# Comparative Sequence Analysis of a Gene-Rich Cluster at Human Chromosome 12p13 and its Syntenic Region in Mouse Chromosome 6

M. Ali Ansari-Lari, John C. Oeltjen, Scott Schwartz, et al.

*Genome Res.* 1998 8: 29-40

RESEARCH

# Comparative Sequence Analysis of a Gene-Rich Cluster at Human Chromosome 12p13 and its Syntenic Region in Mouse Chromosome 6

**M. Ali Ansari-Lari,[1] John C. Oeltjen,[1] Scott Schwartz,[2] Zheng Zhang,[2] Donna M. Muzny,[1] Jing Lu,[1] James H. Gorrell,[1] A. Craig Chinault,[1] John W. Belmont,[1] Webb Miller,[2] and Richard A. Gibbs[1,3]**

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030 USA;
[2]Department of Computer Science and Engineering, Penn State University, University Park, Pennsylvania 16802 USA

The Human Genome Project has created a formidable challenge: the extraction of biological information from extensive amounts of raw sequence. With the increasing availability of genomic sequence from other species, one approach to extracting coding and regulatory element information is through cross-species sequence comparison. To assess the strengths and weaknesses of this methodology for large-scale sequence analysis, 227 kb of mouse sequence syntenic to a gene-rich cluster on human chromosome 12p13 was obtained. Primarily through percent identity plots (PIPs) of SIM comparative sequence alignments, the sequence of coding regions, putative alternative exons, conserved noncoding regions, and correlation in repetitive element insertions were easily determined. The analysis demonstrated that the number, order, and orientation of all 17 genes are conserved between the two species, whereas two human pseudogenes are absent in mouse. In addition, apart from MIRs, no direct correlation of distribution or position of the majority of repetitive elements between the two species is seen. Finally, in examining the synonymous and nonsynonymous substitution rates in the conserved genes, a large variation in nonsynonymous rates is observed indicating that the genes in this region are diverging at different rates. This study indicates the utility and strength of large-scale cross-species sequence comparisons in the extraction of biological information from raw sequence, especially when combined with other computational tools such as GRAIL and BLAST.

[The sequence data described in this paper have been submitted to the GenBank data library under accession nos. AC002393 and AC002397.]

While the human genomic sequence is being generated at an increasing rate, the extraction of biological information remains a formidable challenge. Together with exon prediction and gene modeling programs, data bases of ESTs (Hillier et al. 1996) have greatly improved the initial identification and characterization of the sequence. However, the complexity of the human genome requires additional approaches to efficiently determine complete gene structure and elucidate regulatory elements in a manner amenable to large-scale sequence analysis.

Large-scale cross-species DNA sequence comparisons are being used increasingly to address structural and evolutionary questions. Several studies have compared large segments (>30 kb) of human and rodent sequences, indicating that coding domains are generally well conserved, whereas noncoding (intronic and intergenic) domains exhibit highly variable levels of sequence conservation (Koop 1995; Hardison et al. 1997). Sequence comparison of the human and mouse T-cell receptor C-δ and C-α regions (Koop and Hood 1994), the human and hamster α- and β-myosin heavy chain genes (Epp et al. 1995), and the human and murine Bruton's tyrosine kinase loci (Oeltjen et al. 1997) revealed a high level of sequence similarity at coding and noncoding regions. However, sequence com-

[3]Corresponding author.
E-MAIL agibbs@bcm.tmc.edu; FAX (713) 798-5741.

parison of the human and mouse β-globin gene cluster (Collins and Weissman 1984; Sheehee et al. 1989), the human and rat γ-crystallin genes (den Dunnen et al. 1989), the human and mouse *XRRC1* DNA repair gene regions (Lamerdin et al. 1995), and the human, mouse, and hamster *ERCC2* gene regions (Lamerdin et al. 1996) showed less conservation of noncoding sequence. A mixed pattern of sequence similarity has been observed in the human and mouse immunoglobulin C(Mu)–C(δ) heavy chain loci (Koop et al. 1996). These studies suggest that portions of human and rodent genomes evolve at different rates (Koop 1995; Hardison et al. 1997).

