric acid (GABA) β receptor, 1 (AL078630)

[a]Newly identified.

RING Finger and Other Nonclass I Genes

RING finger domains are defined by the consensus sequence CX2CX(9-39)CX(1-3)HX(2-3)C/HX2CX(4-48)CX2C, where the Cys and His residues bind zinc (Saurin et al. 1996). The RING finger proteins interact with other proteins in a range of biological processes, including development, oncogenesis, apoptosis, and viral replication. Five previously identified RING finger genes (*Trim39*, *Trim26*, *Trim15*, *Trim10*, and *Trim31*; Orimo et al. 2000; Reymond et al. 2001) were localized to the *H2-M* region, and two new RING finger genes (*ZfpU* and *ZfpW*) were predicted by GenScan and EST database searches (Table 1). Six of these form a cluster (*Trim26*, *Trim15*, *Trim10*, *ZfpU*, *Trim31*, and *ZfpW*) in the 120-kb region between *H2-M10.8* and *Ppp1r11* (or *Tctex5*; Supplemental Research Data). These genes are under further investigation (C. Amadou, T. Takada, A. Kumánovics, and K. Fischer Lindahl, in prep.). The corresponding 307-kb region around *Trim26* (or *Zfp173*) from pig has been sequenced, and four RING finger genes were detected. The pig has an extra RING finger gene centromeric to *Trim26* (or *Zfp173*) compared with mouse and man (Renard et al. 2001).

Trim39, which encodes a zinc-binding protein abundant in testis (Orimo et al. 2000), and *Rpp21*, which encodes a subunit of the ribonuclease P holoenzyme, span the 18 kb in the boundary between the *H2-T* and *H2-M* class I genes (Table 1).

The 20 kb between *ZfpW* and *H2-M6* contains three "T complex testis-expressed" (*Tctex*) genes, *Tctex5* (or *Ppp1r11*), *-6*, and *-4* (Ha et al. 1991; Table 1). The 80-kb telomeric to *H2-M5* contains a Krüppel-like zinc finger gene, *Zfp57*, and two previously mapped and analyzed genes, encoding the myelin-oligodendrocyte glycoprotein gene (*Mog*; Gardinier et al. 1992; Pham-Dinh et al. 1995), and a receptor for γ-aminobutyric acid (*Gabbr1*; Grifa et al. 1998).

Class I Genes

The sequence presented here contains three clusters of *Mhc* class I genes. In the first 52 kb of the centromeric H1 isochore (Fig. 1), which belongs to the *H2-T* region, there are two genes, three pseudogenes and some fragments. Because allelic relationships are difficult to establish for the *H2-T* class I genes, which differ in number between haplotypes and have been incompletely characterized, we have tentatively named these genes *H2-Tw* with a number counting from the telomeric end (*w* for work in progress). $Tw5^{bc}$ is most similar to the *H2-Bl* (for blastocyst) gene previously sequenced from a strain 129/SvJ cosmid (Sipes et al. 1996) and mapped to the center of the *H2-T* region (Amadou et al. 1999). *H2-Bl* is expressed in mouse preimplantation blastocysts and in the placenta. $Tw3^{bc}$ is very similar to the *H2-T3* gene already sequenced from the *b*, *d*, and *k* haplotypes (Obata et al. 1985; Pontarotti et al. 1986; Mashimo et al. 1992), where it encodes the TL antigen expressed in the intestinal epithelium (Wu et al. 1991). Both *Tw5* and *Tw3* encode the consensus residues characteristic of peptide-binding class I molecules (Fig. 2), and they are more closely related to classical class Ia genes than to any of the *H2-M* class I genes (Fig. 3).

In the 30-kb telomeric H1 isochore (Fig. 1), we found one class I gene, *H2-M5*, and two pseudogenes, *H2-M4* and *H2-M6*. These class I genes were previously sequenced from two overlapping cosmids from strain BALB/c (haplotype *d*; Wang and Fisher Lindahl 1993). *M4* and *M6* are also pseudogenes in the *d* haplotype; $M5^{bc}$ differs from $M5^{d}$ by four amino acids (A30D, P49A, D122E, F270L; *d* to *bc*). Although the open reading frame of *M5* is conserved, transcripts have been hard to find in the mouse (Wang and Fischer Lindahl 1993). The significance, if any, of the M5 polymorphism is not known.

The central L2 isochore (Fig. 1) contains nine *H2-M* class I genes, four pseudogenes and several fragments in 520 kb. These class I genes can be divided into two families, named after previously sequenced members: The *M1* family of three genes, two pseudogenes and two fragments, flanked by five members of the *M10* family on the centromeric side and three on the telomeric side (Fig. 1 and Supplemental Research Data). To clarify the relationships among the ten intact *H2-M* genes described here plus *H2-M3* and *H2-M2*, which are located telomeric to the region discussed here, we have constructed a neighbor-joining tree based on the α3 domains encoded by exon 4 (Fig. 3). The genes from the L2 isochore formed two distinct families, a tight cluster of the M10 family and a more open cluster of the M1 family. M2, M3, and M5 were closer to the class Ia genes than the M1 and M10 families and did not form a tight family.

Of the intact members of the *M1* family, *H2-M1* (Singer et al. 1988) and *H2-M9* (Arepalli et al. 1998) were described previously, as was the pseudogene *H2-M7* and the fragment *H2-M8* (Wang and Fischer Lindahl 1993). The newly identified open reading frame, *H2-M7.2*, is most closely related to the pseudogene, *H2-M7*; hence the name. Figure 2 aligns the deduced amino acid sequence of the *bc* allelic forms of these three molecules with $M1^{b}$ from C57BL/6 and $M9^{d}$ from BALB/c. M9 is identical between the *bc* and *d* haplotypes, and M1 differs between *b* and *bc* by only one amino acid (Gly135 to Ala), predicted to be located in the loop before the α2 helix. We chose to name the members of the *M10* family by their linear order along the chromosome, just like genes of the *H2-Q* and *H2-T* regions were previously named (Klein et al. 1990). The first *M10* gene to be cloned came from strain BALB/c (Arepalli et al. 1998). It is now clear that this gene represents the *d* allele of *M10.2*; the alleles differ by only three amino acids.

Class Ia molecules share canonical amino acids that interact with the amino- and carboxy-termini of bound peptides and are located at the ends of the peptide binding groove. Of these, only Trp147 is conserved in the M1 family (Fig. 2). Substitution of Trp167 is seen in class I-like molecules, such as M3 and FcRn, in which the A pocket is closed (Burmeister et al. 1994; Wang et al. 1995a). Therefore, if the molecules of the M1 family bind peptides at all, it is not in the classical manner. Like the M1 family of molecules, the M10 proteins have a substitution at position 167 (Trp to Arg), and they lack most consensus residues involved in standard peptide binding. All four cysteine residues that form the essential disulfide bridges in the α2 and α3 domains of class I molecules are conserved in the M1/M10 families.

Expansion of Class I Genes

To understand the organization and evolution of the *H2-M* region, we compared the 951,695 bp sequence (after masking of genome-wide repeats) to itself as a dot matrix (Fig. 4A). Many regions of similarity were found, represented by dots and lines parallel (direct repeats) or perpendicular (inverted repeats) to the diagonal. Because the *Mhc* class I genes are similar to each other, they show up as short diagonal lines. For example, in Figure 4A, the similarity between the *H2-T* and *H2-M* class I sequences is visible; the *H2-T* and *H2-M* class I genes are separated by a 90-kb region of extensive duplication of "chromatin-related" pseudogenes (Figs. 4A,B).

