Analysis of the Gene-Dense Major Histocompatibility Complex Class III Region and Its Comparison to Mouse

Tao Xie,[1,4,7] Lee Rowen,[1,7] Begoña Aguado,[2,5] Mary Ellen Ahearn,[3] Anup Madan,[1,6] Shizhen Qin,[1] R. Duncan Campbell,[2] and Leroy Hood[1,8]

[1]*Institute for Systems Biology, Seattle, Washington 98103, USA;* [2]*MRC Rosalind Franklin Center for Genomics Research (formerly HGMP Resource Center), Hinxton, Cambridge CB10 1SB, UK;* [3]*Department of Pediatrics, University of Miami School of Medicine, Miami, Florida 33136, USA*

In mammals, the Major Histocompatibility Complex class I and II gene clusters are separated by an ~700-kb stretch of sequence called the MHC class III region, which has been associated with susceptibility to numerous diseases. To facilitate understanding of this medically important and architecturally interesting portion of the genome, we have sequenced and analyzed both the human and mouse class III regions. The cross-species comparison has facilitated the identification of 60 genes in human and 61 in mouse, including a potential RNA gene for which the introns are more conserved across species than the exons. Delineation of global organization, gene structure, alternative splice forms, protein similarities, and potential *cis*-regulatory elements leads to several conclusions: (1) The human MHC class III region is the most gene-dense region of the human genome: >14% of the sequence is coding, ~72% of the region is transcribed, and there is an average of 8.5 genes per 100 kb. (2) Gene sizes, number of exons, and intergenic distances are for the most part similar in both species, implying that interspersed repeats have had little impact in disrupting the tight organization of this densely packed set of genes. (3) The region contains a heterogeneous mixture of genes, only a few of which have a clearly defined and proven function. Although many of the genes are of ancient origin, some appear to exist only in mammals and fish, implying they might be specific to vertebrates. (4) Conserved noncoding sequences are found primarily in or near the 5′-UTR or the first intron of genes, and seldom in the intergenic regions. Many of these conserved blocks are likely to be *cis*-regulatory elements.

[Supplemental material is available online at www.genome.org and http://www.systemsbiology.org. The nucleotide sequences of human and mouse cosmid or BAC clones have been submitted to GenBank as a series of separate entries. The accession numbers are as follows: AC007080, AF109719, AF109905, AF109906, AF049850, and AF030001 (mouse); and AF129756, AF134726, AF019413, U89337, U89336, and U89335 (human). The following individuals kindly provided reagents, samples, or unpublished information as indicated in the paper: T. Spies.]

The ~4-Mb human major histocompatibility complex (MHC) on Chromosome 6p21.3 contains genes encoding the highly polymorphic class I and II MHC polypeptides required for the presentation of antigenic peptides to T-cells in the adaptive immune response (Beck and Trowsdale 2000). In addition to these narrowly defined MHC class I and II genes, the MHC region or locus contains hundreds of other genes and pseudogenes (The MHC Sequencing Consortium 1999). The MHC class III region consists of a dense array of genes sandwiched between the MHC class I and II regions in primates and rodents. Some of these genes, for example, the complement fixation genes C4, C2, and factor B, play a role in the innate as opposed to the adaptive immune system. Others, for example, valine tRNA synthetase, appear to have no specialized function in the immune response but instead play other key roles in the life of a cell.

Present addresses: [4]Hartwell Center for Bioinformatics and Biotechnology, St. Jude's Children's Research Hospital, Memphis, TN 38105, USA; [5]Centro Nacional de Biotecnologia (CNB), CSIC Campus Universidad Autonoma 28049, Madrid, Spain; [6]Neurogenomics Research Laboratory, University of Iowa, Iowa City, IA 52246, USA.
[7]These authors contributed equally to this work.
[8]Corresponding author.
E-MAIL lhood@systemsbiology.org; FAX (206) 732-1254.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1736803.

Linkage analysis studies in various populations have identified the MHC class III region as a likely target for predisposing genes for several diseases, including autoimmune diseases such as type 1 diabetes (Nishimura et al. 2003), rheumatoid arthritis (Okamoto et al. 2003), and lupus erythematosus (Gruen and Weissman 2001). However, in addition to the conundra associated with replicating marker association results in different patient populations, there is the added challenge of identifying which specific gene or set of genes is responsible for a given disease process. As of yet, only a handful of clear-cut linkages of gene-to-function-to-disease have been established. These include congenital adrenal hyperplasia, which is associated with mutations in the cytochrome P450 *CYP21A2* gene that is required for cortisol biosynthesis (Chiou et al. 1990), and Ehlers-Danlos Syndrome, a connective tissue disorder of which one variant is caused by mutations in tenascin X (*TNX*; Burch et al. 1997). For many other diseases, for example, narcolepsy (Miyagawa et al. 2000) and myocardial infarction (Ozaki et al. 2002), candidate genes in the MHC class III region are offered, but the causal relationship is unproven.

To facilitate research into candidate gene identification and function, we have sequenced and analyzed the human and mouse class III portion of the MHC locus. In contrast to the human genome as a whole, for which the average gene size is

thought to vary from ~27 kb (International Human Genome Sequencing Consortium 2001) to >45 kb (Heilig et al. 2003; Scherer et al. 2003) and the number of genes per megabase is typically <11 (International Human Genome Sequencing Consortium 2001), we report here that the MHC class III region contains 60–61 genes in ~700 kb, with an average gene size of ~8.5 kb and an average intergenic distance of just under 3 kb. It is the most gene-dense region in the human genome. Not surprisingly, computational gene prediction programs such as GenScan (Burge and Karlin 1997) encounter difficulties in regions of this sort: With such extensive transcription, one real gene may be split into two predicted genes, or two or three real genes combined into one predicted gene. In principle, given the numbers of full-length cDNA sequences now available in the public databases (Okazaki et al. 2002; Strausberg et al. 2002), a more accurate identification of individual genes can be made, on the assumption that the transcriptional machinery in the cell correctly reads the start and stop signals so as not to create read-through transcripts from adjacent genes. To address this issue and, more generally, to understand better the dynamics at work in gene-dense regions of the genome likely to be rife with disease associations, we have undertaken an analysis of the human and mouse MHC class III region as a paradigm case.

RESULTS AND DISCUSSION

Identification of the Most Gene-Dense Regions of the Human Genome

Gene-dense regions of the genome are characterized by a large gene count per unit length of genomic sequence. Using a megabase (Mb) as the unit length, with a sliding window offset of 250 kb, the regions of the human genome with the highest gene count were determined (Table 1). Regions that scored high in successive 250-kb offsets were combined into one longer region. Because the gene count has been based on genomic alignments of the longest reviewed or provisional RefSeq associated with the LocusLink entry for a given gene, the actual number of genes in these regions is likely to be somewhat higher, as not all of the genes are represented in RefSeq. It is also possible that some gene-dense regions have been missed owing to acquisition bias inherent to the use of RefSeqs for counting genes. Finally, regions such as protocadherins and T-cell receptors that contain rearranging gene elements strewn across hundreds of kilobases of sequence have been eliminated from the analysis as, in a sense, these can be considered to be one gene. With these caveats, the MHC class III region is established as the most gene-dense region of the human genome (Table 1). In the 1.25-Mb region at 6p21.3 that scored highest in gene density, all but eight of the genes were located within the 700-kb class III region.

The most gene-dense regions of the genome fall into two categories. Seven of the sequences contain a heterogeneous mixture of closely spaced, smaller than average-sized genes with disparate functions. In addition to the MHC class III, the regions on Chromosome 12 (Ansari-Lari et al. 1996), Chromosome 16 (Daniels et al. 2001), and the X-chromosome (Chen et al. 1996) have previously been characterized as being remarkably gene-dense (Table 1). Unlike the "hodgepodge" type of gene organization, three of the sequences contain functionally related multigene families that have expanded their membership through genomic duplication: histones on Chromosome 6 (Albig and Doenecke 1997), hair protein keratins on Chromosome 17 (Rogers et al. 1997), and the leukocyte receptor cluster on Chromosome 19 (Wende et al. 2000). Three of the most gene-dense regions lie at the subtelomeric regions of the chromosome, one each on Chromosomes 11, 16, and the X-chromosome. One particularly dense portion of Chromosome 19 spans almost 2 Mb.

Although generally GC-rich, the most gene-dense megabases of the genome are not those with the highest GC content. Regions with more than 20 genes per megabase are scattered across a wide range of GC percentage, from 37% to 60% (Fig. 1). Likewise, there is significant variation in the interspersed-repeat content of the gene-dense regions (Table 1). The region on Chromosome 16 presents the most extreme case, with 60% GC and 28% interspersed-repeat content. The ~700-kb MHC class III region, taken alone, has an overall GC content of 51% and an overall interspersed-repeat content of 36.7% in human and 29.1% in mouse (Supplemental Table 1, available online at www.genome.org). The frequency of SINEs (Short Interspersed Nuclear Elements) in the MHC class III region is significantly higher than other repeat classes for both human (24.68%) and mouse (15.12%), which is consistent with what has been found for many other GC-rich, gene-dense regions of the genome (Mouse Genome Sequencing Consortium 2002).

Global Comparison of the Human and Mouse MHC Class III Region

The MHC class III region was so named because it lies between the class I and II genes in mammalian genomes. Prior to the genomic sequencing, the boundaries of the class III region were ill-defined. For example, several genes of unknown function (e.g., *BAT1*, *BAT2*) were designated as HLA B-associated transcripts, in light of their proximity to the class I *HLAB* gene (Spies et al. 1989), leaving it open as to whether they belonged to class I or class III. That there is in fact a solid contextual basis for

Table 1. Human Genome Top 10 Gene-Dense Regions

GoldenPath location	Region	%GC	% repeats	Genes/Mb	Comments
chr6:31250001–32500000	HLAC–HLADRB3	47	47	48.8	Includes MHC class III region
chr6:25500001–26500000	FLJ20048–BTN2A3	41	43	44.0	Includes histone families
chr12:6250001–7250000	FLJ10665–PXR1	46	41	43.1	Includes CD4, complement 1
chr17:39000001–40000000	KRT23–ACLY	46	44	43.0	Includes keratin families
chr19:53250001–55000000	ELSPBP1–TCBAP0758	52	57	42.3	Includes CD37
chr16:250001–1500000	DKFZP761D0211–KIAA0683	60	28	40.8	GC rich
chr11:250001–1500000	AP2A2–HCCA2	53	36	40.2	Gap in sequence; includes IRF7, TOLLIP
chr17:7000001–8000000	ASGR1–PER1	51	43	39.0	Includes TNSF12, 13; CD68; TP53
chrX:150500001–151500000	DUSP9–GAB3	53	43	39.0	Includes G6PD; IRAK1
chr19:59250001–60250000	OSCAR–RDH13	49	53	36.0	Includes KIR, ILT, LILR families

Using a window offset of 250 kb, the number of genes per megabase and GC content were calculated as described in Figure 1. If a region appeared in the top 20 hits more than once (e.g., chr16:250001–250000 and chr16:5000001–1,500000), the regions were combined. "Region" indicates the outermost genes within the GoldenPath span.