The importance of cross-species sequence comparison for detection of regulatory elements has been suggested in numerous studies (Koop and Hood 1994; Lamerdin et al. 1995; Oeltjen et al. 1997). Increased conservation of noncoding segments has been observed 5′ and 3′ of the first exon of the genes in the BTK region. The conserved region flanking the first exon of BTK regulates, at least in part, the lineage-specific expression pattern of BTK (Oeltjen et al. 1997). Some of the noncoding sequence conservation could be attributable to constraint by splicing, chromatin condensation, matrix association, and replication origins (Koop 1995). However, some of the conserved noncoding regions clearly represent promoter and transcriptional regulatory elements (Koop 1995; Hardison et al. 1997; Oeltjen et al. 1997).

Although a great deal of attention has been directed toward computer approaches for annotation and gene predictions in a genomic sequence, there is no agreement about the precise goals of large-scale comparative sequence analysis, the computer methods to attain them, or how to report and/or annotate the results. These issues need to be resolved before the benefits of comparative sequence analysis can be accurately assessed.

Previously, we identified a gene-rich cluster comprising 17 complete genes, one partial gene, two likely pseudogenes, and other transcribed regions, at the *CD4* locus on human chromosome 12p13 (Ansari-Lari et al. 1996, 1997). The genes in this region possess diverse expression patterns and functions, ranging from signal transduction and glycolysis, to regulation of cell proliferation and ubiquitin-dependent proteolysis (Ansari-Lari et al. 1996, 1997). To gain insight into the evolution of coding and noncoding segments of this region and as a first step in elucidation of the regulatory elements of the genes, we isolated and sequenced the syntenic region in mouse chromosome 6. Several computational tools were utilized for human and mouse sequence comparison. This study indicates that genes closely clustered in one region have evolved at different rates.

## RESULTS AND DISCUSSION

### Organization of the Genes in the Mouse Sequence

Several mouse BAC clones corresponding to the gene-rich cluster at human chromosome 12p13 were isolated (Ansari-Lari et al. 1996, 1997). Based on PCR and hybridization analyses, one of the clones appeared to contain the majority of genes corresponding to the human cluster. The sequence of this clone was obtained using a modified shotgun M13 sequencing strategy (Richards et al. 1994). The mouse sequence is composed of two contigs totaling 226,708 bp, separated by a small gap of several hundred bases in a repeat-rich region. From the first exon of *CD4* to the penultimate exon of *C6*, the GC content of the human sequence is 51.35%, whereas that of the mouse sequence is 49.68%. The number of CpG islands estimated by GRAIL is 22 and 14 for the human and mouse sequences, respectively. It has been estimated that 20% of the CpG islands in humans are absent in mouse (Antequera and Bird 1993).

Initial analysis of the mouse sequence with PowerBLAST (Zhang and Madden 1997) revealed the presence of all the genes in the human cluster, with the exception of the two putative human pseudogenes. The number, order, and orientation of the genes are conserved between human and mouse. *RPL13-2* and *Destrin-2*, which appear to be pseudogenes in the human (Ansari-Lari et al. 1997), are absent in the mouse sequence, suggesting their formation subsequent to divergence of primates and rodents. The region encompassing *RPL13-2* and *Destrin-2* contains an unusually large number of repeat elements. In mouse, this region is also repeat-rich and contains an ETn mouse early transposon (Sonigo et al. 1987), which is absent in the human sequence.

The boundaries of the coding exons could be determined accurately for all the genes in this region. The exact exon–intron organizations of the mouse *CD4*, *TPI*, *ENO-2*, *DRPLA*, *U7* snRNA, *PTPN6*, and *BAP* genes were determined, based on the available published mouse cDNA sequences. In the absence of murine cDNA sequence information, the exact exon–intron organization of genes *A*, *B*, *GNB3*, *ISOT*, *B7*, *C2f*, and *C3f* were determined using homology to human cDNA sequences.