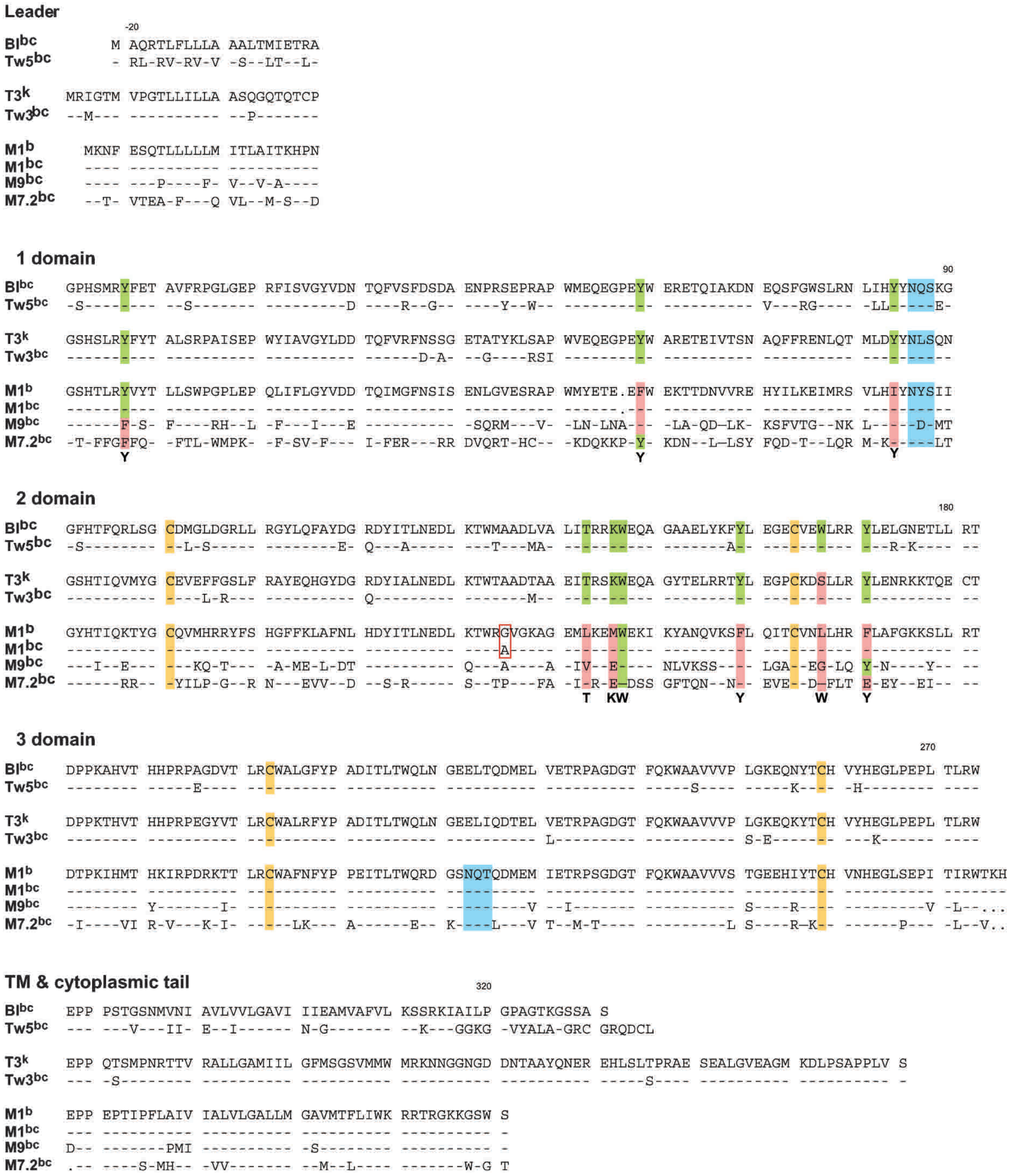
Leader
-20
Bl^bc M AQRTLFLLLA AALTMIETRA
Tw5^bc - RL-RV-RV-V -S--LT--L-
T3^k MRIGTM VPGTLLILLA ASQGQTQTCP
Tw3^bc --M--- ---------- --P-------
M1^b MKNF ESQTLLLLLM ITLAITKHPN
M1^bc ---- ---------- ----------
M9^bc ---- ---P----F- V--V-A----
M7.2^bc --T- VTEA-F---Q VL--M-S--D
1 domain
90
Bl^bc GPHSMRYFET AVFRPGLGEP RFISVGYVDN TQFVSFDSDA ENPRSEPRAP WMEQEGPEYW ERETQIAKDN EQSFGWSLRN LIHYYNQSKG
Tw5^bc -S-------- --S------- ---------D ----R--G-- ----Y--W-- ---------- ---------- V---RG---- -LL-----E-
T3^k GSHSLRYFYT ALSRPAISEP WYIAVGYLDD TQFVRFNSSG ETATYKLSAP WVEQEGPEYW ARETEIVTSN AQFFRENLQT MLDYYNLSQN
Tw3^bc ---------- ---------- ---------- ------D-A- --G----RSI ---------- ---------- ---------- ----------
M1^b GSHTLRYVYT LLSWPGPLEP QLIFLGYVDD TQIMGFNSIS ENLGVESRAP WMYETE.EFW EKTTDNVVRE HYILKEIMRS VLHIYNYSII
M1^bc ---------- ---------- ---------- ---------- ---------- ------.--- ---------- ---------- ----------
M9^bc ------F-S- F----RH--L -F---I---E ---------- -SQRM---V- -LN-LNA--- -LA-QD—LK- KSFVTG--NK L-----D-MT
M7.2^bc -T-FFGFFQ- -FTL-WMPK- -F-SV-F--- I-FER---RR DVQRT-HC-- -KDQKKP-Y- KDN--L—LSY FQD-T--LQR M-K-----LT
Y Y Y
2 domain
180
Bl^bc GFHTFQRLSG CDMGLDGRLL RGYLQFAYDG RDYITLNEDL KTWMAADLVA LITRRKWEQA GAAELYKFYL EGECVEWLRR YLELGNETLL RT
Tw5^bc -S-------- --L-S----- --------E- Q---A----- ---T---MA- ---------- -------A-- ---------- ---R-K---- --
T3^k GSHTIQVMYG CEVEFFGSLF RAYEQHGYDG RDYIALNEDL KTWTAADTAA EITRSKWEQA GYTELRRTYL EGPCKDSLLR YLENRKKTQE CT
Tw3^bc ---------- ----L-R--- ---------- Q--------- -------M-- ---------- ---------- ---------- ---------- --
M1^b GYHTIQKTYG CQVMHRRYFS HGFFKLAFNL HDYITLNEDL KTWRGVGKAG EMLKEMWEKI KYANQVKSFL QITCVNLLHR FLAFGKKSLL RT
M1^bc ---------- ---------- ---------- ---------- ----A----- ---------- ---------- ---------- ---------- --
M9^bc ---I--E--- ---KQ-T--- -A-ME-L-DT ---------- Q---A----A -IV--E---- NLVKSS---- LGA--EG-LQ Y-N----Y-- --
M7.2^bc ------RR-- -YILP-G--R N---EVV--D --S-R----- S--TP---FA -I-R-E—DSS GFTQN--N-- EVE--D—FLT E-EY--EI-- --
T KW Y W Y
3 domain
270
Bl^bc DPPKAHVT HHPRPAGDVT LRCWALGFYP ADITLTWQLN GEELTQDMEL VETRPAGDGT FQKWAAVVVP LGKEQNYTCH VYHEGLPEPL TLRW
Tw5^bc -------- -----E---- ---------- ---------- ---------- ---------- -----S---- -----K---- -H-------- ----
T3^k DPPKTHVT HHPRPEGYVT LRCWALRFYP ADITLTWQLN GEELIQDTEL VETRPAGDGT FQKWAAVVVP LGKEQKYTCH VYHEGLPEPL TLRW
Tw3^bc -------- ---------- ---------- ---------- ---------- L--------- ---------- S-E------- ---K------ ----
M1^b DTPKIHMT HKIRPDRKTT LRCWAFNFYP PEITLTWQRD GSNQTQDMEM IETRPSGDGT FQKWAAVVVS TGEEHIYTCH VNHEGLSEPI TIRWTKH
M1^bc -------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- -------
M9^bc -------- Y-------I- ---------- ---------- ---------V --I------- ---------- S----R---- ---------V -L--...
M7.2^bc -I----VI R-V---K-I- -----LK--- A------E-- K----L---V T--M-T---- ---------L S----R—K-- ------P--- -L--V..
TM & cytoplasmic tail
320
Bl^bc EPP PSTGSNMVNI AVLVVLGAVI IIEAMVAFVL KSSRKIAILP GPAGTKGSSA S
Tw5^bc --- ---V---II- E--I------ N-G------- --K---GGKG -VYALA-GRC GRQDCL
T3^k EPP QTSMPNRTTV RALLGAMIIL GFMSGSVMMW MRKNNGGNGD DNTAAYQNER EHLSLTPRAE SEALGVEAGM KDLPSAPPLV S
Tw3^bc --- -S-------- ---------- ---------- ---------- ---------- -----S---- ---------- ---------- -
M1^b EPP EPTIPFLAIV IALVLGALLM GAVMTFLIWK RRTRGKKGSW S
M1^bc --- ---------- ---------- ---------- ---------- -
M9^bc D-- -------PMI ---------- -S-------- ---------- -
M7.2^bc .-- ----S-MH-- -VV------- --M--L---- -------W-G T

Figure 2 Alignment of the deduced amino acid sequences of two H2-T proteins and the H2-M1 family with their closest relatives. Tw5 is compared with H2-Bl^{bc} of strain 129 (accession no. NM_008199) and Tw3 with T3^{k} of strain C3H (accession no. NM_008208). M1, M9, and M7.2 of haplotype *bc* are compared with M1 of the *b* haplotype (accession no. M20985); the purple box highlights the only allelic variation. M9 of the *d* haplotype (accession no. NM_008205) is identical to M9^{bc} and not shown. The position of the highly conserved cysteine residues are marked in yellow; putative sites for N-linked glycosylation are marked in light blue; residues normally found at the ends of the peptide-binding site in Mhc class Ia molecules are listed below the sequences, with green highlights where conserved and pink where not.