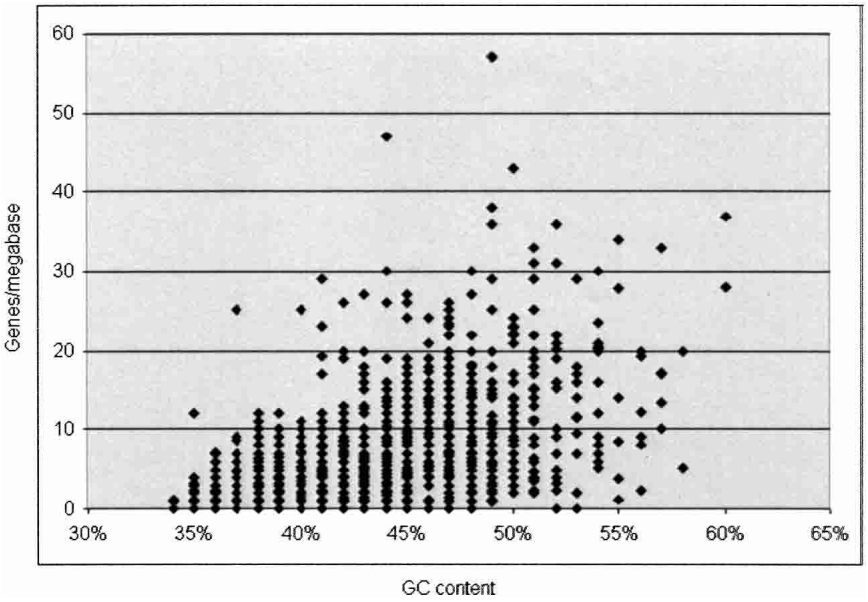

Figure 1 Using LocusLink entries for reviewed and provisional RefSeqs belonging to a given gene, and assigning the longest alignable RefSeq to the gene so that each gene is counted only once, the number of genes per megabase and GC content were plotted for the human genome using the April 2003 GoldenPath assembly (http://genome.cse.ucsc.edu). The nonoverlapping megabase-sized windows begin at 0, 250,000, 500,000, and 750,000 for each chromosome. Results for the window beginning at 250,000 are shown. The uppermost point on the graph represents Chromosome 6:31250001–32250000, which includes the entire MHC class III region.

defining the class III region is shown in a dot-plot comparison of the human and mouse sequences (Fig. 2). Two landscape features set off the class III region from its genomic surroundings: a rise in GC content demarcating a GC-rich isochore; and a well-defined block of conserved sequences between human and mouse that extends across the same 700-kb region. As will be discussed below, these turn out to be mostly exons. With one exception, the human and mouse sequences are nearly collinear across the entire length of the GC-rich isochore, indicating little disruption of overall gene organization by interspersed repeats or genomic deletions. The exception consists of a 40-kb sequence consisting in part of an old (27.9% divergent) L1MA6 repeat found in the mouse genome between the two copies of the complement *C4-CYP21* duplication (Fig. 2; Yang et al. 1999). With these data in hand, the boundaries of the class III region are most naturally defined as lying just outside the two genes found at the ends of the conserved region—*BAT1*, near class I, and *NOTCH4*, near class II.

Since divergence of human and mouse from the last common ancestor, class III has remained a relatively stable region of the genome, standing in marked contrast to the dynamically evolving neighboring regions. On the telomeric side of *BAT1*, the human and mouse sequences diverge because of differences in the gene duplications and copy number of the class I-associated gene. Between *BAT1* and *HLAB*, the human genome has two copies of the *MIC* gene, which is not found in mouse. On the other hand, in place of the two class I genes found in human—*HLA-B* and *HLA-C*—mouse contains multiple copies of the class I *H2Q* and one copy of *H2D* (Kumanovics et al. 2002). On the centromeric side of *NOTCH4*, mouse has expanded the copy number of the butyrophilin gene family in comparison to human (Stammers et al. 2000). Human, while having only 1 butyrophilin gene, has extensively duplicated the class II *DR*β genes. The theme of a conserved class III cluster of genes relative to the dynamically evolving adjacent class I and II regions is repeated in the rhesus monkey sequence as well (D. Geraghty and L. Rowen, unpubl.). Whole-genome comparisons between human and mouse (Mouse Genome Sequencing Consortium 2002) indicate that the conserved class III region fits the norm, and the divergent class I and II constitute exceptions, which would be consistent with the finding that immune-response genes are well represented in the group of rapidly evolving genes (Mouse Genome Sequencing Consortium 2002).

To more precisely identify conserved segments, the human and mouse MHC class III sequences were aligned using the GLASS program (Batzoglou et al. 2000), and the resulting alignment was visualized using the VISTA tool (Fig. 3; Mayor et al. 2000). The alignment program in PIPMaker (Schwartz et al. 2000), based on local alignments instead of a global alignment, revealed no significant differences with the GLASS alignment (data not shown). The genes in the class III region (see below; Fig. 3) have been identified and mapped onto the alignment. As expected, there are discrete regions with different levels of similarities across the class III region. Most exons are conserved between the orthologous genes, but most introns and intergenic regions cannot be aligned. Some noncoding regions, mainly in the 5′-UTRs or in the first introns of individual genes, are also conserved. These conserved noncoding sequences (CNSs) may be of great biological significance if they contain gene regulatory elements (discussed below). Regions containing interspersed repeats were not alignable for the most part. However, four of the MIRs (mammalian interspersed repeats) were conserved, implying that insertion of these particular repeats predated the divergence between human and mouse ~80 million years ago.

Organization, Function, and Regulation of the Class III Region Genes

Identification of Genes

When the MHC class III sequences were obtained initially, full-length cDNAs existed for only a small subset of the genes. Therefore, GenScan, ESTs, and predicted protein similarity were used to identify genes and propose transcript models. In this regard, having the genomic sequences from both human and mouse was enormously useful for providing support for the gene models, in that potential exons surrounded by splice sites could be identified from conserved sequence blocks in a local region.

The recent explosion of cDNA and EST sequences has made the task of gene identification both more precise and more confusing: more precise because predicted exons could be confirmed or discounted and more confusing because of "read-through" transcripts between pairs of adjacent genes. For example, cDNAs and ESTs exist that conjoin *MSH5* and *C6orf26*, *G6F* and *LY6G6D*, and *PPT2* and *EGFL8*, among others (Table 2). In a known gene such as *MSH5*, where several 3′ ESTs contain the polyadenylation signal that terminates the transcript, sorting out the genes is straightforward. But for novel genes or genes of unknown function, the read-through transcripts could indicate a single gene with alternative splicing, alternative promoters, and alternative

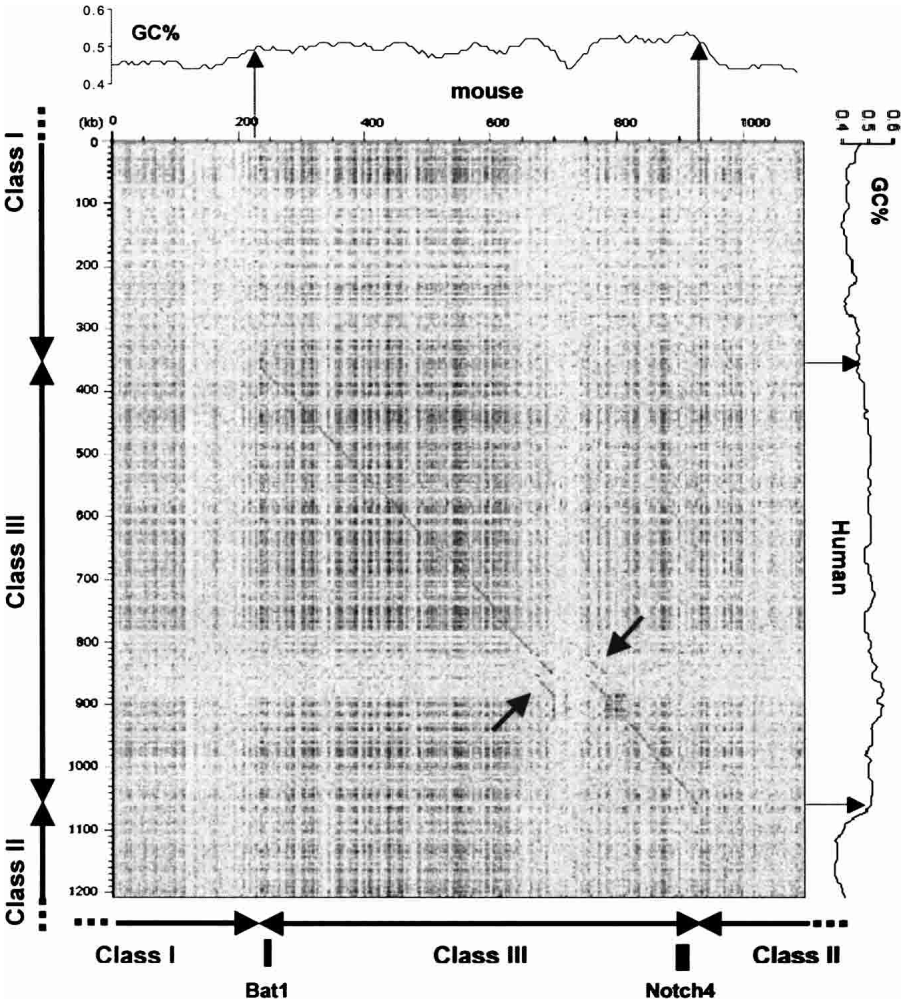

Figure 2 DOTTER (version 3.1, default setting) dot-matrix comparison of the extended human and mouse MHC class III regions. The *X*-axis represents the mouse sequence, the *Y*-axis the human. The depicted human and mouse sequences begin at *POU5F1* (in the class I region, for both human and mouse) and extend to *TSBP*. The location of the class III region to class I and II is indicated using arrows outside the *top X*-axis and *left Y*-axis. Only a portion of the human and mouse class I and II regions is included as indicated by the open arrows. GC contents were also determined and are shown *above* the axes. Arrows in the GC plots define the class III region as a GC-rich isochore. In mouse, the locations of the first class III gene, *Bat1*, and the last class III gene, *Notch4*, are shown as black boxes. There are a pair of short diagonal lines (see the arrows inside dot plots) away from the main diagonal axis, that correspond to the C4-CYP21 module (see text). Interspersed repeats are represented by the background dots.

terminations rather than two genes. *C6orf31*, for example, is supported by three cDNAs (BC013201, AK054885, AL050203) and several ESTs (e.g., BM544221, BI765381, BM786281, AW134492), which give different gene models with unrelated translations, depending on which subset of exons is used for the model. However, only some of the exons predicted by the spliced mRNAs are present in mouse; hence, we based the gene model on those exons in common between the two species. Additional complications in constructing gene models come from exons with nonstandard splice sites (e.g., *C6orf29*), clusters of spliced ESTs with no obvious translation (*C6orf48*), and genes with BLASTX similarity to other genes but which lack sufficient cDNA/ESTs to construct a definitive model (e.g., *TNXB*).

C6orf48 (mouse *G8*) provides an interesting example of the usefulness of the cross-species comparison for defining a gene model. This gene is present between the *NEU1* and *HSPA1B* genes in both species. Human *C6orf48* consists of four exons that encode a putative 75-amino-acid polypeptide, whereas the mouse ortholog has five exons that encode a smaller polypeptide. No significant sequence similarity could be detected either in their coding DNA sequences or in their protein sequences. Therefore, they seemed unrelated at first glance. However, these two genes share two conserved noncoding sequences (CNSs), both in introns. In them, CNS1 can be transcribed as part of a 5′-UTR in some alternatively spliced isoforms both in human and in mouse, whereas CNS2 seems always to be transcribed (Fig. 4). Human ESTs containing CNS1 were obtained from CD34+ hematopoietic stem/progenitor cells (two ESTs) and parathyroid tumor (one EST), respectively. These may indicate the tissue-specific expression of different alternatively spliced variants. After searching GenBank using the two CNSs, we found a small nuclear RNA, U52 (Kiss-Laszlo et al. 1996), encoded in CNS1, and another small nuclear RNA, U48 (Kiss-Laszlo et al. 1996), encoded in CNS2. U52 and U48 are RNA components of snoRNPs that play roles in rRNA maturation (Mattaj 1993). These two introns are much more conserved than the coding regions of the corresponding genes (Eddy 1999), indicating that human *C6orf48* and mouse *G8* are really orthologs, with the two small nuclear RNAs being the main products encoded by these genes. It is unlikely that the putative protein-coding regions are themselves physiologically relevant.