The nucleotide similarity for the coding do-

mains of the genes in this region ranges from 70.2% for *CD4* to 91.9% for *ISOT-1,* whereas the percent amino acid identity is between 56.3% for *CD4* and 100% for *BAP* (Table 1). These ranges of values are consistent with earlier observations for other human and murine genes (Makalowski et al. 1996). The sizes of the majority of the coding exons are conserved between human and mouse. Of the 153 coding and noncoding exons in this region (excluding the *C6* gene and the two putative pseudogenes), 141 are coding. Twelve of the coding exons in this region do not maintain the same number of nucleotides between the two species, although they differ by multiples of three nucleotides, conserving the reading frame. The sizes of 5′- and 3′-untranslated exons are not highly conserved between the two species. Sequence comparison may not allow the unequivocal identification of the 5′-untranslated boundaries of the genes; however, polyadenylation signals that are present in human are conserved in mouse. Similarly, the genes that are lacking recognizable polyadenylation signals are in common between the two species.

Several new EST matches that do not correspond to previously annotated exons of genes in this region have been observed. Some of these ESTs could be artifactual, whereas others could represent antisense transcripts, which have been hypothesized to play important roles in the regulation of gene expression (Nellen and Lichtenstein 1993). Alternatively, these ESTs could be differentially spliced exons or 5′-untranslated exons. An example of an alternatively spliced exon is seen in gene *B7,* where the human nucleotide positions 124740–125037 (see Fig. 2, below) show 76% sequence identity to the mouse sequence, and the ORF for this putative exon shows 87% amino acid similarity between the two species. GRAIL-2 (Xu et al. 1994) also predicts an exon corresponding to this region. An EST match (accession no. U25931) indicates splicing of this putative exon to the last exon of *B7.* This putative exon was not present in the original cDNA sequence for *B7;* however, human/mouse sequence comparison clearly directed us to look more closely at this region. An example of a putative antisense transcript for gene *C3f* is the IMAGE cDNA clone 40871 (accession no. U72507). The transcription of this clone, which contains a polyadenylation signal at its 3′ end, is opposite of gene *C3f* (Ansari-Lari et al. 1997).

Table 1. Nucleotide and Amino Acid Sequence Comparison Between the Coding Segments of the Genes

| Gene | Amino acid similarity (%) | Amino acid identity (%) | Nucleotide similarity (% coding cDNA) |
|---|---|---|---|
| *CD4* | 62.9 | 56.3 | 70.2 |
| *A-2* | 96.8 | 95.9 | 89.5 |
| *B* | 90.9 | 87.6 | 86.4 |
| *GNB3* | 98.2 | 97.0 | 89.8 |
| *C8* | 75.9 | 73.3 | 81.5 |
| *ISOT-1* | 98.6 | 98.4 | 91.9 |
| *TPI* | 96.0 | 96.0 | 88.7 |
| *C9* | 92.0 | 90.9 | 87.4 |
| *B7* | 86.6 | 79.7 | 81.6 |
| *ENO-2* | 99.1 | 98.6 | 91.7 |
| *DRPLA* | 95.1 | 94.3 | 88.9 |
| *C10* | 97.6 | 97.6 | 88.7 |
| *PTPN6 (hem)* | 96.1 | 94.4 | 89.2 |
| *BAP* | 100 | 100 | 91.3 |
| *C2f* | 92.2 | 88.9 | 85.0 |
| *C3f* | 93.2 | 89.8 | 88.6 |

For *B7*, the new identified exon was included. Sequence comparison was performed using the GCC BestFit program.

## Human/Mouse Sequence Comparison

Initially, the human and mouse sequences were compared by dot plot analysis using DOTTER software (Fig. 1) (Sonnhammer and Durbin 1995). The diagonal line represents conserved coding and noncoding segments of the genes. Three major repeat-rich regions can be observed at the beginning, middle, and end of the sequence, interrupting the diagonal line of sequence conservation. In addition, repetitive elements throughout the sequence are represented by matches away from the diagonal. The dot plot indicated a substantial global sequence conservation across the entire segment and the potential for large-scale alignment within well-defined domains.