The 520-kb region from *H2-M10.1* to *H2-M10.8* shows evidence of extensive duplications and inversions. This region was compared without repeat-masking to itself in Figure 4C and to its complement in Figure 4D. The *H2-M1* family genes, between 428 and 590 kb, are flanked by two regions of *H2-M10* genes, at 174–360 kb and 613–686 kb. Segmental duplication is evident as longer diagonal lines, including not only class I coding regions but also noncoding and repeat regions (Fig. 4C). The class I genes from *M10.1* through *M1-like* are all in the opposite orientation (telomere-to-

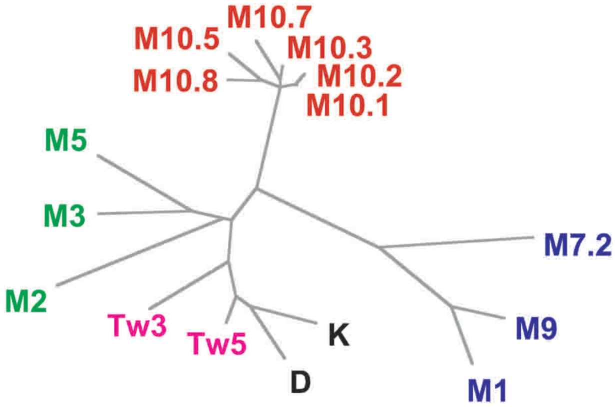

Figure 3 Neighbor-joining tree, obtained by using MEGA2 (Kumar et al. 2001), is based on the amino acid sequence of the $\alpha3$ domain of all H2-M class Ib molecules and Tw3 and Tw5. The classical class Ia proteins, H2-K^{bc} (accession no. AF100956) and H2-D^{b} (accession no. AAA39601) were included for comparison. The sequence of H2-M3 is from the *b* haplotype (accession no. NM_013819) and that of H2-M2 is from the *d* haplotype (accession no. M26156); both genes are telomeric to *Gabbr1* (Amadou et al. 1999).

centromere) relative to the class I genes in the *M9 – M10.8* segment (Fig. 1). The L1 elements of the LINE family are frequent in this region, including one potentially transposable unit with full-length open reading frames between 333,891 and 338,892 bp. L1 elements that are present in the *M1* family region are also there in the *M10* region, suggesting that they could have contributed to the expansion of the region.

Comparison of the Telomeric *Mhc* Class I Region of Mouse and Man

We identified an 853-kb segment from the sequenced human *Mhc* (accession no. NT_001520; MHC Sequencing Consortium 1999), which represents the region of conserved synteny to our mouse *M*-region sequence. The dot matrix comparison of these human and mouse sequences (Fig. 5) recognized three well conserved regions: 18 kb including *Rpp21* and *Trim39* (or *Rnf23*), 130 kb including *Trim26* (or *Zfp173*) to *Tctex4*, and 80 kb including *Zfp57* to *Gabbr1* genes. These regions appear as long diagonal lines, boxed in red, green and blue in Figure 5.

The human and mouse class I genes are sufficiently similar to appear as short lines in this dot-matrix analysis, but they are not orthologous and have evolved independently in the two species (Hughes and Nei 1989; Hughes 1991; Klein and O'hUigin 1994; Kumánovics et al. 2003). Between *Trim39* (or *Rnf23*) and *Trim26* (or *Zfp173*), the mouse has nine class I genes and four pseudogenes and no other kinds of genes in 520 kb. In the corresponding 85-kb interval in man, there is only a single class I pseudogene, *HLA-92* (Geraghty et al. 1992; MHC Sequencing Consortium 1999). On the other hand, the mouse has only one intact class I gene, *H2-M5*, and two pseudogenes in 30 kb between *Tctex4* and *Zfp57*, whereas the corresponding interval in man is 390 kb long and contains three expressed class I genes (*HLA-A*, *-G*, and *-F*), eight class I pseudogenes or fragments, three class-I-like *MIC* pseudogenes, 13 *HCG* genes/pseudogenes and numerous gene fragments (Geraghty et al. 1992; MHC Sequencing Consortium 1999).

DISCUSSION

The entire mouse *Mhc* (*H2*) has been cloned from strain 129/Sv (Amadou et al. 1999), the strain of choice for gene targeting and deletion. Almost two megabases of sequence, from the *H2-K* region through the class II/III regions to the *H2-D/Q* region is already determined (Kumánovics et al. 2003). Here we provide the complete genomic sequence of 951 kb of the telomeric part of the mouse *Mhc*, from *H2-Tw5* through *Gabbr1*. Another 879 kb of sequence from *Gabbr1* through the olfactory receptor region is being analyzed (C. Amadou, R.M. Younger, S. Sims, L.H. Matthews, J. Rogers, A. Ziegler, S. Beck, and K. Fischer Lindahl, in prep.; Younger et al. 2000). The sequencing of the remaining ~800 kb, including the *H2-T* region, is hindered by the highly repetitive nature of the region, but it should, nonetheless, soon be completed (A. Kumánovics, E. P. Jones and K. Fischer Lindahl, unpubl.).

Gene Content: The Class I Genes

Strain 129, haplotype *bc*, is similar in the centromeric 1.5-Mb, classical part of *H2* from *H2-K* through *H2-D* to haplotype *b* (e.g., C57BL/6). For example, the H2-K and H2-D proteins are identical between haplotypes *b* and *bc* (Kumánovics et al. 2002). But it was also long known that haplotype *bc* differs in the telomeric *H2,* for example, the *H2-Q*, *-T* and *-M* regions, from haplotype *b*, hence the designation *bc* (Snell et al. 1971). The number of class I genes in the *Q* and the *T* regions differs between these two haplotypes (Fischer Lindahl 1997; Kumánovics et al. 2002). The *H2-M* region, on the other hand, appears to be stable among the investigated strains (Wang and Fischer Lindahl 1993; Jones et al. 1995; Jones et al. 1999). Moreover, the class I genes are orthologous between the rat and mouse *M* regions (Wang et al. 1995b; Günther and Walter, 2001).

By sequencing, we identified 12 class I genes from *H2-Tw5* to *Gabbr1* that have complete open reading frames and are therefore predicted to be functional. In the *H2-T* region segment, we found two genes and three pseudogenes. *H2-Tw3* appears to be an allele of *H2-T3*, which encodes the serologically defined TL (for thymus-leukemia) antigen (Old et al. 1963; Wu et al. 1991). TL has been studied quite extensively, yet the function remains elusive. The strong interaction between the TL molecule expressed by intestinal epithelial cells and the CD8$\alpha\alpha$ homodimer expressed by intraepithelial lymphocytes (IEL) could regulate the behavior of IELs (Leishman et al. 2001).

H2-Tw5 is most similar to *H2-Bl* (Fig. 2). *H2-Bl* is expressed in preimplantation embryos and the placenta (Sipes et al. 1996). *H2-Bl* maps to the center of the *H2-T* region in haplotype *bc* (Amadou et al. 1999). *Tw5* and *Bl* are presumably related to each other by one of the strain-specific large-scale duplications that are characteristic of the *H2-T* region (Teitell et al. 1994). We have to wait for the complete sequence of the *H2-T* region to define this duplication in strain 129.

The class I genes of the *H2-M* region fall into three groups, based on the comparison of their $\alpha3$ domains (Fig. 3). The M1 family has three potentially functional members (*M1*, *M7.2* and *M9*), and the M10 family has six (*M10.1*, *M10.2*, *M10.3*, *M10.5*, *M10.7* and *M10.8*; Figs. 2, 3). The rest of the *M* region class I genes (*M2*, *M3*, and *M5*) do not form a close family (Fig. 3). In the *M6–M4–M5* class I cluster, only *M5* has the potential to encode a class I molecule; the others have frame-shifts and in-frame stop codons rendering them pseudogenes (Wang and Fischer Lindahl 1993). Unlike in mouse,