By integrating analyses based on ab initio gene prediction, similarity searches and conserved regions revealed by genomic comparison, we have identified 60 genes in human and 61 genes in mouse (see Fig. 3; details are in Supplemental Table 2). There are only two pseudogenes (*CYP21A1* and *LY6G6E*; Mallya et al. 2002) in human and three (*Cyp21a2-ps*, *Slp*, and *Ncr3*) in mouse, in contrast with large numbers of pseudogenes in the class I and class II regions (Fig. 3; The MHC Sequencing Consortium 1999). We also identified a processed pseudogene with a high degree of similarity to the human *NHP2L1* gene (nonhistone chromosome protein 2-like 1; Saito et al. 1996), in the intergenic region between the *G8* and the *Hspa1b* genes in mouse. There is not an equivalent processed pseudogene in the human MHC class III region. *NHP2L1* and the two gene fragments *TNXA* and *RP2* (see location in Fig. 3) are not counted as genes in the following analyses.

Gene Organization

The average sizes of the MHC class III coding sequences are 1.79 kb (human) and 1.71 kb (mouse); the 5′-UTRs are 223 bp (human) and 171 bp (mouse); and the 3′-UTRs are 413 bp (human) and 355 bp (mouse; see Table 2 and Supplemental Table 2). The average numbers of exons in the human and mouse genes in this region are 11.2 and 11.0, respectively. We found that the num-

Table 2. Features of the Human Class III Genes

Human gene	mRNA length	Evidence for gene model	ATG exon	CDS length	Alternative promoter	Alternative splice forms	Read-through transcripts	Notes
BAT1	2128	mRNA/ESTs	1 or 2	1284	Start at exon 1 or 2	Alternative exon 2 length	mRNA/ESTs	
ATP6V1G2	1631	ESTs	1 or 2	354	Start at exon 1 or 2	Alternative exon 1 length	mRNA/ESTs	May be brain-specific
NFKBIL1	1394	mRNA/ESTs	2	1074	Two alternative exon 1	No	No	Mouse may have an extra 3′ exon
LTA	1408	mRNA/ESTs	2	615	Two start sites for exon 1	Alternative exon 1 length; combine exons 2 and 3	No	
TNF	1675	mRNA/ESTs	1	699	No	No	No	
LTB	1542	mRNA/ESTS	1	918	No	Exon 2 skipped	No	3 or 4 exons
LST1	818	mRNA/ESTs	2	291	Four alternative exon 1	Exons 3 and 4 skipped	No	2, 3, 4, or 5 exons
NCR3	1043	mRNA/ESTS	1	603	No	Alternative location for exon 4	No	
AIF1	691	mRNA/ESTs	1	441	No	Alternative exon 1 length; other minor splice variants	No	
BAT2	6879	mRNA/ESTs	2	6468	No	Alternative exon 2 length	No	
BAT3	3696	mRNA/ESTs	2	3378	Four alternative exon 1	Alternative exon 7 length; exon 24 skipped in several ESTs	No	
APOM	854	mRNA/ESTs	2	670	Two alternative exon 1	No	No	
C6orf47	2476	mRNAs (2)	1	882	ND	No	No	Single exon gene
BAT4	1778	mRNA/ESTs	2 or 3	1068	Start at exon 2 or 3	Alternative exon 2 length	No	
CSNK2B	1055	mRNA/ESTs	2	645	Two alternative exon 1	No	No	
LY6G5B	2589	mRNA/ESTs	1 or 2	438	Start at exon 1 or 2	Exon 1 and 2 conjoined	No	
LY6G5C	951	mRNA/partial ESTs	5′ partial	675	Four alternative exon 1	Variable location for exon 2	No	No mouse ESTs/ATG start unclear
BAT5	2065	mRNA/ESTs	1	1674	Two alternative exon 1	No	No	
G6F	ND	GenScan	ND	873	ND	ND	mRNA	
LY6G6E	383	mRNAs (3)	ND	378	No	Exon 2 and 3 conjoined	mRNA	Probable pseudogene
LY6G6D	1266	mRNAs (3)	1	399	No	Alternative exon 3 length; exons 1 and 2 conjoined	mRNA	
LY6G6C	991	mRNA/ESTs	1	375	Two alternative exon 1	Extra exon in one EST	No	
C6orf25	2394	mRNAs (8)	1	711	No	Exons 3 and 4 skipped	No	
DDAH2	1685	mRNA/ESTs	2	855	Two alternative exon 1	No	No	
CLIC1	1222	mRNA/ESTs	1 or 2	723	Start at exon 1 or 2	ND	No	
MSH5	2887	mRNA/ESTs	2	2505	No	Alternative exons 1 and 7 length	mRNA/ESTs	
C6orf26	1055	ESTs	1	444	Two alternative exon 1	Exons 4 and 5 conjoined	mRNA/ESTs	
C6orf27	4337	mRNA (2)	2	2673	ND	ND	No	
VARS2	4169	mRNA/ESTs	2	3795	No	No	No	GC splice donor at exon 19
LSM2	864	mRNA/ESTs	1	285	Two minor EST variants	No	No	
HSPA1L	2539	mRNA/ESTs	2	1923	No	No	No	
HSPA1A	2337	mRNA/ESTs	1	1923	No	No	No	
HSPA1B	2528	mRNA/ESTs	1	1923	No	No	No	
C6orf48(G8)	1050	mRNA/ESTs	ND	ND	Three alternative exon 1	Variable location for exon 3/exon 3 skipped	No	Possible pseudogene
NEU1	2045	mRNA/ESTs	1	1245	No	No	mRNA	
C6orf29	2577	mRNA/ESTs	1	2127	No	No	mRNA	GC splice acceptor at exon 13
BAT8	3940	mRNA/ESTs	1	3570	Two alternative exon 1	Exon 10 skipped	No	
C6orf46(G10)	1773	mRNA (1)	2	1281	ND	ND	mRNA/EST	Spliced to exon 4 of C2

(continued)

Table 2. *Continued*

Human gene	mRNA length	Evidence for gene model	ATG exon	CDS length	Alternative promoter	Alternative splice forms	Read-through transcripts	Notes
C2	2772	mRNA/ESTs	1	2256	No	Alternatie exon 8 length	mRNA/EST	
BF	2861	mRNA/ESTs	1	2292	No	Alternative exons 2 and 3 length in some ESTs	No	
RDBP	1476	mRNA/ESTs	2	1140	No	No (only one minor EST variant)	No	
SKIV2L	3961	mRNA/ESTs	1	3738	No	No	No	
DOM3Z	1619	mRNA/ESTs	1 or 2	1188	Start at exon 1 or 2	Alternative exon 4 length	No	
STK19	1720	mRNA/ESTs	1	1092	No	Exon 3 skipped	No	
C4A	5459	mRNA/ESTs	1	5232	No	Minor EST variants	No	
CYP21A1P	ND		ND	ND	ND	ND		Pseudogene
C4B	5459	mRNA/ESTs	1	5232	No	Minor EST variants	No	
CYP21A2	2108	mRNA/ESTs	1	1485	No	Alternative exon 2 length; exon 2 skipped	No	
TNXB	13268	GenScan/partial mRNA/ESTs	2	12867	No	Alternative exon 4 length	mRNA	
CREBL1	2655	mRNA/ESTs	1	2100	No	Minor EST variants	mRNA	
FKBPL	1344	mRNA/ESTs	2	1047	No	Extra exon in 1 EST	No	
C6orf31	1917	mRNA/ESTs	2	918	Three alternative exon 1	Highly complex; internal terminations	No	No clear gene model
PPT2	2022	mRNA/ESTs	2	906	Three alternative exon 1	No	mRNA/ESTs	
EGFL8	1262	mRNA/ESTs	2	879	No	Alternative exon 1 and 6 length	mRNA/ESTs	
AGPAT1	1989	mRNA/ESTs	2	849	Four alternative exon 1	Exons 3 and 4 skipped in 1 mRNA	No	
RNF5	1172	mRNA/ESTs	1	540	No	Minor EST variants	No	
AGER	1489	mRNA/ESTs	1	1212	No	1 mRNA has exon 3 length variation and skips exon 8; some ESTs skip exon 11	No	
PBX2	3202	mRNA/ESTs	1	1290	Minor EST variants	No	No	
C6orf9(short)	1477	mRNA/ESTs	1	480	Two alternative exon 1	Four extra exons in 1 mRNA variant	No	
NOTCH4	6742	mRNA/partial ESTs	1	6006	ND	ND	No	GC splice donor at exon 20

Gene models were constructed based on the best supporting evidence, giving highest weight to full-length cDNA sequences (mRNAs). For genes without supporting EST evidence, the number of mRNAs is given in parentheses. The length of the coding sequences was determined from the longest inframe translation. For the cases of pseudogenes (*CYP21A1P*), RNA genes (*C6orf48*), and genes for which there was minimal supporting evidence (e.g., *G6F, LY6G6E, C6orf46, C6orf27, Notch4*), gene models of alternative splice variants cannot be precisely determined (ND). cDNA/EST sequences were identified from the April 2003 assembly found in http://genome.cse.ucsc.edu.

bers and the splice junctions of exons in most genes are well conserved, and the conserved exons in human and mouse are generally of very similar lengths (normally within 9 bp). However, based on present cDNA/EST data, there are 14 genes showing different numbers of exons between human and mouse, caused by 5′ noncoding exons, differences in the splice variants supported by mRNAs, or deletions of nonessential exons (see Supplemental Table 2). We also calculated average intron lengths for genes having orthologs in both organisms: 0.63 kb for human and 0.58 kb for mouse. These values are much shorter than the average human (4.7 kb) and mouse (3.9 kb) introns, based on 1289 pairs of human/mouse orthologous genes (Mouse Genome Sequencing Consortium 2002) This strongly demonstrates that the class III region is very compact. On the other hand, the mouse introns are ~8% shorter than their human counterparts in the MHC class III region, which agrees with the trend found in the genome as a whole (Mouse Genome Sequencing Consortium 2002).

Although the average number of exons per gene and size of the mRNA are fairly typical of the human genome (International Human Genome Sequencing Consortium 2001), the average size of the class III genes in human is 8.5 kb (7.7 kb in mouse; see Fig. 5 and Supplemental Table 2), significantly smaller than the genome average, of at least 27 kb (International Human Genome Sequencing Consortium 2001). Only one gene, *TNXB*, is larger than 30 kb. Likewise, the average intergenic distances, 2.99 kb (human) and 2.80 kb (mouse), are much less than the genome average (see Fig. 5), with only four pairs of genes having an intergenic distance >10 kb in both species. In terms of smallest intergenic distance, nine pairs of adjacent human genes and six pairs of mouse genes have overlapping transcripts (e.g., *CSNK2B*/*BAT4*, *DOM3Z*/*STK19*; see Supplemental Table 2), generally with