A more detailed approach to cross-species sequence comparison is depicted in Figure 2. The repeat elements were masked by RepeatMasker [A.F.A. Smit and P. Green (1995–1997) http://ftp.genome.washington.edu/

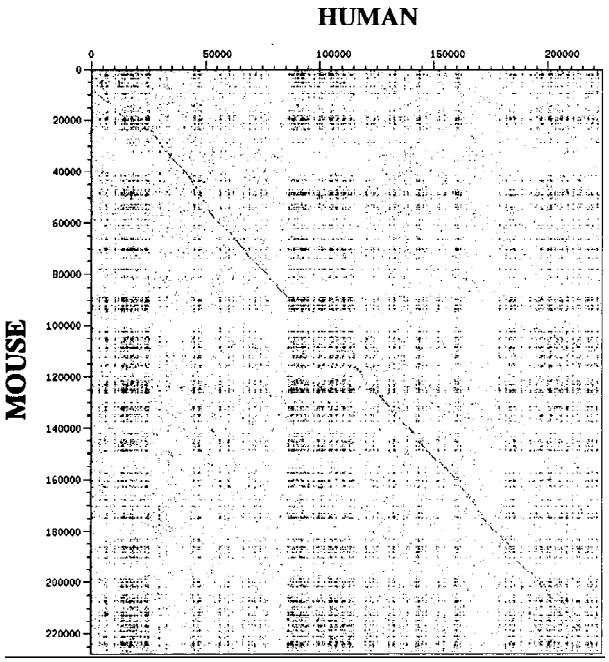


Figure 1 Human and mouse sequence comparison by DOTTER. The human and mouse nucleotide sequence positions are indicated on the axes.

cgi-bin/RepeatMasker], and using a modified version of SIM (similar) (Huang et al. 1990), the local alignments scoring above 95% in approximate significance (Schwartz et al. 1991) were determined. The gap-free alignments were subsequently converted into segments of percent identity relative to positions of human sequence, and the resulting data drawn as a percent identity plot (PIP), using a modified version of the local alignment to postscript (LAPS) program (Schwartz et al. 1991). In addition, EST matches and exons predicted by GRAIL-2 were mapped on the PIP.

The majority of the exons could be identified by visual inspection of the PIP, even in the absence of EST and GRAIL-2 matches. In particular, with the exception of 5′- and 3′-untranslated regions, every exon of an active gene shows conservation between human and mouse that contrasts sharply with the typical noncoding region. Alternatively, the PIP could be used to improve predictions made by other means. For instance, of 146 exons predicted by GRAIL-2, the requirement that a predicted exon contain a gap-free aligned segment of at least 40 nucleotides eliminates 21, which are all false positives. Sequence conservation in intronic and intergenic regions is observed throughout the region. In particular, of 171 gap-free aligned regions of at least 100 bp in length, 34 do not overlap an identified exon. However, significantly more sequence conservation is observed upstream of the first exon and within the first intron of each gene. There are regions with strikingly high matches in other introns, such as intron 6 of *B7*, and introns 4 and 12 of *PTPN6*. The PIP indicates that the genes in this region have diverged at different rates, with *CD4* being the most divergent. Furthermore, interruption in sequence alignment is observed in highly repeat-rich regions in intron 3 of *CD4*, intergenic region between *C9* and *B7*, and the region downstream of *C3f*.

Based on sequence conservation identified by the PIP, three regions with potentially new genes were considered (Fig. 2). All three regions contain at least one EST match and exhibit 60%–86% sequence conservation, at least in part, between the two species. The corresponding human and mouse sequences were analyzed by GRAIL-2 (Xu et al. 1994), FGENEH (Solovyev et al. 1995), and Genie (Kulp et al. 1996) exon prediction/gene modeling programs (data not shown). None of the programs predicted a consistent gene model that was common to the two species in any of these three regions. However this analysis does not exclude the possibility of the presence of a gene in either of these regions. One of these is an intergenic segment between *DRPLA* and *ENO-2*, containing an EST match (accession no. M78236). This EST could represent the 5′-untranslated region of *DRPLA*. The second region is located between the 3′ ends of *PTPN6* and *BAP* genes. In this region there are two mouse overlapping EST matches (accession nos. AA423763 and AA272334). No continuous ORF is present in either EST. The third region is between the *C3f* and *C6f* genes, containing an EST (accession no. T83233) partially conserved between the two species. The segments of sequence conservation in these three regions could represent regulatory elements. Further experiments are required to resolve the significance of these three regions.