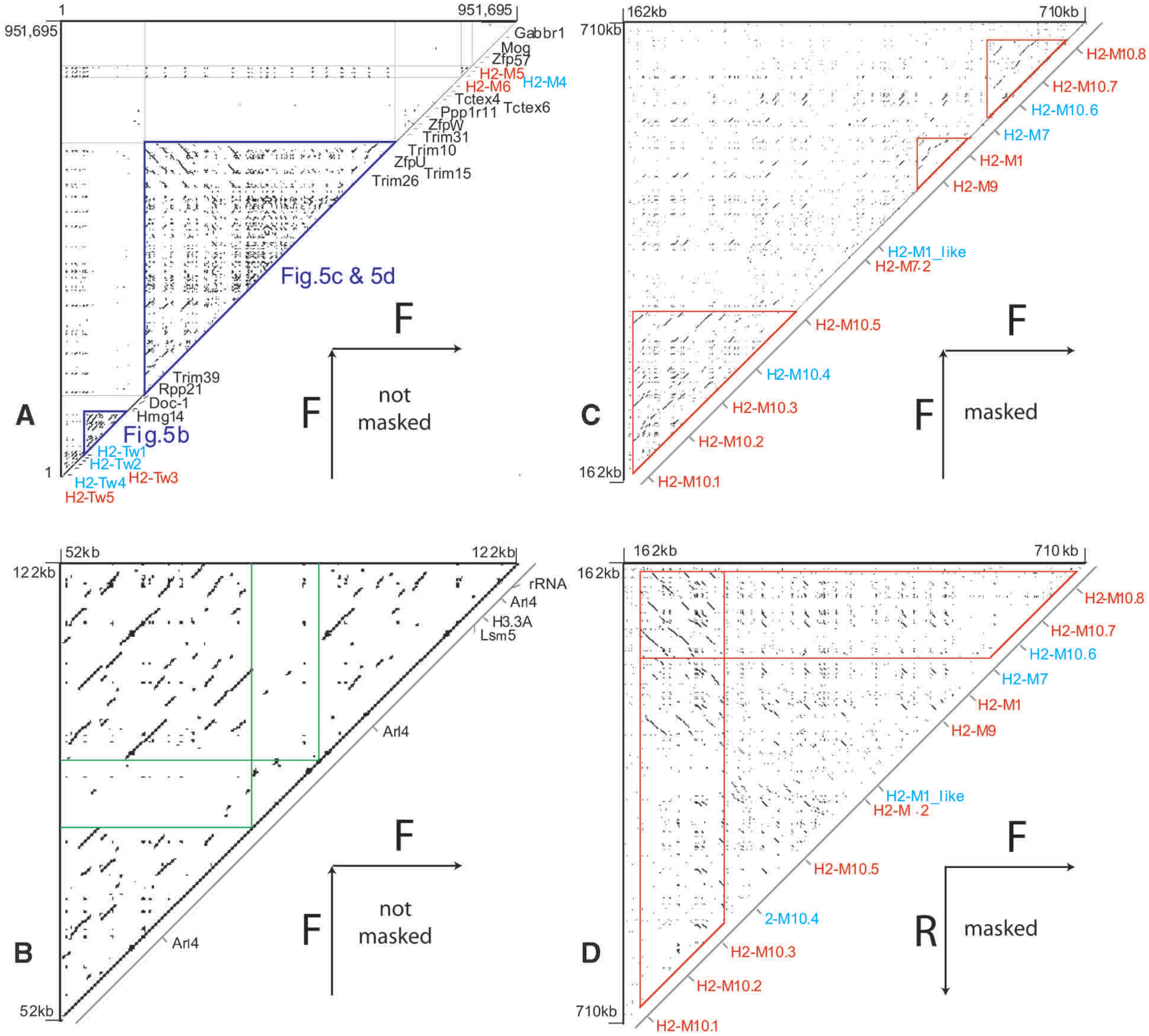

Figure 4 Dot matrix analysis of the 951,695 bp sequence of the *H2-M* region against itself. (*A*) Segmental duplication and inversion of the class I genes is evident in the central box. This plot was generated with PipMaker, set to compare both strands and show all matches. (*B*) Self-comparison of the 70-kb region rich in pseudogenes for nuclear proteins. (*C*) Forward and (*D*) reverse self-comparison of the 520-kb *M1-M10* region. The genome-wide repeats are masked in plots (*C*) and (*D*), whereas plots (*A*) and (*B*) are not masked.

the rat *M4*, *M5*, and *M6* all have open reading frames, and transcripts can be detected by RT-PCR (D. Lambracht-Washington and K. Fischer Lindahl, in prep.). In mouse, *M5* transcripts have been detected in the thymus by the same method.

The *M1* and *M10* families encode novel and divergent *Mhc* class I proteins. Restricted expression pattern and low polymorphism define the nonclassical or class Ib molecules (Stroynowski and Fischer Lindahl 1994). The *M1/M10* families conform to both. Their expression appears to be strictly regulated. *M1* expression appears to be actively repressed in adult tissues (Howcroft et al. 1996). We did not find convincing *M10* expression in mouse embryos or major adult mouse organs (E.P. Jones, unpubl; Arepalli et al. 1998). There is only limited information on the *M1/M10* polymorphism (Singer et al. 1988; Arepalli et al. 1998), but based on the three sequences from haplotype *d,* the *M1/M10* genes appear to be oligo- or monomorphic (Fig. 2). Most of the residues forming the M1/M10 antigen binding groove do not follow the class I consensus (Fig. 2), suggesting a function different from the classical antigen presentation. *M1/M10* have no equivalent in human or any other species, so far, except in rat.

Gene Content: The Nonclass I Genes

One of the main interests in the *Mhc* is its association to hundreds of diseases. Sequencing and detailed mapping of the *Mhc* revealed that the class I regions from human, mouse, rat, and pig encode a large number of nonclass I genes (this study; Amadou et al. 1999; MHC Sequencing Consortium 1999;

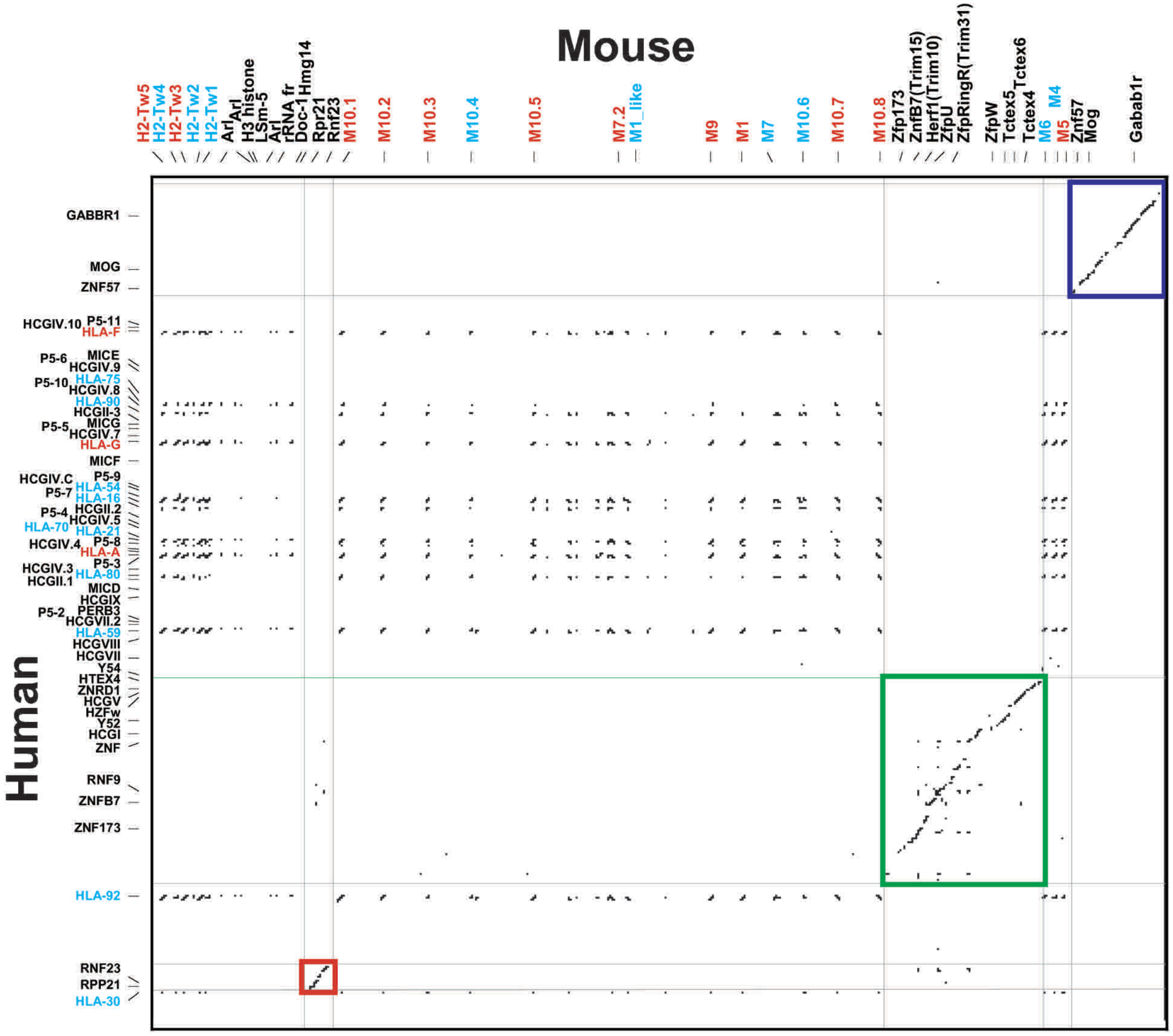

Figure 5 Dot matrix analysis of the telomeric class I region of mouse and man. The 951,695-bp sequence of the centromeric *H2-M* region is compared with 853,185 bp of the *HLA* region (accession no. NF_001520). Class I genes are red, class I pseudogenes are blue; other genes and pseudogenes are black. The red, green, and blue boxes mark the three framework regions that are conserved syntenic regions of orthologous nonclass I genes, which yield clear diagonal lines on the dot matrix. The class I genes give only short lines of similarity between paralogous genes.