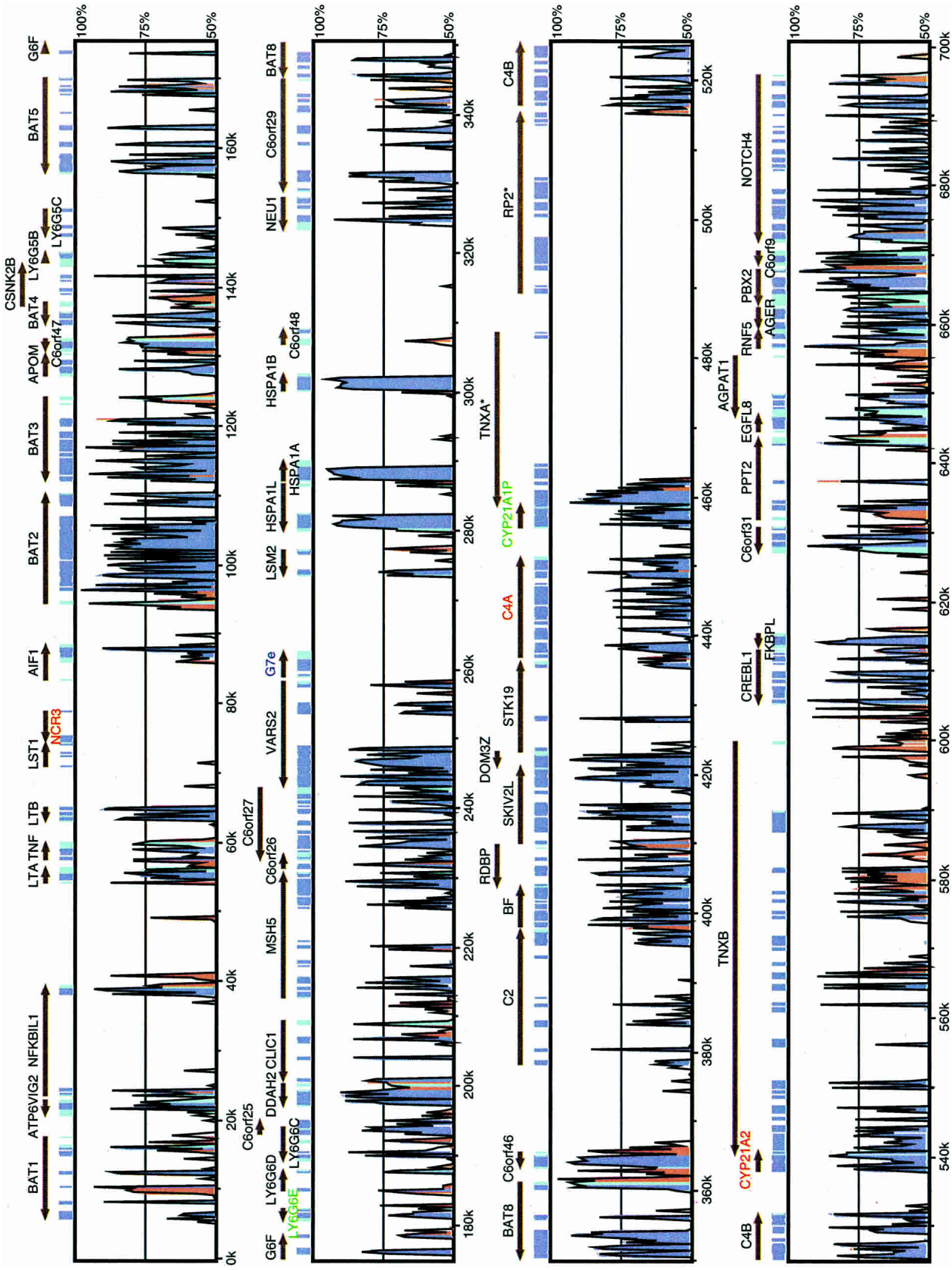

Figure 3 (Legend on next page)

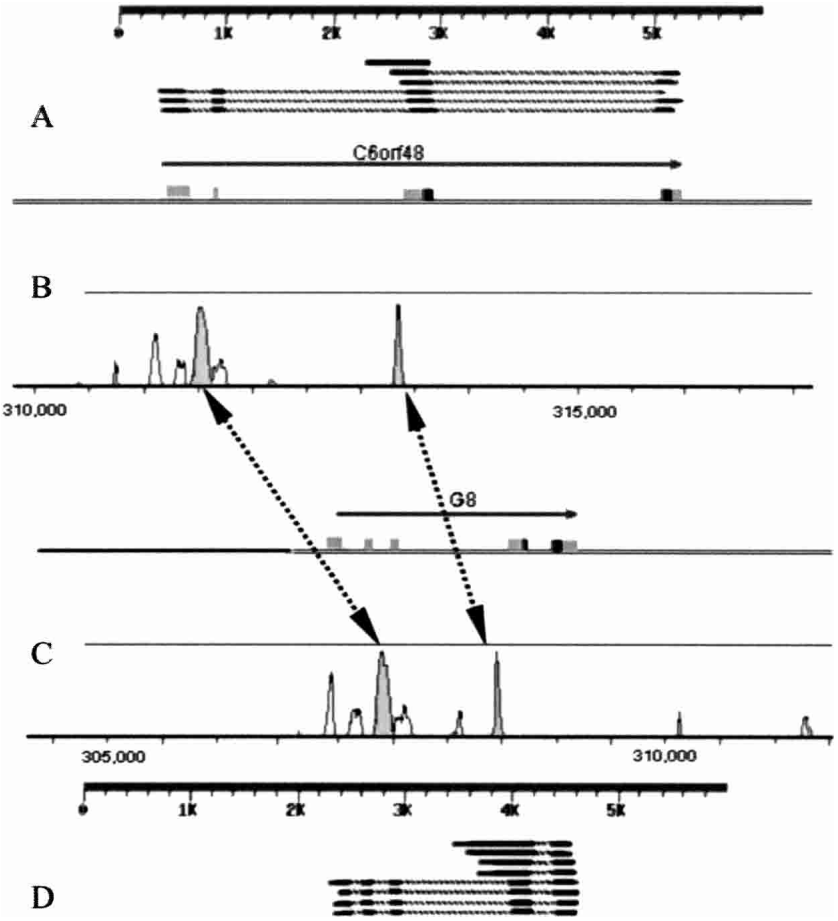

Figure 4 Conserved noncoding regions and nonconserved coding regions. (*A*) The graphic representation of the BLASTN alignment between human EST hits and the genomic DNA region around *C6orf48* is superimposed based on the DNA coordinates. (*B*) VISTA output of the same region showing the intron–exon structure of *C6orf48*. (*C*,*D*) The VISTA and BLASTN outputs for the corresponding gene, *G8*, in mouse. Two conserved noncoding regions (indicated by dotted arrows) are found to encode two snRNAs.

opposite transcriptional orientations, but not always (e.g., *DDAH2*/*CLIC1*). These data stand in marked contrast to Chromosome 14 (Heilig et al. 2003), for example, which has an average gene size of 45.7 kb and an average intergenic distance of 51.2 kb, indicating that only ~44% of the chromosome is transcribed. Here we show that as much as 72% of the MHC class III region is transcribed. These results establish that the MHC class III region is extremely gene-dense, not only in terms of number of genes per megabase, but also with respect to the coding sequence and extent of the region transcribed. The smaller average lengths of mouse 5′- and 3′-UTRs and a lower number of overlapping gene pairs might be partially due to the existence of fewer mouse ESTs. In some cases, however, it is clear that the mouse has deleted DNA relative to human. For example, numerous ESTs and mRNAs indicate that *STK19* has eight exons in human and seven in mouse. The local alignment of the gene indicates that mouse is missing DNA corresponding to human exon 1, the intron, and the beginning of exon 2, giving a significantly longer open reading frame for human (1092 vs. 762 bases) and a different translation for the first several amino acids of the mouse protein.

Alternatively Spliced Genes

Several MHC class III region genes, for example, *LST1*, exhibit alternatively spliced isoforms (de Baey et al. 1997; Neville and Campbell 1997; Rollinger-Holzinger et al. 2000). Through comparisons of EST or cDNA sequences with the genomic sequence, we found potential alternative splicing in at least two-thirds of the human class III genes (see Table 2), many of which have not been previously known to use alternative splicing. Several genes have alternative first exons, indicating different possibilities for gene regulation. Human *APOM* contains one variant whose exon 1 overlaps the 5′-UTR of *BAT3* and another variant whose exon 1 is in the intergenic region closer to exon 2 of *APOM*. Whereas the ESTs supporting the second variant are predominately from liver tissue, the ESTs supporting the first variant are from a wide variety of tissues, indicating perhaps that the opening up of this region of the chromosome to transcribe *BAT3* also allows for unregulated transcription of *APOM*. As another example, the *AGPAT1* gene encodes a lysophosphatidic acid acyltransferase that is also present in bacteria, yeast, and plants. We (Aguado and Campbell 1998) and other groups (Stamps et al. 1997; West et al. 1997) identified this seven-exon MHC class III region gene in human from full-length cDNA clones. From genomic sequence comparison of human and mouse class III sequences, we found several conserved noncoding regions (CNSs) around the first exon of the *AGPAT1* gene in human and in mouse. After analyzing the BLASTN hits for the *AGPAT1* gene in dbEST, we identified ESTs that represent at least five different spliced isoforms in human and three in mouse. Figure 6 shows the three alternatively spliced isoforms of the mouse *Agpat1* gene. More importantly, these new alternatively spliced exons all are located in previously identified CNSs. Thus, CNSs provide a powerful way to identify additional exons in genomic sequences. Some alternatively spliced exons in one species do not have counterparts in the other species based on sequence similarity. These species-specific exons might originate after the divergence of the ancestors of human and mouse. Another possibility is that some alternatively spliced exons might not be detected in a cross-species comparison because they are evolving much faster than constitutively spliced exons, as has been observed on a genome scale (Modrek and Lee 2003).

Figure 3 VISTA plot of the human and mouse MHC class III regions. Conserved sequences (percent identity >50%) are shown in different colors according to the type of their sequence: blue for coding regions, turquoise for UTRs, and red for CNSs. Two names with "*" (*TNXA*, *RP2*) denote gene fragments. Gene names are given for the human orthologs. Three gene names (*NCR3*, *C4A*, *CYP21A2*) are painted in red, to indicate their mouse orthologs are pseudogenes, and the names of two human pseudogenes (*CYP21A1P* and *LY6G6E*) are green. The two regions where the sequences do not align are due to a unique gene in mouse, *G7e* (name in blue), which resembles a viral envelope gene (Snoek et al. 1996) at 260 kb; and to several transposable elements that have inserted into the mouse genome between 470 and 510 kb. The approximate positions on the April 03 Goldenpath assemblies are chr6:31550009–32223670 (human) and chr17:33160937–33875007 (mouse).