## Extent of Coding and Noncoding Sequence Conservation

Using a modified version of CONSERVED (Oeltjen et al. 1997), the gap-free segments (≥50 bp) showing sequence identity (≥60%) were determined from the overall SIM alignment. Weighted percent identities were calculated for transcribed (coding and noncoding exons) and intronic noncoding regions between the first and last exon of each gene (Table 2). Using these CONSERVED parameters, 85.90% of the total weighted transcribed region,

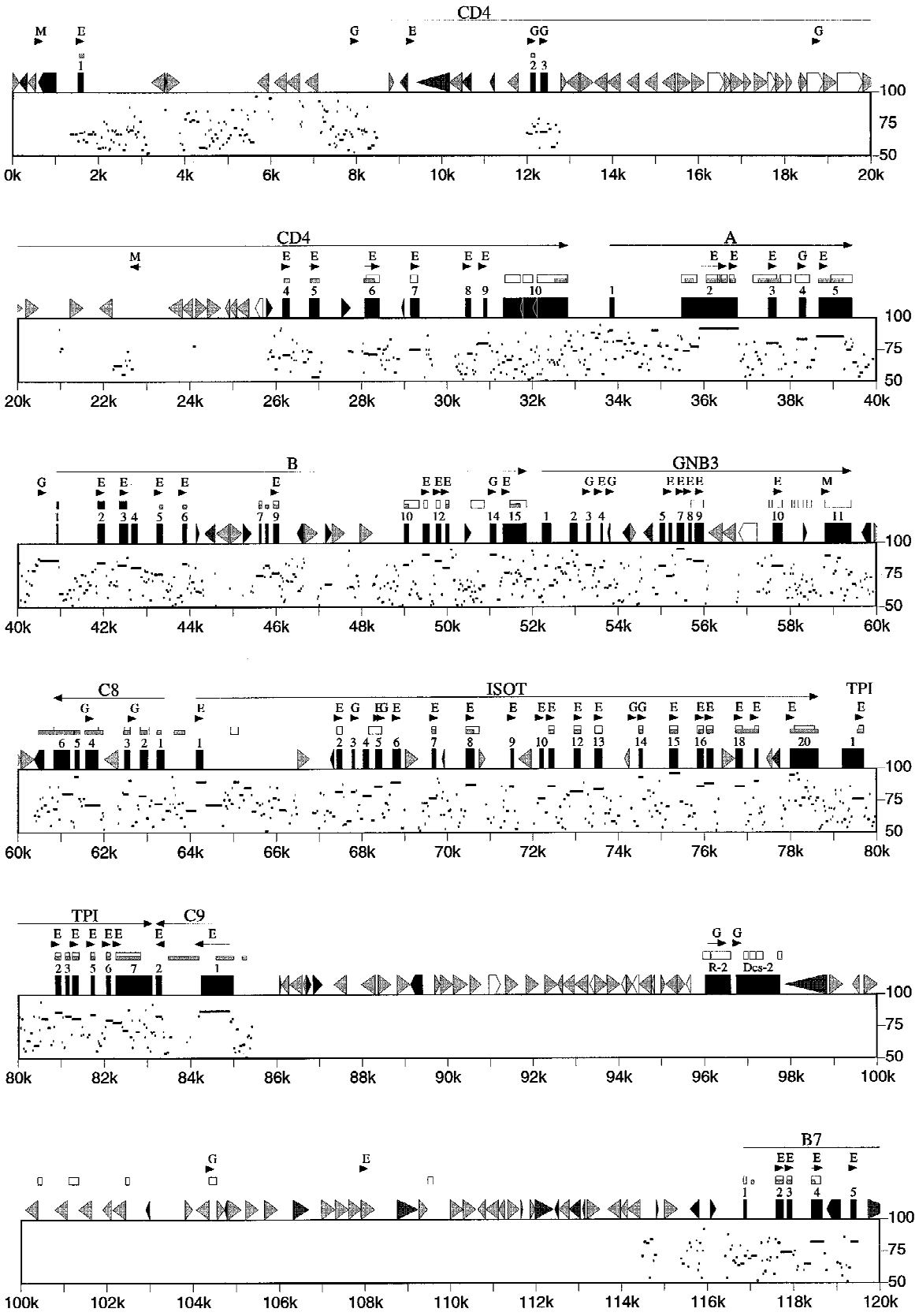


Figure 2 (*See following page for legend.*)

and 72.65% of the total weighted noncoding region are conserved. For *CD4*, the ratio of transcribed to noncoding sequence conservation has approached unity (Table 2). This is mainly due to extensive di-