Chardon et al. 2001; Ioannidu et al. 2001; Renard et al. 2001). Mapping of corneodesmosin (*CDSN* or *S* gene), a nonclass I gene expressed in skin, to the class I region provided the first hint that not all class I-linked diseases are necessarily mediated by the class I genes themselves (Zhou and Chaplin 1993). Psoriasis vulgaris, a skin disorder, is associated with alleles of *HLA-C*, but the gene (or genes) responsible for the disease most likely maps telomeric to *HLA-C*, where, among others, corneodesmosin maps (Oka et al. 1999; Nair et al. 2000). The presence of nonclass I genes in the class I region not only helps to explain the evolution of the *Mhc,* but provides new clues to the molecular bases of a large number of diseases.

There are 14 nonclass I genes in our sequence (Table 1). Three of them have known function: *Rpp21* is a subunit of the nuclear ribonuclease P holoenzyme, the tRNA processing enzyme (Jarrous et al. 2001). *Mog* encodes the myelin-oligodendrocyte glycoprotein. Mog protein is specific to the central nervous system and it is the major autoantigen in experimental autoimmune encephalomyelitis (EAE), the widely used animal model for autoimmune demyelinating diseases such as multiple sclerosis (Gardinier et al. 1992; Pham-Dinh et al. 1995). *Gabbr1* is a receptor for the neurotransmitter γ-amino-butyric acid (GABA). Intriguingly, susceptibility loci for multiple sclerosis, epilepsy, and schizophrenia have been suggested to map in the region of *GABBR1* (Grifa et al. 1998). *Herf1* (hematopoietic RING finger 1) is required for terminal differentiation of erythroid cells (Harada et al. 1999). ZfpRingR or TRIM31 might participate in the retinoid-induced growth arrest of MCF-7 breast carcinoma cells (Dokmanovic et al. 2002). *Tctex5* is a regulatory (inhibitor) subunit of protein phosphatase 1 (Ppp1r11; LocusID: 76497).

The other nonclass I genes in our region have no identified function (Table 1). But the *Trim26* (or *Zfp173*) to *Tctex4* cluster (Table 1, Fig. 5) is an interesting one even without known functions: (1) Overlapping and oppositely oriented genes generate possible sense–antisense gene pairs, such as *ZfpW* (or *HZFw*) and *Ppp1r11* (*Tctex5* or *HCGV*) at their 3′ end UTR (untranslated region), and *Tctex6* (or *HTEX6*) and *Tctex4* (or *HTEX4*) at their 5′ end (Coriton et al. 2000). (2) The region has a high gene density generated by the overlapping genes; and, finally, (3) alternative splicing has been detected for most of them (Lepourcelet et al. 1996; Lepourcelet et al. 1998; Coriton et al. 2000). Sense–antisense transcript pairs can form double-stranded RNA duplexes, which are increasingly thought to be an important part of the gene regulatory process. Natural antisense transcripts play a role in RNA interference, genomic imprinting, translational regulation, alternative splicing, X chromosome inactivation, and RNA editing, and finally, sense–antisense gene pairs can exhibit reciprocal expression patterns (Lehner et al. 2002; Shendure and Church 2002). Natural antisense transcripts have been predicted in the *Mhc*, in the class III region, which is also conserved between human and mouse and has a high gene density (Beck and Trowsdale 2000; Lehner et al. 2002).

Evolution of the *Mhc*

The *Mhc* class I and II proteins present antigens. An important question about the *Mhc* is how the *Mhc* can maintain the ability to present antigens from a large variety of infectious agents in diverse environments. Analysis of the class I genes showed that the basic structure and function of the *Mhc* and the class I and class II proteins are conserved throughout evolution (Trowsdale 1995; Ohta et al. 2000). Despite this fundamental conservation, the class I and class II gene sets are replaced with every major radiation as a result of a dynamic birth-and-death process (Klein et al. 1992; Nei et al. 1997; Klein et al. 1998). Orthologous genes are separated by speciation, as opposed to paralogs that are separated by gene duplication. Among mammals, orthologous class I genes are only found within the same order, such as in primates or in rodents (Hughes and Nei 1989; Hughes 1991). The human and mouse class I genes group separately in a phylogenetic tree, in a species-specific manner (Hughes and Nei 1989; Hughes 1991; Yeager et al. 1997; Kumánovics et al. 2003).

The mapping of nonclass I genes in the class I region complements the previous studies on class I evolution by providing evidence that the class I regions are orthologous, even if the class I genes are not (Amadou 1999; Amadou et al. 1999). We compared about half of the *Mhc* class I region of mouse (951 kb) and human (853 kb), and showed that, unlike the class I genes, all the nonclass I genes are conserved between the two species (Fig. 5). Without the nonclass I genes, the two class I regions cannot be aligned. The use of these nonclass I genes helps to understand the genomic organization and evolution of the *Mhc* class I region. The species-specific expansions of class I genes occurred in the same framework of nonclass I genes in both man and mouse (Amadou 1999; Amadou et al. 1999), and, most likely, in all other mammals too (Chardon et al. 2001; Ioannidu et al. 2001; Renard et al. 2001; Di Palma et al. 2002). In other words, the class I genes expanded in the same position in the genomes, but from different ancestors. For example, there are class I genes between *Trim39* (or *Rnf23*) and *Trim26* (or *Zfp173*) in both man and mouse, but the expanding gene in mouse was the ancestor of the *M1/M10* families occupying 520 kb, whereas in human the sole *HLA-92* pseudogene in 120 kb is unrelated to the mouse *M1/M10* genes. The 390-kb long *HLA-A* region expansion occurred between *HTEX4* and *ZFP57*. In mouse, the 37-kb *Tctex4–Zfp57* region contains the *H2-M4, 5, 6* genes, which are unrelated to *HLA-A, G, F* and to the pseudogenes from the corresponding human region.

How did the class I expansion happen? The self dot-plot comparison of the *M* region clearly shows that the *M1/M10* segment is the result of a series of segmental duplications followed by a large-scale inversion (Fig. 1 and 4D). The region is rich (29.7%) in L1-like transposons, including not only the usual fragmented L1 sequences (Smit 1999), but also one full length L1. In human, HERV-16 retroviral sequences are thought to mediate duplications by recombination between homologous retroelements (Kulski et al. 1997; Dawkins et al. 1999; Shiina et al. 1999b). Similarly, in the *M1/M10* segment the L1 elements are abundant and may have contributed to the expansion.

There is one major difference between man and mouse in this respect: In human, the same duplication unit can explain most of the class I region, whereas in mouse, the segments undergoing expansion are entirely locus specific. That is, we do not find common elements, outside of the class I genes, shared between the *K/D/Q*, *T*, *M1/M10* and *M4/5/6* expansion (this paper; Kumánovics et al. 2002), whereas in human, an ancient duplication unit for the entire class I region can be deduced as a segment containing a MIC gene, a HERV element, and the ancestor of the human class I genes (Dawkins et al. 1999).

The human class I duplication unit includes nonclass I genes, such as the *HCG* (hemochromatosis candidate gene) series (Pichon et al. 1996). In mouse, the *H2-Q* region duplication included *archain* pseudogenes (Kumánovics et al. 2002). The *M1/M10* expansion contains no nonclass I genes or gene fragments. There is a 70-kb, highly repetitive segment (Fig. 4B) between *H2-Tw1* and *Rpp21* (Fig. 4A), which contains only gene fragments and pseudogenes derived from class I and nonclass I genes. Despite the plentiful L1s and duplicated segments, the *M*-region appears stable among the mouse strains and is largely shared with rats (Jones et al. 1995; Wang et al. 1995b; Jones et al. 1999; Günther and Walter, 2001).

METHODS

Sequencing of BAC Clones

BAC clones were all from the CITB-CJ7-B library of strain 129/SvJ, distributed by Research Genetics. The minimal overlap fragment between clones 585c7 and 10i1 was isolated from clone 255d16 by treatment with *Xho*I and used to prepare sequencing templates.

Mechanically sheared BAC DNA fragments, 1–4-kb long, were ligated into M13 or pUC18 vectors (Bankier et al. 1987) and sequenced with M13 reverse primer until greater than 6× random shotgun coverage was achieved, based on the estimated BAC clone size. The DNA sequence was determined by the enzymatic dideoxy chain termination chemistry with automated ABI 377 or ABI 3700 sequencers (Applied Biosystems). Sequence base calling, contig assembly, quality clipping and screening for vectors were performed with PHRED/PHRAP and CONSED software (Ewing and Green, 1998; Ewing et al. 1998; Gordon et al. 1998).

We used two strategies in finishing to fill gaps and clarify ambiguities. Selected clones were sequenced with M13 forward or custom primers. We also used custom primers to generate PCR products, which were then sequenced directly.