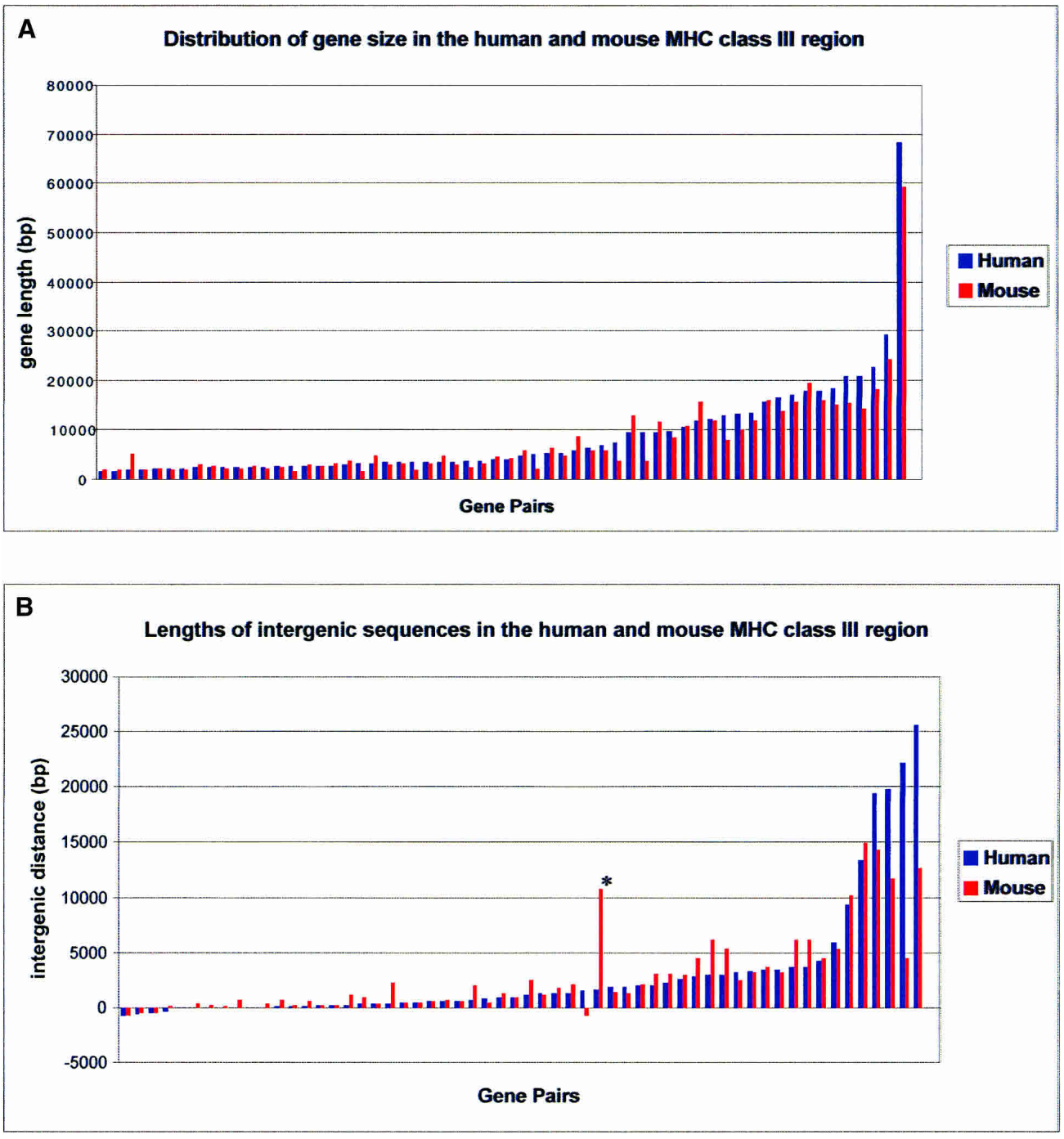

Figure 5 Distribution of gene size (*A*) and intergenic sequences (*B*) for the human and mouse MHC class III gene pairs. The *X*-axis coordinate is the rank order of the human genes. The outlier in *B* (indicated by *) results from a mouse-specific insertion between the *Lsm2* and *Vars2* genes, which also harbors the mouse unique gene, *G7e*. The intergenic distance of mouse *Lsm2* to its closest upstream gene, *G7e*, is 10,829 bp; whereas there are only 1588 bp from the human *LSM2* gene to its closest upstream gene, *VARS2*. Supporting data can be found in Supplemental Table 2.

Other forms of alternative splicing include variations in exon length and in the number of exons found in the transcripts (see Table 2). Whereas some alternative splice variants can alter the protein sequence in ways that increase functional possibilities, others result in apparently nonfunctional proteins. For example, human lymphotoxin β (*LTB*) exists in two forms, each supported by several ESTs. One form is missing exon 2 with the result that the transcript contains a stop codon that truncates the protein's extracellular domain, thereby inhibiting its ability to bind to lymphotoxin α. However, it cannot be ruled out that a truncated protein of this sort serves a different physiological function. In mouse, interestingly, *Ltb* appears not to be alternatively spliced. There are only three exons, yet the amino acid sequence is longer because the mouse DNA orthologous to the intron between human exons 2 and 3 is represented in the *Ltb* mRNA and gives an in-frame translation, whereas the human

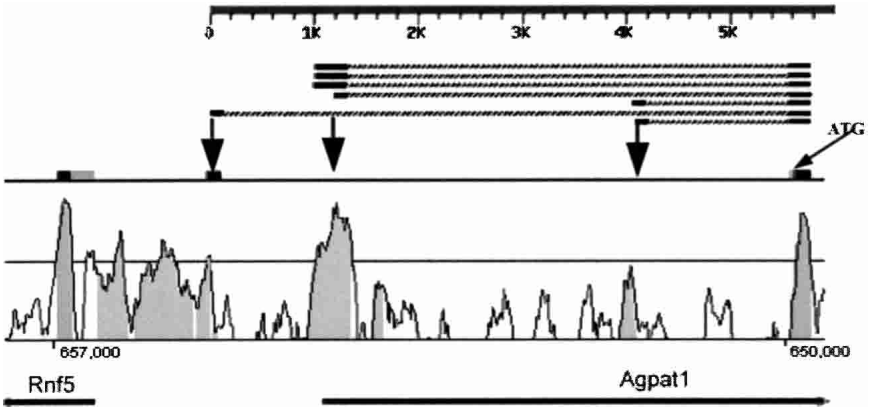

Figure 6 Alternatively spliced exons of the mouse *Agpat1* gene found by genome comparison. The upper part of the figure is the graphic representation of the alignment from the result of BLASTN search, against the mouse EST database. The query sequence is the 6-kb upstream sequence from the first coding exon of the gene *Agpat1*. The graph below is the VISTA output of this region. Three alternatively spliced exons can be clearly identified and mapped to the conserved genomic sequence (see arrows). The coding sequence begins from the eleventh base of the second exon (see the ATG sign).

intron contains stop codons. Mouse DNA corresponding to human exon 2 lacks a functional splice donor (AT rather than GT), whereas the DNA corresponding to exon 3 and its splice sites is well conserved.

Gene/Protein Functions

The MHC has been linked to susceptibility to many diseases, and often these associations cannot be fully explained by variation in the class I and II regions (Gruen and Weissman 2001). The characterization of the class III region of the MHC at the protein level could provide insights into understanding these diseases. The functions of nearly half of the MHC class III region genes are still unknown, although this region has been intensively studied for decades. We performed similarity searches and motif/profile analyses to provide functional insights into the proteins encoded by the genes of unknown function. The results are shown in Table 3. Some of these proteins with uncharacterized function showed no significant similarity to any sequence in the NR database. However, they were found to contain motifs/domains that could indicate a potential function. For example, *G6F* and *C6orf25* contain potential Ig domains, indicating that they could encode potential cell surface receptors involved in the immune system and/or inflammation (de Vet et al. 2001). Similarly, *C6orf46* was found to contain a Broad Complex Tramtrack and Bric Brac (BTB) motif and four Zn fingers, features of a transcription factor, whereas others such as *BAT4* contain several ANK repeats, indicating potential roles in protein–protein interactions within the cell. For some other proteins, such as *C6orf47*, the only information that we could obtain is the presence of two potential transmembrane domains, whereas *C6orf26* does not show any known motifs/domains at all.

In contrast, other gene products of unknown function showed significant similarity to proteins found in other species (e.g., *DOM3Z*), proteins that have a described biochemical activity (e.g., *PPT2*), or members of multigene families (e.g., *C6orf29*). These genes may have specialized, but similar, functions to their counterparts.

Several of the genes within the MHC class III region play roles in the innate immune system, including members of the complement fixation cascade (*C4*, *C2*, *BF*) and the tumor necrosis factor family (*TNF*, *LTB*, *LTA*), which are downstream components of the immune response initiated by Toll receptors. Other genes (e.g., the LY6 family members, *LST1*, *NCR3*, and *AIF1*) are likely to function as part of the immune/inflammatory response as well.

Several of the genes in the class III region appear to act on DNA or RNA: *MSH5* is thought to be involved in chromosomal pairing during meiosis; *LSM2*, in pre-mRNA splicing; *BAT1* is a member of the ATP-dependent RNA helicase family; *RD* may interact with the basal transcriptional apparatus; and *VARS2* is an aminoacyl tRNA synthetase.

The remaining genes include a heterogeneous mixture of metabolic enzymes, transcription factors, protein-modification enzymes, and the like.

In terms of gene expression (Table 3), there is not a consistent pattern. Several genes are highly expressed in a wide variety of tissues (e.g., *BAT1*, *CLIC1*, *DDAH2*). Other genes appear to be restricted in their expression. For example, based on ESTs, the vacuolar ATPase G2 subunit is expressed primarily in the brain. As a general trend, the genes with ancient origin (meaning that they have counterparts in yeast and bacteria) are highly and widely expressed, as one would expect if they perform cellular "housekeeping" functions. Only a few of the mammal-specific genes appear to be highly and widely expressed (e.g., *CREBL1*). Others, for example, *NCR3* and *CYP21A2*, are more restricted, being expressed in natural killer cells and the adrenal gland, respectively.

Evolutionary Conservation

The MHC class III region contains both ancient genes and genes that may have emerged recently (Table 3). For example, *AGPAT1*, lysophosphatidic acid acyl transferase α, an enzyme that participates in phosphatidic acid biosynthesis, appears to be conserved across the whole spectrum of organisms even at the nucleotide level, as judged by BLASTN similarities. On the other hand, *NCR3*, a putative natural killer cell receptor, is likely to play a specialized role in the vertebrate immune system. For this gene and others in the "mammals only" class, the lack of apparent orthologs in pufferfish could be due to the absence of the gene or to an evolution of the gene in the two species that is so rapid that protein similarity searches do not yield significant matches (Aparicio et al. 2002).

In Table 3, we show the percent identity of pairwise protein alignments for 54 pairs of human and mouse orthologs (see "PIP" column). The average is 83.1%, which is higher than the average percent identity of 70.1% found for 12,845 human–mouse orthologs analyzed in the draft mouse genome paper (Mouse Genome Sequencing Consortium 2002). In class III, the evolutionarily most conserved proteins show the highest percent similarity between human and mouse, and the genes found only in mammals show the least amount of conservation. One possible explanation for this observation is that some of the mammal-specific genes, for example, *LY6G6E,* are not functional. (Mallya et al. 2002). Alternatively, these pairs of genes may be novel and rapidly evolving. Nine of these gene pairs, of which seven are believed to have a potential immune-related role (*APOM*, *C6orf47*, *LY6G5B*, *LY6G5C*, *G6F*, *LY6G6E*, *LY6G6D*, *LY6G6C*, and *C6orf25*), lie within a 60-kb region of the class III region, that is, in the region between *BAT3* and *DDAH2* (Fig. 3). Unlike the situation for many of the class III genes, which have paralogs on other human chromosomes, most frequently on 1, 9, and 19