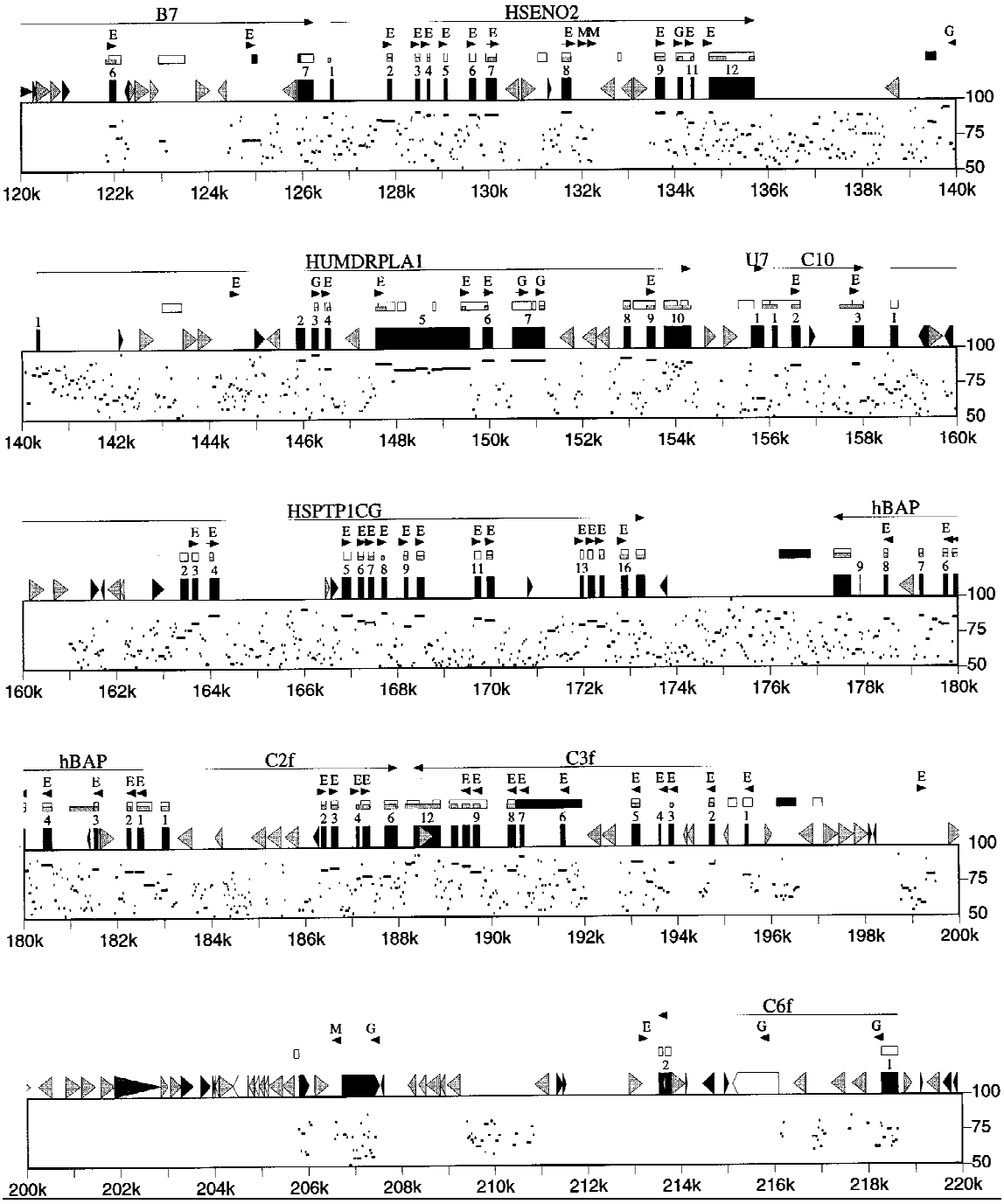


Figure 2 The PIP. The nucleotide position for the human sequence is shown in the *x*-axis, and the percent sequence identity (50%–100%) is shown in the *y*-axis. Exons (numbered black boxes), and repeats (SINEs other than MIR are light gray triangles pointing toward the A-rich 3′ end; LINE1s are open arrow boxes; MIR and LINE2 elements are black triangles and pointed boxes, respectively; other interspersed repeats are dark gray triangles) are indicated above the main boxes; BLAST hits in dbEST as shorter white (human), gray (mouse); or black (ESTs discussed in the text) boxes above that; and GRAIL-2 exon predictions (email to grail@ornl.gov) above EST matches. The GRAIL-2 hits show direction, extent, and quality [(E) excellent; (G) good; (M) marginal].

Table 2. Sequence Conservation of Transcribed and Noncoding Regions

| | Sum length (transcribed) | Sum length (noncoding) | Weighted conservation (% transcribed) | Weighted conservation (% noncoding) |
|---|---|---|---|---|
| *CD4* | 1492 | 2435 | 71.93 | 69.60 |
| *A* | 2310 | 1245 | 86.89 | 75.66 |
| *B* | 1648 | 1347 | 86.15 | 69.27 |
| *GNB3* | 1319 | 846 | 87.28 | 69.89 |
| *C8* | 1073 | 298 | 77.99 | 70.76 |
| *ISOT* | 2792 | 2961 | 90.49 | 72.81 |
| *TPI* | 1139 | 761 | 81.96 | 70.80 |
| *C9* | 791 | 83 | 87.35 | 66.41 |