To guard against sequence misassembly, we digested each BAC clone with restriction enzyme (*Bam*HI*, Bgl*II*, Eco*RI*, Hin*dIII *or Xba*I), separated the fragments by pulsed-field gel electrophoresis (CHEF-MAPPER, BioRad, Hercules, CA), and compared the pattern to a virtual digest of the consensus sequence. In all cases, the predictions were consistent with the digest pattern (not shown).

The sequences reported here have been submitted to GenBank under the following clone name and accession numbers: Citb585c7–AF532116; Citb255d16–AF532113; Citb10i1–AF532111; Citb592j14–AC005413; Citb76k14–AC005665; Citb592j14–AF532114; Citb9k22–AF532117; 553n23–AF532115; 20h22–AF532112; and 544e14–AF532114. The final sequence assembles as follows: Citb585c7 is from 1 to 192,871, 255d16 fragment is from 188,743 to 215,704, 10i1 is from 212,656 to 390,909, 592j14 is from 390,371 to 513,741, 76k14 is from 402,819 to 527,004, 9k22 is from 499,424 to 672,711, 553n23 is from 630,719 to 792,447, 20h22 is from 703,411 to 875,609, and 544e14 is from 797,824 to 951,695.

The quality score for each clone was calculated by CONSED before manual finishing: 585c7: 0.23×10^{-4}, 10i1: 0.06×10^{-4}, 20h22: 0.04×10^{-4}, 544e14: 0.12×10^{-4}, 255d16: 13.45×10^{-4}, 592j14: 2.6×10^{-4}, 76k14: 4.3×10^{-4}, 9k22: 9.17×10^{-4} and 533n23: 3.7×10^{-4}. Each BAC was finished separately, but in overlap regions, use was made of the information from the clearest reads. Small discrepancies were found in the overlap between 20h22 and 544e14. Three bases differ between them: T at position 107,794 and AC at position 113,728–113,729 of 20h22 are missing in 544e14. All these differences are located within repeat sequences. We used 20h22 for analysis, because it had a better quality score than 544e14.

544e14 is the most telomeric BAC in our contig and it overlaps with 573k1 (Younger et al. 2000), which was sequenced and assembled independently in the Sanger Institute (accession no. AL078630). There are five discrepancies in the 35,882 bp overlap between 544e14 and 573k1 (listed as 544e14 to 573k1, and the positions are given in 544e14): A to C at position 119,514; G insertion at position 119,964; G to N at position 121,847; CTC to GCT at position 122,414–122,416; GA insertion at position 151,811–151,812.

Sequence Analysis Programs

Interspersed and simple repeat sequences were identified and masked by RepeatMasker (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker). The masked sequence was compared against various databases using the BLAST programs (http://www.ncbi.nlm.nih.gov/BLAST; Altschul et al. 1997). GC-content was calculated with the ISOCHORE program (http://www.emboss.org). Coding regions and the gene structure were predicted by GENSCAN (http://genes.mit.edu/GENSCAN.html; Burge and Karlin 1997). EST alignments were also used to determine the exon–intron boundaries of predicted genes. CLUSTALW (Thompson et al. 1994) was used for multiple alignments, and the neighbor-joining trees were constructed with MEGA2 (Kumar et al. 2001). Dot matrix comparisons were carried out with PipMaker. To compare the human and mouse sequence we searched "both strands" and "showed all matches" in the output. To compare the sequence to itself we used "both strands" and "single strand" searches (http://bio.cse.psu.edu/pipmaker; Schwartz et al. 2000).

ACKNOWLEDGMENTS

This work was supported in part by NIH grants AI37818 and NHGRI HG00202. We thank Dr. Clive Slaughter and the HHMI Biopolymer Facility, in particular Teresa Bannister and Brad Carmical, as well as members of the Genome Science and Technology Center at University of Texas Southwestern Medical Center, in particular Pablo Aguayo, Cynthia English, Meiyu Loo and Jeff Schageman for their technical support.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

NOTE ADDED IN PROOF

The March 7, 2003, issue of *Cell* describes expression of M10 and M1 families in the vomeronasal organ (Loconto, J., Papes, F., Chang, E., Stowers, L., Jones, E.P., Takada, T., Kumánovics, A., Fischer Lindahl, K., and Dulac., C. Functional expression of murine V2R pheromone receptors involves selective association with the M10 and M1 families of class Ib molecules).

REFERENCES

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Amadou, C. 1999. Evolution of the Mhc class I region: The framework hypothesis. *Immunogenetics* **49:** 362–367.

Amadou, C., Kumánovics, A., Jones, E.P., Lambracht-Washington, D., Yoshino, M., and Fischer Lindahl, K. 1999. The mouse major histocompatibility complex: Some assembly required. *Immunol. Rev.* **167:** 211–221.

Arepalli, S.R., Jones, E.P., Howcroft, T.K., Carlo, I., Wang, C., Fischer Lindahl, K., Singer, D.S., and Rudikoff, S. 1998. Characterization of two class I genes from the *H2-M* region: Evidence for a new subfamily. *Immunogenetics* **47:** 264–271.

Bahram, S., Bresnahan, M., Geraghty, D.E., and Spies, T. 1994. A second lineage of mammalian major histocompatibility complex class I genes. *Proc. Natl. Acad. Sci.* **91:** 6259–6263.

Bankier, A.T., Weston, K.M., and Barrell, B.G. 1987. Random cloning and sequencing by the M13/dideoxynucleotide chain termination method. *Methods Enzymol.* **155:** 51–93.

Beck, S. and Trowsdale, J. 2000. The human major histocompatability complex: Lessons from the DNA sequence. *Annu. Rev. Genomics Hum. Genet.* **1:** 117–137.

Beck, T.W., Menninger, J., Voigt, G., Newmann, K., Nishigaki, Y., Nash, W.G., Stephens, R.M., Wang, Y., de Jong, P.J., O'Brien, S.J., et al. 2001. Comparative feline genomics: A BAC/PAC contig map of the major histocompatibility complex class II region. *Genomics* **71:** 282–295.

Bernardi, G. 2000. Isochores and the evolutionary genomics of vertebrates. *Gene* **241:** 3–17.

Braud, V.M., Allan, D.S., O'Callaghan, C.A., Soderstrom, K., D'Andrea, A., Ogg, G.S., Lazetic, S., Young, N.T., Bell, J.I., Phillips, J.H., et al. 1998. HLA-E binds to natural killer cell receptors CD94/NKG2A, B and C. *Nature* **391:** 795–799.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Burmeister, W.P., Gastinel, L.N., Simister, N.E., Blum, M.L., and Bjorkman, P.J. 1994. Crystal structure at 2.2Å resolution of the MHC-related neonatal Fc receptor. *Nature* **372:** 336–343.

Chardon, P., Rogel-Gaillard, C., Cattolico, L., Duprat, S., Vaiman, M., and Renard, C. 2001. Sequence of the swine major histocompatibility complex region containing all nonclassical class I genes. *Tissue Antigens* **57:** 55–65.

Clark, M.S., Shaw, L., Kelly, A., Snell, P., and Elgar, G. 2001. Characterization of the MHC class I region of the Japanese pufferfish (*Fugu rubripes*). *Immunogenetics* **52:** 174–185.

Coriton, O., Lepourcelet, M., Hampe, A., Galibert, F., and Mosser, J. 2000. Transcriptional analysis of the 69-kb sequence centromeric to HLA-J: A dense and complex structure of five genes. *Mamm. Genome* **11:** 1127–1131.

Dawkins, R., Leelayuwat, C., Gaudieri, S., Tay, G., Hui, J., Cattley, S., Martinez, P., and Kulski, J. 1999. Genomics of the major histocompatibility complex: Haplotypes, duplication, retroviruses and disease. *Immunol Rev.* **167:** 275–304.

Di Palma, F., Archibald, S.D., Young, J.R., and Ellis, S.A. 2002. A BAC contig of approximately 400 kb contains the classical class I major histocompatibility complex (MHC) genes of cattle. *Eur. J. Immunogenet* **29:** 65–68.

Dokmanovic, M., Chang, B.D., Fang, J., and Roninson, I.B. 2002. Retinoid-induced growth arrest of breast carcinoma cells involves coactivation of multiple growth-inhibitory genes. *Cancer Biol. Ther.* **1:** 24–27.

Ewing, B., and Green, P. 1998. Base-calling of automated sequencer traces using *Phred*. II. Error probabilities. *Genome Res.* **8:** 186–194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using *Phred*. I. Accuracy assessment. *Genome Res.* **8:** 175–185.

Eyre-Walker, A. and Hurst, L.D. 2001. The evolution of isochores. *Nat. Rev. Genet.* **2:** 549–555.

Fischer Lindahl, K. 1997. On naming *H2* haplotypes: Functional significance of MHC class Ib alleles. *Immunogenetics* **46:** 53–62.