Table 3. Evolutionary Conservation of Genes in the MHC Class III Region

Name (HUGO)	Human ESTs	Mouse ESTs	PIP	FC	Function (total average PIP = 83.1)
					Similarity found only to mouse/rat or other mammals, 16 genes, average PIP = 70.7
NFKBIL1	>100	30	86.4	T	Inhibitor of Rel family transcription factors?
LST1	30	43	48.5	U*	Unknown, Leukocyte-specific transcript 1
NCR3	6	0	—	I	Natural Killer (NK) Receptor
APOM	80	74	80.5	U	Apolipoprotein M. Unknown, Trmb domain, Lipid transport?
C6orf47	10	0	75.5	U	Unknown, 2–3 Trmb domains
L Y6G5B	9	2	58.2	U*	Ly6 family member
LY6G5C	9	0	67.3	U*	Ly6 family member
G6F	1	2	65.4	U*	Immunoglobulin gene superfamily member
LY6G6E	0	13	—	U*	Ly6 family member
LY6G6D	2	9	60.3	U*	Ly6 family member
LY6G6C	3	9	84.1	U*	Ly6 family member
C6orf25	3	9	57.9	U*	Immunoglobulin gene superfamily member
C6orf26	0	4	58.7	U	Unknown
CYP21A2	14	2	—	O	Cytochrome P450 steroid 21 hydroxylase
CREBL1	>100	>100	88.1	T	cAMP response element binding protein motif, BRLZ and BZIP motif
C6orf9	35	39	88.1	U*	Unknown, Proline rich, GoLoco motif (Gα/β motif). G protein signaling?
					Similarity found to *fugu* fish, 4 genes, average PIP = 80.0
LTA	6	5	72.2	I	Cytokine; role in lymphoid organ development and germinal center formation
TNF	12	41	78.8	I	Cytokine; antitumour activity; roles in inflammation of immunomodulation
LTB	38	42	79.3	I	Cytokine; anchors LTA to cell membrane
BAT2	>100	>100	89.7	U	Unknown
					Similarity found to worm and insects 20 genes, average PIP = 85.1
BAT3	>100	>100	92.1	U	Unknown. 2–3 Tmb domains, Ubiquitin motif, CAP motif
BAT4	71	81	78.5	U	Unknown. ANK repeat, G patch motif
BAT5	>100	>100	96.1	U	Unknown. Signal peptide, Trm domain, α-β hydrolase motif
CLIC1	>100	>100	98.3	T	Nuclear chloride ion channel protein? Regulation of cell cycle?
C6orf27	3	0	78.7	U	Unknown. Von Willebrand factor type A domain
C6orf48	>100	66	—	U	encode two snRNA genes
C6orf46	1	33	90.6	U	4 Zn finger, Broad Complex Tramtrack and Bric Brac (BTB) protein–protein interaction motif
C2	>100	>100	75.0	I	Complement classical pathway serine protease
BF	>100	>100	83.6	I	Complement classical pathway serine protease
RDBP	>100	>100	90.5	T	Subunit of NELF (negative elongation factor). Inhibits transcription elongation
STK19	57	62	83.8	O*	Serine threonine kinase 19
C4	>100	>100	76.3	I	Complement classical pathway thioester containing protein
TNXB	75	71	69.2	O	Extracellular matrix protein. Connective-tissue structure/function? Development blood vessels?
FKBPL	47	52	73.4	U	Unknown. FK506-binding protein like. Immunophilin like, 3 TRR repeats
C6orf31	9	32	97.4	U	Unknown. Signal peptide, 2-3 Trmb domain, Proline rich
EGFL8	37	13	79.9	U*	Unknown. Signal peptide, Trmb domain, 2EGF motifs
RNF5	>100	93	97.8	U	Ubiquitin Ligase E3, Ring finger. Trmb domain
AGER	36	13	78.2	T*	Receptor for advanced glycosylation end products of proteins
PBX2	71	69	97.9	T	Homeobox domain; transcriptional regulation
NOTCH4	11	89	80.2	T*	Cell differentiation; Cell proliferation? Regulation of cell fate determination? Morphogenesis?
					Similarity found to yeast, fungi 11 genes, average PIP = 89.2
BAT1	>100	>100	99.3	T	56kD U2AF56 associated protein UAP56. Essential splicing factor. DEAD-box domain
ATP6VIG2	11	89	94.9	O	Vacuolar ATPase G synthetase subunit
AIF1	>100	43	88.4	U*	Allograft inflammatory factor; macrophage activation?, EF hand, Ca^{2+} binding motif
CSNK2B	>100	>100	100	T	Casein kinase II β subunit; cell growth?
DDAH2	>100	>100	97.2	O	NG-dimethylarginine dimethylamino hydrolase II
LSM2	85	92	72.5	T	Like Sm protein 2. Binds specifically to the 3′-terminal U-track of U6 snRNA
NEU1	>100	>100	82.0	O*	Sialidase enzyme
C6orf29	>100	>100	82.0	U	hCTL4, Choline transporter-like 4
BAT8	>100	>100	81.2	U	Histone Methyltransferase (HMTase) ANK repeats
DOM3Z	96	>100	89.7	U	Unknown. Proliferation and viability in *C. eleg.* (similar to DOM3Z). 5.8S rRNA processing in yeast
PPT2	94	65	93.7	O*	Thioesterase activity
					Similarity found to bacteria, 7 genes, average PIP = 92.6
MSH5	73	22	88.7	T	MutS homolog 5. Chromosome pairing in meiosis. Heterooligomer with MSH4
VARS2	>100	>100	91.7	O	valyl tRNA synthetase
HSPA1L	>100	>100	94.9	O*	heat-shock protein, constitutive HSP70
HSPA1A	>100	>100	95.2	O*	heat-shock protein, chaperone in recovery of cells from stress

(continued)

Table 3. *Continued*

Name (HUGO)	Human ESTs	Mouse ESTs	PIP	FC	Function (total average PIP = 83.1)
HSPA1B	>100	>100	95.0	O*	heat-shock protein, chaperone in recovery of cells from stress
SKIV2L	>100	>100	92.9	T*	RNA helicase
AGPAT1	>100	56	89.9	O*	Lysophosphatidic acid acyltransferase

The number in columns "human" and "mouse" show how many ESTs (dbEST_human and dbEST_mouse, version 5/18/2003 from NCBI) were found by BLASTN with scores >200 and alignments >100 bp. In "PIP" (percent identity of protein pairwise alignment) column, those numbers come from alignments of the same splicing forms between human and mouse: (—) pseudogene, so no PIP value is provided. Also, no PIP value was given for *C6orf48,* because it is an snRNA gene. The FC column shows four function categories: (I) known immune-related and inflammatory genes; (T) transcription/regulation/signaling related genes; (O) other known functional genes; and (U) genes of unknown function. For motifs: (UBQ) ubiquitin homolog; (ANK) ankyrin repeats; (G-patch) glycine-rich nucleic acid-binding domain; (VWF) Von Willebrand factor A domains; (BTB) Broad-Complex, Tramtrack and Bric a brac (BTB) protein–protein interaction motif, also known as POZ (poxvirus and Zinc finger) domain; (TPR) tetratricopeptide repeat; (BRLZ) basic region leucine zipper; (bZIP) basic leucine zipper.
*Genes of unknown function that are thought to have a potential role in the immune/inflammatory response.

(Abi-Rached et al. 2002), most genes in this block appear to be unique. To decipher the physiological roles in vivo of these genes needs further experimental study.

Conserved Noncoding Sequences (CNS) and Gene Regulation

Conserved noncoding regions near exons may identify regulatory elements in mammals (Koop and Hood 1994). Algorithms for identifying regulatory elements are not very effective; therefore, orthologous cross-species comparisons of noncoding regions is a strategy used to identify potential regulatory elements (Oeltjen et al. 1997; Loots et al. 2000; Touchman et al. 2001). One such element, in a noncoding region of the mouse and human T-cell receptor Jα region, was demonstrated to have enhancer activities (Kuo et al. 1998). In the present study, we scanned the CNSs found by GLASS for known and unknown regulatory elements. When using an arbitrary cutoff, >65% identity and extending >120 bp, we identified 149 CNSs, encompassing 36,187 bp in the human (5.1%) and 35,211 bp in the mouse (5.0%) class III region, respectively. About two-thirds of the CNSs (97 out of 149) are located close to the first exon of the genes. Another 22 CNSs are found in middle introns, whereas the remaining CNSs are located in intergenic regions (Fig. 3). Some CNSs exhibit >80% identity. For example, the CNS in the upstream region of the *RNF5* gene shows an ungapped 547-bp alignment with an identity of 80.6%. The longest CNS found by GLASS, 1219 bp long in mouse and 1221 bp long in human, exhibits an identity of 75.20% and lies in the first intron of the *DDAH2* gene. Some CNSs are close to each other (<200 bp), indicating that a set of discrete transcription-factor-binding elements may combine to form a regulatory module spanning hundreds of base pairs.

We also examined several known *cis*-regulatory elements in several well-studied genes. Among them, the human *CYP21A2* gene has a distal transcriptional regulatory element "Z promoter," which lies in intron 35 of the upstream gene *C4B* (Wijesuriya et al. 1999). From the genomic comparisons, we found a conserved noncoding DNA segment exactly at the correct position at intron 35 in the human *C4B* gene as well as in the mouse *C4B* gene (Fig. 7). This CNS is 162 bp long with a 67.7% identity. This indicates that the "Z promoter" is really conserved between human and mouse. This example demonstrates the power of using cross-species comparisons to screen for unknown but conserved *cis*-elements.

We also used PromoterInspector (see Methods) to detect promoters in this region. In total, it found 22 promoters in human and 17 in mouse. Among them, 12 promoters in human and 12 in mouse are associated with the same CNSs; five other promoters in human and two in mouse seem not to overlap with any CNSs. Interestingly, another five promoters in human and three in mouse overlap with some CNSs, but no promoter was found overlapping with those CNSs in the other species. Moreover, the Z promoter could not be determined by PromoterInspector. These analyses imply that CNS identification using cross-species comparisons is a powerful computational approach for prescreening candidates for regulatory elements in rapidly evolving regions of the genome. However, it is difficult to use this approach on slowly evolving regions such as many AT-rich, gene-poor regions because there are an abundance of CNSs, most of which are probably not functional.

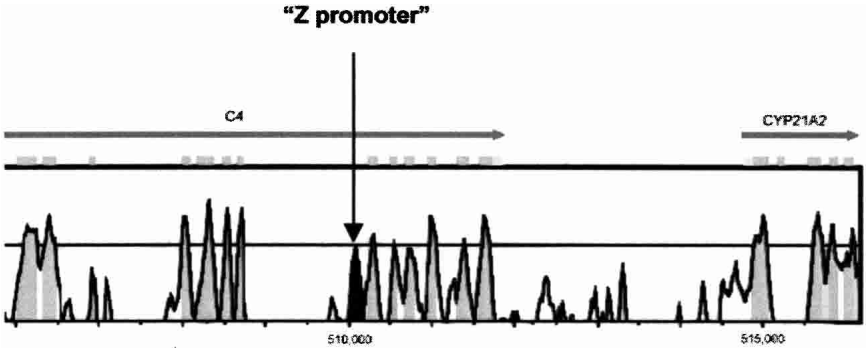

Figure 7 Distal *cis*-element found by CNS analysis. The Z promoter (Wijesuriya et al. 1999) of the human *CYP21A2P* gene is located in the 35th intron of its upstream gene, *C4B*. This promoter sits exactly in a conserved noncoding sequence, represented by a black peak.

Polymorphisms in the MHC Class III Region

Polymorphic markers in the class III region may facilitate the identification of gene loci that are involved in susceptibility to numerous diseases. Unlike the highly polymorphic MHC class I and class II regions, the MHC class III region shows sequence variations that are typical of the genome as a whole. To assess the type and frequency of SNPs and microsatellite variations, we compared our human class III consensus sequence with the overlapping DNA segment sequenced by Shiina et al. (1999) (GenBank accession nos. AP000502–AP000506, where the overlap is ~420 kb long, from the genes

BAT1 to *C2*). The two 420-kb genomic sequences were aligned using the AVID program. There are 38 genes in this region, covering 252.5 kb, including 319 exons (51.8 kb), 76 UTRs (14.0 kb), and 357 introns (186.7 kb). Using the Sputnik program, 32 di-, 41 tri-, 64 tetra-, and 41 pentanucleotide repeats were identified in the 420-kb region. Among them, 16 di-, 3 tri-, 5 tetra-, and 4 pentanucleotide repeats vary between the two sequences, and therefore could be potential polymorphic markers. Of these microsatellites, 12 were localized in intergenic regions, whereas others were in introns. Matsuzaka et al. (2001) found that 17 microsatellites (all are trinucleotide repeats) were located within the coding region of expressed genes in the human MHC class III region. Here we did not find any polymorphic microsatellites in coding exons or UTRs. We also found >120 short indels from the alignment.

The density of SNPs in the human genome is estimated to be 1 to 10 per 1000 nucleotides, when comparing two chromosomes (Kruglyak 1997), and a more recent estimate is ~1 SNP/1.9 kb, reported by the International SNP Map Working Group (Sachidanandam et al. 2001). Geraghty et al. (1999) identified >10,000 SNPs in the 2.2-Mb DNA segment that includes all of the class I region. Ribas et al. (2001) detected a density of one SNP per 489 bp in an 18-kb DNA segment in the human class III region. Here, by screening the alignment of the two genomic sequences, we found a total of 371 SNPs therein, representing 0.9 SNPs/kb. Of these 371 SNPs, 217 (58.5%) were found in genes, including 23 in coding regions (cSNP), 11 in UTRs, and 183 in introns. Of the 23 cSNPs, 13 (56.5%) were nonsynonymous mutations, whereas another 10 (43.5%) SNPs were synonymous mutations. These SNPs are distributed relatively evenly in the genomic sequence. There are 254 transitions (68%) and 117 (32%) transversions. The frequencies of the substitutions were A/C (7.3%), A/G (28.3%), A/T (6.7%), C/G (9.2%), C/T (40.2%), and G/T (8.4%).