Fischer Lindahl, K., Byers, D.E., Dabhi, V.M., Hovik, R., Jones, E.P., Smith, G.P., Wang, C.R., Xiao, H., and Yoshino, M. 1997. H2-M3, a full-service class Ib histocompatibility antigen. *Annu. Rev. Immunol.* **15:** 851–879.

Gardinier, M.V., Amiguet, P., Linington, C., and Matthieu, J.M. 1992. Myelin/oligodendrocyte glycoprotein is a unique member of the immunoglobulin superfamily. *J. Neurosci. Res.* **33:** 177–187.

Geraghty, D.E., Koller, B.H., Pei, J., and Hansen, J.A. 1992. Examination of four HLA class I pseudogenes. Common events in the evolution of HLA genes and pseudogenes. *J. Immunol.* **149:** 1947–1956.

Gordon, D., Abajian, C., and Green, P. 1998. *Consed*: A graphical tool for sequence finishing. *Genome Res.* **8:** 195–202.

Grifa, A., Totaro, A., Rommens, J.M., Carella, M., Roetto, A., Borgato, L., Zelante, L., and Gasparini, P. 1998. GABA (γ-amino-butyric acid) neurotransmission: Identification and fine mapping of the human GABAB receptor gene. *Biochem. Biophys. Res. Commun.* **250:** 240–245.

Günther, E. and Walter, L. 2001. The major histocompatibility complex of the rat (*Rattus norvegicus*). *Immunogenetics* **53:** 520–542.

Ha, H., Howard, C.A., Yeom, Y.I., Abe, K., Uehara, H., Artzt, K., and Bennett, D. 1991. Several testis-expressed genes in the mouse t-complex have expression differences between wild-type and t-mutant mice. *Dev. Genet.* **12:** 318–332.

Harada, H., Harada, Y., O'Brien, D.P., Rice, D.S., Naeve, C.W., and Downing, J.R. 1999. HERF1, a novel hematopoiesis-specific RING finger protein, is required for terminal differentiation of erythroid cells. *Mol. Cell. Biol.* **19:** 3808–3815.

Howcroft, T.K., Weissman, J.D., Rudikoff, S., Frels, W.I., and Singer, D.S. 1996. Repression of the nonclassical MHC class I gene H2-M1 by *cis*-acting silencer DNA elements. *Immunogenetics* **44:** 268–274.

Hughes, A.L. 1991. Independent gene duplications, not concerted evolution, explain relationships among class I MHC genes of murine rodents. *Immunogenetics* **33:** 367–373.

Hughes, A.L. and Nei, M. 1989. Evolution of the major histocompatibility complex: Independent origin of nonclassical class I genes in different groups of mammals. *Mol. Biol. Evol.* **6:** 559–579.

Ioannidu, S., Walter, L., Dressel, R., and Günther, E. 2001. Physical map and expression profile of genes of the telomeric class I gene region of the rat MHC. *J. Immunol.* **166:** 3957–3965.

Jarrous, N., Reiner, R., Wesolowski, D., Mann, H., Guerrier-Takada, C., and Altman, S. 2001. Function and subnuclear distribution of Rpp21, a protein subunit of the human ribonucleoprotein ribonuclease P. *RNA* **7:** 1153–1164.

Johnson, K.R., Cook, S.A., Bustin, M., and Davisson, M.T. 1992. Genetic mapping of the murine gene and 14 related sequences encoding chromosomal protein HMG-14. *Mamm. Genome* **3:** 625–632.

Jones, E.P., Xiao, H., Schultz, R.A., Flaherty, L., Trachtulec, Z., Vincek, V., Larin, Z., Lehrach, H., and Fischer Lindahl, K. 1995. MHC class I gene organization in >1.5-Mb YAC contigs from the *H2-M* region. *Genomics* **27:** 40–51.

Jones, E.P., Kumánovics, A., Yoshino, M., and Fischer Lindahl, K. 1999. *Mhc* class I and nonclass I gene organization in the proximal *H2-M* region of the mouse. *Immunogenetics* **49:** 183–195.

Kaufman, J., Milne, S., Gobel, T.W., Walker, B.A., Jacob, J.P., Auffray, C., Zoorob, R., and Beck, S. 1999. The chicken B locus is a minimal essential major histocompatibility complex. *Nature* **401:** 923–925.

Klein, J. and O'hUigin, C. 1994. The conundrum of nonclassical major histocompatibility complex genes. *Proc. Natl. Acad. Sci.* **91:** 6251–6252.

Klein, J., Benoist, C., David, C.S., Demant, P., Fischer Lindahl, K., Flaherty, L., Flavell, R.A., Hammerling, U., Hood, L.E., Hunt III, S.W., et al. 1990. Revised nomenclature of mouse *H-2* genes. *Immunogenetics* **32:** 147–149.

Klein, J., Ono, H., Klein, D., and O'hUigin, C. 1992. The accordion model of *Mhc* evolution. In *Progress in Immunology* (ed. J. Gergely), Vol. III, pp. 137–143. Springer, Berlin, Germany.

Klein, J., Sato, A., and O'hUigin, C. 1998. Evolution by gene duplication in the major histocompatibility complex. *Cytogenet. Cell Genet.* **80:** 123–127.

Kulski, J.K., Gaudieri, S., Bellgard, M., Balmer, L., Giles, K., Inoko, H., and Dawkins, R.L. 1997. The evolution of MHC diversity by segmental duplication and transposition of retroelements. *J. Mol. Evol.* **45:** 599–609.

Kumánovics, A., Madan, A., Qin, S., Rowen, L., Hood, L., and Fischer Lindahl, K. 2002. *Quod erat faciendum:* sequence analysis of the *H2-D* and *H2-Q* regions of 129/SvJ mice. *Immunogenetics* **54:** 479–489.

Kumánovics, A., Takada, T., and Fischer Lindahl, K. 2003. Genomic organization of the mammalian *Mhc*. *Annu. Rev. Immunol.* **21:** 629–657.

Kumar, S., Tamura, K., Jakobsen, I.B., and Nei, M. 2001. MEGA2: Molecular evolutionary genetics analysis software. *Bioinformatics* **17:** 1244–1245.

Kundu, T.K. and Rao, M.R. 1999. CpG islands in chromatin organization and gene expression. *J. Biochem. (Tokyo)* **125:** 217–222.

Kuroda, N., Figueroa, F., O'hUigin, C., and Klein, J. 2002. Evidence that the separation of *Mhc* class II from class I loci in the zebrafish, *Danio rerio*, occurred by translocation. *Immunogenetics* **54:** 418–430.

Lehner, B., Williams, G., Campbell, R.D., and Sanderson, C.M. 2002. Antisense transcripts in the human genome. *Trends Genet.* **18:** 63–65.

Leishman, A.J., Naidenko, O.V., Attinger, A., Koning, F., Lena, C.J., Xiong, Y., Chang, H.C., Reinherz, E., Kronenberg, M., and Cheroutre, H. 2001. T cell responses modulated through interaction between CD8αα and the nonclassical MHC class I molecule, TL. *Science* **294:** 1936–1939.

Lennon-Duménil, A.M., Bakker, A.H., Wolf-Bryant, P., Ploegh, H.L., and Lagaudrière-Gesbert, C. 2002. A closer look at proteolysis and MHC-class-II-restricted antigen presentation. *Curr. Opin. Immunol.* **14:** 15–21.

Lepourcelet, M., Andrieux, N., Giffon, T., Pichon, L., Hampe, A., Galibert, F., and Mosser, J. 1996. Systematic sequencing of the human HLA-A/HLA-F region: Establishment of a cosmid contig and identification of a new gene cluster within 37 kb of sequence. *Genomics* **37:** 316–326.

Lepourcelet, M., Coriton, O., Hampe, A., Galibert, F., and Mosser, J. 1998. *HTEX4*, a new human gene in the MHC class I region, undergoes alternative splicing and polyadenylation processes in testis. *Immunogenetics* **47:** 491–496.

Mashimo, H., Chorney, M.J., Pontarotti, P., Fisher, D.A., Hood, L., and Nathenson, S.G. 1992. Nucleotide sequence of the BALB/c H-2T region gene, T3d. *Immunogenetics* **36:** 326–332.

Matsuo, M.Y., Asakawa, S., Shimizu, N., Kimura, H., and Nonaka, M. 2002. Nucleotide sequence of the MHC class I genomic region of a teleost, the medaka (*Oryzias latipes*). *Immunogenetics* **53:** 930–940.

MHC Sequencing Consortium, 1999. Complete sequence and gene map of a human major histocompatibility complex. *Nature* **401:** 921–923.

Nair, R.P., Stuart, P., Henseler, T., Jenisch, S., Chia, N.V., Westphal, E., Schork, N.J., Kim, J., Lim, H.W., Christophers, E., et al. 2000. Localization of psoriasis-susceptibility locus PSORS1 to a 60-kb interval telomeric to HLA-C. *Am. J. Hum. Genet.* **66:** 1833–1844.

Nei, M., Gu, X., and Sitnikova, T. 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc. Natl. Acad. Sci.* **94:** 7799–7806.

Obata, Y., Chen, Y.T., Stockert, E., and Old, L.J. 1985. Structural analysis of TL genes of the mouse. *Proc. Natl. Acad. Sci.* **82:** 5475–5479.

Ohta, Y., Okamura, K., McKinney, E.C., Bartl, S., Hashimoto, K., and Flajnik, M.F. 2000. Primitive synteny of vertebrate major histocompatibility complex class I and class II genes. *Proc. Natl. Acad. Sci.* **97:** 4712–4717.

Oka, A., Tamiya, G., Tomizawa, M., Ota, M., Katsuyama, Y., Makino, S., Shiina, T., Yoshitome, M., Iizuka, M., Sasao, Y., et al. 1999. Association analysis using refined microsatellite markers localizes a susceptibility locus for psoriasis vulgaris within a 111-kb segment telomeric to the HLA-C gene. *Hum. Mol. Genet.* **8:** 2165–2170.

Old, L.J., Boyse, E.A., and Stockert, E. 1963. Antigenic properties of experimental leukemias. I. Serological studies in vitro with spontaneous and radiation-induced leukemias. *J. Natl. Cancer Inst.* **31:** 977–986.

Orimo, A., Yamagishi, T., Tominaga, N., Yamauchi, Y., Hishinuma, T., Okada, K., Suzuki, M., Sato, M., Nogi, Y., Suzuki, H., et al. 2000. Molecular cloning of testis-abundant finger protein/ring finger protein 23 (RNF23), a novel RING-B box-coiled coil-B30.2 protein on the class I region of the human MHC. *Biochem. Biophys. Res. Commun.* **276:** 45–51.

Pham-Dinh, D., Jones, E.P., Pitiot, G., Della Gaspera, B., Daubas, P., Mallet, J., Le Paslier, D., Fischer Lindahl, K., and Dautigny, A. 1995. Physical mapping of the human and mouse MOG gene at the distal end of the MHC class Ib region. *Immunogenetics* **42:** 386–391.

Pichon, L., Carn, G., Bouric, P., Giffon, T., Chauvel, B., Lepourcelet, M., Mosser, J., Legall, J.Y., and David, V. 1996. Structural analysis of the HLA-A/HLA-F subregion: Precise localization of two new multigene families closely associated with the HLA class I sequences. *Genomics* **32:** 236–244.

Pontarotti, P.A., Mashimo, H., Zeff, R.A., Fisher, D.A., Hood, L., Mellor, A., Flavell, R.A., and Nathenson, S.G. 1986. Conservation and diversity in the class I genes of the major histocompatibility complex: Sequence analysis of a Tla^{b} gene and comparison with a Tla^{c} gene. *Proc. Natl. Acad. Sci.* **83:** 1782–1786.

Renard, C., Vaiman, M., Chiannilkulchai, N., Cattolico, L., Robert, C., and Chardon, P. 2001. Sequence of the pig major histocompatibility region containing the classical class I genes. *Immunogenetics* **53:** 490–500.

Reymond, A., Meroni, G., Fantozzi, A., Merla, G., Cairo, S., Luzi, L., Riganelli, D., Zanaria, E., Messali, S., Cainarca, S., et al. 2001. The tripartite motif family identifies cell compartments. *EMBO J.* **20:** 2140–2151.

Rogel-Gaillard, C., Piumi, F., Billault, A., Bourgeaux, N., Save, J.C., Urien, C., Salmon, J., and Chardon, P. 2001. Construction of a rabbit bacterial artificial chromosome (BAC) library: Application to the mapping of the major histocompatibility complex to position 12q1.1. *Mamm. Genome* **12:** 253–255.

Saurin, A.J., Borden, K.L., Boddy, M.N., and Freemont, P.S. 1996. Does this have a familiar RING? *Trends Biochem. Sci.* **21:** 208–214.

Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. 2000. PipMaker—A web server for aligning two genomic DNA sequences. *Genome Res.* **10:** 577–586.

Shendure, J. and Church, G.M. 2002. Computational discovery of sense–antisense transcription in the human and mouse genomes. *Genome Biol.* **3:** research00044.00041–00014.

Shiina, T., Shimizu, C., Oka, A., Teraoka, Y., Imanishi, T., Gojobori, T., Hanzawa, K., Watanabe, S., and Inoko, H. 1999a. Gene organization of the quail major histocompatibility complex (*MhcCoja*) class I gene region. *Immunogenetics* **49:** 384–394.

Shiina, T., Tamiya, G., Oka, A., Takishima, N., Yamagata, T., Kikkawa, E., Iwata, K., Tomizawa, M., Okuaki, N., Kuwano, Y., et al. 1999b. Molecular dynamics of MHC genesis unraveled by sequence analysis of the 1,796,938-bp HLA class I region. *Proc. Natl. Acad. Sci.* **96:** 13282–13287.

Singer, D.S., Hare, J., Golding, H., Flaherty, L., and Rudikoff, S. 1988. Characterization of a new subfamily of class I genes in the H-2 complex of the mouse. *Immunogentics* **28:** 13–21.

Sipes, S.L., Medaglia, M.V., Stabley, D.L., DeBruyn, C.S., Alden, M.S., Catenacci, V., and Landel, C.P. 1996. A new major histocompatibility complex class Ib gene expressed in the mouse blastocyst and placenta. *Immunogenetics* **45:** 108–120.

Smit, A.F. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9:** 657–663.

Snell, G.D. and Higgins, G.F. 1951. Alleles at the histocompatibility-2 locus in the mouse as determined by tumor transplantation. *Genetics* **36:** 306–310.

Snell, G.D., Graff, R.J., and Cherry, M. 1971. Histocompatibility genes of mice. XI. Evidence establishing a new histocompatibility locus, *H-12*, and new *H-2* allele, *H-2*^{bc}. *Transplantation* **11:** 525–530.

Spies, T. 2002. Induction of T cell alertness by bacterial colonization of intestinal epithelium. *Proc. Natl. Acad. Sci.* **99:** 2584–2586.

Stroynowski, I. and Fischer Lindahl, K. 1994. Antigen presentation by nonclassical class I molecules. *Curr. Opin. Immunol.* **6:** 38–44.

Teitell, M., Cheroutre, H., Panwala, C., Holcombe, H., Eghtesady, P., and Kronenberg, M. 1994. Structure and function of *H-2 T* (*Tla*) region class I MHC molecules. *Crit. Rev. Immunol.* **14:** 1–27.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.

Trowsdale, J. 1995. "Both man & bird & beast": Comparative organization of MHC genes. *Immunogenetics* **41:** 1–17.

Wang, C.-R. and Fischer Lindahl, K. 1993. Organization and structure of the *H-2M4-M8* class I genes in the mouse major histocompatibility complex. *Immunogenetics* **38:** 258–271.

Wang, C.-R., Castaño, A.R., Peterson, P.A., Slaughter, C., Fischer Lindahl, K., and Deisenhofer, J. 1995a. Nonclassical binding of formylated peptide in crystal structure of the MHC class Ib molecule H2-M3. *Cell* **82:** 655–664.

Wang, C.-R., Lambracht, D., Wonigeit, K., Howard, J.C., and Fischer Lindahl, K. 1995b. Rat RT1 orthologs of mouse H2-M class Ib genes. *Immunogenetics* **42:** 63–67.

Wu, M., van Kaer, L., Itohara, S., and Tonegawa, S. 1991. Highly restricted expression of the thymus leukemia antigens on intestinal epithelial cells. *J. Exp. Med.* **174:** 213–218.

Yeager, M., Kumar, S., and Hughes, A.L. 1997. Sequence convergence in the peptide-binding region of primate and rodent MHC class Ib molecules. *Mol. Biol. Evol.* **14:** 1035–1041.

Yewdell, J.W. and Bennink, J.R. 2001. Cut and trim: Generating MHC class I peptide ligands. *Curr. Opin. Immunol.* **13:** 13–18.

Younger, R.M., Amadou, C., Bethel, G., Ehlers, A., Fischer Lindahl, K., Forbes, S., Horton, R., Milne, S., Mungall, A.J., Trowsdale, J., et al. 2000. Characterization of clustered MHC-linked olfactory receptor genes in human and mouse. *Genome Res.* **11:** 519–530.

Zhou, Y. and Chaplin, D.D. 1993. Identification in the HLA class I region of a gene expressed late in keratinocyte differentiation. *Proc. Natl. Acad. Sci.* **90:** 9470–9474.

WEB SITE REFERENCES

http://ftp.genome.washington.edu/cgi-bin/RepeatMasker; RepeatMasker home page.

http://www.ncbi.nlm.nih.gov/BLAST; NCBI BLAST home page.

http://www.emboss.org; EMBOSS (European Molecular Biology Software Suite) home page.

http://genes.mit.edu/GENSCAN.html; GENSCAN home page.

http://bio.cse.psu.edu/pipmaker; PIPMAKER home page.

Received November 7, 2002; accepted in revised form January 28, 2003.