Recently, the Sanger Institute released two sequences from their MHC haplotype project (see http://www.sanger.ac.uk/HGP/Chr6/MHC/). The first consensus sequence, from the PGF cell line, was assembled using AL663061, AL662801, AL662899, AL671762, AL645922, AL772248, AL662884, and AL772153 (706,338 bp); the second sequence, from the COX cell line, included AL662847, AL670886, AL662834, AL662849, AL662828, and AL662830 (667,547 bp). We compared those two consensus sequences to our sequences, and found 787 and 582 SNPs, respectively. The second consensus sequence contains only one copy of the C4-CYP21 module, which might explain why the number of SNPs found between it and our sequence is smaller than that of the first one. To link these SNPs to disease susceptibility requires further work. However, it is interesting to note that the overall frequency of nonsynonymous versus synonymous mutations in the cSNPs and the frequency of transitions versus transversions are very similar between both studies. These SNPs can be accessed at http://db.systemsbiology.net/projects/local/mhc/SNP/.

In mouse, the number of variations in the class III region between the 129SJ strain (this study) and the C57B6 strain used for the draft sequence appears remarkably limited. To obtain an estimate of the variation, we aligned a finished BAC sequence from mouse strain C57BL6/J presented in the UCSC mouse genome browser (AC087117), 222 kb long, to our consensus sequence, and found only 12 SNPs and six indels (in microsatellite regions) between the two strains. It is likely that the two strains share identity by descent in this region (Wade et al. 2002).

Conclusion

The sequence comparison of the ~700-kb human and mouse MHC class III regions showed them to be the most gene-dense region of the human and presumably mouse genomes, 60 and 61 genes, respectively. Although about half of the genes have unknown functions, many of the genes with putative functions encode immune-related activities. The comparative analyses of the human and mouse class III regions identified conserved sequence blocks that provided insights into gene structure, alternative RNA splicing exons, putative regulatory regions, and even the presence of previously unrecognized genes. The class III genes could be divided into evolutionary groups exhibiting homologies that extend to mammals, fish, worms, fruit flies, yeast, and bacteria—and these groups exhibited conservation inversely related to their divergence times. The MHC encodes a predisposition to a wide variety of human immune-related diseases. In principle, any of the MHC class III genes could contain sequence variations that predispose to disease either alone or in combination with disadvantageous alleles in other genes. Fortunately, because of its proximity to MHC class I and class II, this region of the genome is under intensive investigation for SNP detection and haplotype block analysis (Walsh et al. 2003). The delineation of the human and mouse sequences provided by our laboratories and others will assist correlations between specific polymorphisms and sequence information content such as alternatively spliced exons or probable regulatory elements, with the result that both gene function and disease associations will eventually be revealed and understood better.

METHODS

Mapping and Sequencing

Cosmids sequenced for the human MHC class III region (U89335, U89336, U89337, AF019413) were a gift from Thomas Spies (Spies et al. 1989). Human and mouse MHC class III BACs were identified from the human BAC library RPCI 11 and the Genome Systems mouse stain 129SJ BAC library by using conventional hybridization screening techniques with probes prepared from known gene sequences. Candidate clones were tested for internal consistency by restriction digest fingerprinting. Cosmid or BAC clones were sequenced using the high-redundancy shotgun sequencing approach (Rowen et al. 1999). Source clone DNA was prepared in the AutoGen740 and sheared using sonication. After end repair and size selection, insert DNA was subcloned into either M13mp9 or pUC18. Sequences were resolved on Applied Biosystems 373 or 377 sequencers, using a mixture of dye primer and dye terminator chemistries. After obtaining enough reads for approximately eightfold coverage, the sequence data were assembled in Phrap. Finishing was done using either resequencing with an alternative chemistry, directed sequencing with custom oligonucleotide primers, or by subcloning PCR products or restriction digest fragments. The sequence was determined to an accuracy of about one error per 35,000 bp.

The nucleotide sequences of human and mouse cosmid or BAC clones have been submitted to GenBank as a series of separate entries. The accession numbers are as follows: AC007080, AF109719, AF109905, AF109906, AF049850, and AF030001 (mouse); and AF129756, AF134726, AF019413, U89337, U89336, and U89335 (Li et al. 1998; human). When assembling the human sequence for this report, we extracted L26261, M59816, S80811, M59815, M12793, and S38953 from GenBank to rebuild the C4-CYP21 module, also called the RCCX module (*RP-C4-CYP21-TNX*) in Yang et al. (1999). This addition creates a discrepancy between the sequence found in the Goldenpath assembly, which contains only one copy of C4 and CYP21, and that presented in Figures 2 and 3 and Supplemental Table 2. AC004181 and AC006046, which were sequenced by the Geraghty group (The MHC Sequencing Consortium 1999), were also used to reconstruct the human consensus sequence. The sequence we have defined as the human class III region is 706,395 bp long, oriented from telomere to centromere, and the mouse class III sequence is 700,393 bp long in the same orienta-

tion. Discrepancies between the mouse sequence presented here and that found in the Goldenpath assembly are due to strain variation and mistakes with the assembly of the mouse working draft. The reported human and mouse sequences begin just telomeric of *BAT1* and end just centromeric of *NOTCH4*.

Computational Sequence Analysis

Dot-matrix comparisons were performed using the default settings of the DOTTER program running on a SUN workstation (Sonnhammer and Durbin 1995). The GC isochores and CpG islands were identified using the GCG sequence analysis package (Wisconsin package version 10.1, Genetics Computer Group) and GESTALT (Glusman and Lancet 2000), respectively. RepeatMasker (A. Smit and P. Green, unpubl.; local version is 19/06/2001) was used to identify the repeats in the human and mouse sequences. Genomic sequences were aligned using the global sequence alignment tool GLASS (Batzoglou et al. 2000). VISTA (Mayor et al. 2000) was used to generate a static graph of the percentage identity calculated from the human/mouse sequence alignment. Conserved regions (identity of 70% or greater and length of 100 bp or greater) were plotted with different colors (see legend to Fig. 3), coordinated by the reference sequence (the mouse MHC class III region sequence). Poorly matched regions appear blank in the graph.

GenScan (Burge and Karlin 1997) was used to predict the coding regions from the unannotated sequences. We also carried out BLASTN searches against the dbEST and NR databases at NCBI to identify potential expressed regions. Query sequences were successive fragments (each 5 kb plus 1 kb of overlap) of the human or mouse genomic sequences. Only hits with sufficient alignment (total aligned sequence >100 bp) and high identity (>95%) were considered in gene identification. Intron–exon boundaries were predicted by close examination of the sequences for splice junctions (consensus GT–AG). The est2genome program (Mott 1997) was used to align the cDNA or EST sequences to genomic sequence when the splice donor/acceptor sites were not clear by BLASTN alignments. Pairwise alignments of protein and DNA sequences were obtained using the ALIGN program from the FASTA package (Pearson and Lipman 1988). The FASTY tool in the same package was also used to detect sequencing errors in the ESTs. In polymorphism analysis, the AVID program (http://bio.math.berkeley.edu/avid/) was used to generate pairwise alignments and the Sputnik program (from http://rast.abajian.com/sputnik/) for searching DNA microsatellite repeats.

As a supplementary approach to gene identification and discovery of alternative splice variants, the ESTs/mRNAs supporting each gene were visualized in the UCSC genome browser (April 2003 assembly for human and February 2003 assembly for mouse). ACGT (A Comparative Genomics Tool), a Java 1.4 based program developed locally, can retrieve all the annotations of the two sequences and provides dynamical views of the alignment with a user-friendly interface (Xie and Hood 2003). This program can be freely downloaded at http://db.systemsbiology.net/projects/mhc/acgt/.

When searching the motif databases, the Profile Scan server (http://www.isrec.isb-sib.ch/software/PFSCAN_form.html) was used. The HUGO symbols of annotated genes are gathered from the Human Gene Nomenclature Database (http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl). For protein analysis, the protein identification program (PIX; http://www.hgmp.mrc.ac.uk/Registered/Webapp/pix/) was used, together with BLASTP (unfiltered option) and SMART (http://smart.embl-heidelberg.de). PromoterInspector (Scherf et al. 2000) was used to predict possible promoters in genomic sequences. When analyzing evolutionary conservation of class III genes, BLASTP was used in searching their protein products in the well-annotated SWISS-PROT database. Species information of hits with *e*-values less than 0.001 and alignments covering at least half of the queries is extracted from SWISS-PROT and then classified into five clusters in Table 3. To identify orthologous genes in pufferfish, the human genes were translated and searched against the translated May 2002 assembly of the *Fugu* genome (Aparicio et al. 2002). Synteny with neighboring genes was used to support conclusions about orthology. References describing protein function or protein features can be obtained using standard databases (such as NCBI's PubMed).

ACKNOWLEDGMENTS

We thank Anuradha Madan, Stephen Lasky, Carol Loretz, Dale Baskin, Janet Faust, Rose James, and Christian Dankers for their assistance with mapping and sequencing; Nat Goodman for generating the gene count per megabase data; Jared Roach, Brian Birditt, and Gustavo Glusman for helpful discussions; and the Sanger Institute for releasing their MHC haplotype data. We also thank the anonymous reviewers for their valuable comments and suggestions. B.A. and R.D.C. are funded by the UK Medical Research Council. This project was funded by DOE, NIH, and Bill Gates.

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.

REFERENCES

Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P., and Inoko, H. 2002. Evidence of en bloc duplication in vertebrate genomes. *Nat. Genet.* **31:** 100–105.

Aguado, B. and Campbell, R.D. 1998. Characterization of a human lysophosphatidic acid acyltransferase that is encoded by a gene located in the class III region of the human major histocompatibility complex. *J. Biol. Chem.* **273:** 4096–4105.

Albig, W. and Doenecke, D. 1997. The human histone gene cluster at the D6S105 locus. *Hum. Genet.* **101:** 284–294.

Ansari-Lari, M.A., Muzny, D.M., Lu, J., Lu, F., Lilley, C.E., Spanos, S., Malley, T., and Gibbs, R.A. 1996. A gene-rich cluster between the CD4 and triosephosphate isomerase genes at human chromosome 12p13. *Genome Res.* **6:** 314–326.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., et al. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297:** 1301–1310.

Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B., and Lander, E.S. 2000. Human and mouse gene structure: Comparative analysis and application to exon prediction. *Genome Res.* **10:** 950–958.

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

Burch, G.H., Gong, Y., Liu, W., Dettman, R.W., Curry, C.J., Smith, L., Miller, W.L., and Bristow, J. 1997. Tenascin-X deficiency is associated with Ehlers-Danlos syndrome. *Nat. Genet.* **17:** 104–108.

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

Chen, C.N., Su, Y., Baybayan, P., Siruno, A., Nagaraja, R., Mazzarella, R., Schlessinger, D., and Chen, E. 1996. Ordered shotgun sequencing of a 135 kb Xq25 YAC containing ANT2 and four possible genes, including three confirmed by EST matches. *Nucleic Acids Res.* **24:** 4034–4041.

Chiou, S.H., Hu, M.C., and Chung, B.C. 1990. A missense mutation at $Ile^{172} \rightarrow Asn$ or $Arg^{356} \rightarrow Trp$ causes steroid 21-hydroxylase deficiency. *J. Biol. Chem.* **265:** 3549–3552.

Daniels, R.J., Peden, J.F., Lloyd, C., Horsley, S.W., Clark, K., Tufarelli, C., Kearney, L., Buckle, V.J., Doggett, N.A., Flint, J., et al. 2001. Sequence, structure and pathology of the fully annotated terminal 2 Mb of the short arm of human chromosome 16. *Hum. Mol. Genet.* **10:** 339–352.

de Baey, A., Fellerhoff, B., Maier, S., Martinozzi, S., Weidle, U., and Weiss, E.H. 1997. Complex expression pattern of the TNF region gene LST1 through differential regulation, initiation, and alternative splicing. *Genomics* **45:** 591–600.

de Vet, E.C., Aguado, B., and Campbell, R.D. 2001. G6b, a novel immunoglobulin superfamily member encoded in the human major histocompatibility complex, interacts with SHP-1 and SHP-2. *J. Biol. Chem.* **276:** 42070–42076.

Eddy, S.R. 1999. Noncoding RNA genes. *Curr. Opin. Genet. Dev.* **9:** 695–699.

Geraghty, D.E., Vu, Q., Williams, L., Janer, M., Gassner, C., Russell, C., Ishitani, A., and Jasoni, C. 1999. Mapping HLA for single nucleotide polymorphisms. *Rev. Immunogenet.* **1:** 231–238.

Glusman, G. and Lancet, D. 2000. GESTALT: A workbench for automatic integration and visualization of large-scale genomic sequence analyses. *Bioinformatics* **16:** 482–483.
Gruen, J.R. and Weissman, S.M. 2001. Human MHC class III and IV genes and disease associations. *Front. Biosci.* **6:** D960–D972.
Heilig, R., Eckenberg, R., Petit, J.L., Fonknechten, N., Da Silva, C., Cattolico, L., Levy, M., Barbe, V., de Berardinis, V., Ureta-Vidal, A., et al. 2003. The DNA sequence and analysis of human chromosome 14. *Nature* **421:** 601–607.
International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Kiss-Laszlo, Z., Henry, Y., Bachellerie, J.P., Caizergues-Ferrer, M., and Kiss, T. 1996. Site-specific ribose methylation of preribosomal RNA: A novel function for small nucleolar RNAs. *Cell* **85:** 1077–1088.
Koop, B.F. and Hood, L. 1994. Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nat. Genet.* **7:** 48–53.
Kruglyak, L. 1997. The use of a genetic map of biallelic markers in linkage studies. *Nat. Genet.* **17:** 21–24.
Kumanovics, 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.
Kuo, C.L., Chen, M.L., Wang, K., Chou, C.K., Vernooij, B., Seto, D., Koop, B.F., and Hood, L. 1998. A conserved sequence block in murine and human T cell receptor (TCR) Jα region is a composite element that enhances TCR α enhancer activity and binds multiple nuclear factors. *Proc. Natl. Acad. Sci.* **95:** 3839–3844.
Li, L., Huang, G.M., Banta, A.B., Deng, Y., Smith, T., Dong, P., Friedman, C., Chen, L., Trask, B.J., Spies, T., et al. 1998. Cloning, characterization, and the complete 56.8-kilobase DNA sequence of the human NOTCH4 gene. *Genomics* **51:** 45–58.
Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M., and Frazer, K.A. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288:** 136–140.
Mallya, M., Campbell, R.D., and Aguado, B. 2002. Transcriptional analysis of a novel cluster of LY-6 family members in the human and mouse major histocompatibility complex: Five genes with many splice forms. *Genomics* **80:** 113–123.
Matsuzaka, Y., Makino, S., Nakajima, K., Tomizawa, M., Oka, A., Bahram, S., Kulski, J.K., Tamiya, G., and Inoko, H. 2001. New polymorphic microsatellite markers in the human MHC class III region. *Tissue Antigens* **57:** 397–404.
Mattaj, I.W. 1993. RNA recognition: A family matter? *Cell* **73:** 837–840.
Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A., Pachter, L.S., and Dubchak, I. 2000. VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16:** 1046–1047.
The MHC Sequencing Consortium. 1999. Complete sequence and gene map of a human major histocompatibility complex. *Nature* **401:** 921–923.
Miyagawa, T., Hohjoh, H., Honda, Y., Juji, T., and Tokunaga, K. 2000. Identification of a telomeric boundary of the HLA region with potential for predisposition to human narcolepsy. *Immunogenetics* **52:** 12–18.
Modrek, B. and Lee, C.J. 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.* **34:** 177–180.
Mott, R. 1997. EST_GENOME: A program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.* **13:** 477–478.
Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Neville, M.J. and Campbell, R.D. 1997. Alternative splicing of the LST-1 gene located in the Major Histocompatibility Complex on human chromosome 6. *DNA Seq.* **8:** 155–160.
Nishimura, M., Obayashi, H., Mizuta, I., Hara, H., Adachi, T., Ohta, M., Tegoshi, H., Fukui, M., Hasegawa, G., Shigeta, H., et al. 2003. TNF, TNF receptor type 1, and allograft inflammatory factor-1 gene polymorphisms in Japanese patients with type 1 diabetes. *Hum. Immunol.* **64:** 302–309.
Oeltjen, J.C., Malley, T.M., Muzny, D.M., Miller, W., Gibbs, R.A., and Belmont, J.W. 1997. Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res.* **7:** 315–329.
Okamoto, K., Makino, S., Yoshikawa, Y., Takaki, A., Nagatsuka, Y., Ota, M., Tamiya, G., Kimura, A., Bahram, S., and Inoko, H. 2003. Identification of I κ BL as the second major histocompatibility complex-linked susceptibility locus for rheumatoid arthritis. *Am. J. Hum. Genet.* **72:** 303–312.
Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420:** 563–573.
Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., Sato, H., Hori, M., Nakamura, Y., and Tanaka, T. 2002. Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.* **32:** 650–654.
Pearson, W.R. and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **85:** 2444–2448.
Ribas, G., Neville, M.J., and Campbell, R.D. 2001. Single-nucleotide polymorphism detection by denaturing high-performance liquid chromatography and direct sequencing in genes in the MHC class III region encoding novel cell surface molecules. *Immunogenetics* **53:** 369–381.
Rogers, M.A., Langbein, L., Praetzel, S., Moll, I., Krieg, T., Winter, H., and Schweizer, J. 1997. Sequences and differential expression of three novel human type-II hair keratins. *Differentiation* **61:** 187–194.
Rollinger-Holzinger, I., Eibl, B., Pauly, M., Griesser, U., Hentges, F., Auer, B., Pall, G., Schratzberger, P., Niederwieser, D., Weiss, E.H., et al. 2000. LST1: A gene with extensive alternative splicing and immunomodulatory function. *J. Immunol.* **164:** 3169–3176.
Rowen, L., Lasky, S., and Hood, L. 1999. Deciphering genomes through automated large-scale sequencing. *Methods Microbiol.* **28:** 155–192.
Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928–933.
Saito, H., Fujiwara, T., Shin, S., Okui, K., and Nakamura, Y. 1996. Cloning and mapping of a human novel cDNA (NHP2L1) that encodes a protein highly homologous to yeast nuclear protein NHP2. *Cytogenet. Cell Genet.* **72:** 191–193.
Scherer, S.W., Cheung, J., MacDonald, J.R., Osborne, L.R., Nakabayashi, K., Herbrick, J.A., Carson, A.R., Parker-Katiraee, L., Skaug, J., Khaja, R., et al. 2003. Human chromosome 7: DNA sequence and biology. *Science* **300:** 767–772.
Scherf, M., Klingenhoff, A., and Werner, T. 2000. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: A novel context analysis approach. *J. Mol. Biol.* **297:** 599–606.
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.
Shiina, T., Tamiya, G., Oka, A., Takishima, N., Inoko, H. 1999. Genome sequencing analysis of the 1.8 Mb entire human MHC class I region. *Immunol. Rev.* **167:** 193–199.
Snoek, M., van Dinten, L., and van Vugt, H. 1996. A novel gene, G7e, resembling a viral envelope gene, is located at the recombinational hot spot in the class III region of the mouse MHC. *Genomics* **38:** 5–12.
Sonnhammer, E.L. and Durbin, R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167:** GC1–GC10.
Spies, T., Bresnahan, M., and Strominger, J.L. 1989. Human major histocompatibility complex contains a minimum of 19 genes between the complement cluster and HLA-B. *Proc. Natl. Acad. Sci.* **86:** 8955–8958.
Stammers, M., Rowen, L., Rhodes, D., Trowsdale, J., and Beck, S. 2000. BTL-II: A polymorphic locus with homology to the butyrophilin gene family, located at the border of the major histocompatibility complex class II and class III regions in human and mouse. *Immunogenetics* **51:** 373–382.
Stamps, A.C., Elmore, M.A., Hill, M.E., Kelly, K., Makda, A.A., and Finnen, M.J. 1997. A human cDNA sequence with homology to non-mammalian lysophosphatidic acid acyltransferases. *Biochem. J.* **326:** 455–461.
Strausberg, R.L., Feingold, E.A., Grouse, L.H., Derge, J.G., Klausner, R.D., Collins, F.S., Wagner, L., Shenmen, C.M., Schuler, G.D., Altschul, S.F., et al. 2002. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl. Acad. Sci.* **99:** 16899–16903.
Touchman, J.W., Dehejia, A., Chiba-Falek, O., Cabin, D.E., Schwartz, J.R., Orrison, B.M., Polymeropoulos, M.H., and Nussbaum, R.L. 2001. Human and mouse α-synuclein genes: Comparative genomic sequence analysis and identification of a novel gene regulatory element. *Genome Res.* **11:** 78–86.
Wade, C.M., Kulbokas III, E.J., Kirby, A.W., Zody, M.C., Mullikin, J.C., Lander, E.S., Lindblad-Toh, K., and Daly, M.J. 2002. The mosaic structure of variation in the laboratory mouse genome. *Nature* **420:** 574–578.

Walsh, E.C., Mather, K.A., Schaffner, S.F., Farwell, L., Daly, M.J., Patterson, N., Cullen, M., Carrington, M., Bugawan, T.L., Erlich, H., et al. 2003. An integrated haplotype map of the human major histocompatibility complex. *Am. J. Hum. Genet.* **73:** 580–590.
Wende, H., Volz, A., and Ziegler, A. 2000. Extensive gene duplications and a large inversion characterize the human leukocyte receptor cluster. *Immunogenetics* **51:** 703–713.
West, J., Tompkins, C.K., Balantac, N., Nudelman, E., Meengs, B., White, T., Bursten, S., Coleman, J., Kumar, A., Singer, J.W., et al. 1997. Cloning and expression of two human lysophosphatidic acid acyltransferase cDNAs that enhance cytokine-induced signaling responses in cells. *DNA Cell Biol.* **16:** 691–701.
Wijesuriya, S.D., Zhang, G., Dardis, A., and Miller, W.L. 1999. Transcriptional regulatory elements of the human gene for cytochrome P450c21 (steroid 21-hydroxylase) lie within intron 35 of the linked C4B gene. *J. Biol. Chem.* **274:** 38097–38106.
Xie, T. and Hood, L. 2003. ACGT—A comparative genomics tool. *Bioinformatics* **19:** 1039–1040.
Yang, Z., Mendoza, A.R., Welch, T.R., Zipf, W.B., and Yu, C.Y. 1999. Modular variations of the human major histocompatibility complex class III genes for serine/threonine kinase RP, complement component C4, steroid 21-hydroxylase CYP21, and tenascin TNX (the RCCX module). A mechanism for gene deletions and disease associations. *J. Biol. Chem.* **274:** 12147–12156.

WEB SITE REFERENCES

http://bio.math.berkeley.edu/avid/; AVID program.
http://db.systemsbiology.net/projects/local/mhc/SNP/; comparison of SNPs.
http://db.systemsbiology.net/projects/mhc/acgt/; ACGT (A Comparative Genomics Tool).
http://genome.cse.ucsc.edu; GoldenPath assembly.
http://rast.abajian.com/sputnik/; Sputnik program.
http://smart.embl-heidelberg.de; SMART.
http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl; Human Gene Nomenclature Database.
http://www.hgmp.mrc.ac.uk/Registered/Webapp/pix/; PIX protein identification program.
http://www.isrec.isb-sib.ch/software/PFSCAN_form.html; Profile Scan server.
http://www.sanger.ac.uk/HGP/Chr6/MHC/; Sanger Institute MHC haplotype project.
http://www.systemsbiology.org; Authors' Web site.

Received July 9, 2003; accepted in revised form September 18, 2003.