| *B7* | 1117 | 862 | 79.84 | 69.45 |
| *ENO-2* | 1646 | 1507 | 88.39 | 73.40 |
| *DRPLA* | 3853 | 1621 | 89.05 | 73.36 |
| *C10* | 447 | 70 | 85.17 | 74.31 |
| *PTPN6* | 2007 | 1996 | 87.31 | 71.55 |
| *BAP* | 1110 | 1007 | 87.66 | 73.59 |
| *C2f* | 850 | 252 | 82.70 | 70.04 |
| *C3f* | 1343 | 637 | 88.88 | 73.47 |
| Intergenic regions | | 7241 | | 74.63 |
| Totals | 24,937 | 25,169 | 85.90 | 72.65 |

The gap-free segments (≥50 bp) showing sequence identity (≥60%) were extracted from SIM alignments using the modified version of CONSERVED. Based on coordinates of exons, the alignments were divided into transcribed and noncoding regions. ESTs that do not correspond to known gene-specific mRNAs were not included in the calculation for transcribed regions. Because the transcription initiation site for some of the genes in this region is not known, the conserved intergenic noncoding segments were calculated separately. For each gene, a weighted percent identity was calculated by $\Sigma Ci\,Fi/N$, where *Ci* is the length in base pairs of each individual aligned sequence of either transcribed or noncoding sequence, *Fi* is the percent of identity of the aligned sequence, and *N* is the total number in base pairs of compared transcribed or noncoding sequence.

vergence at the coding region rather than higher conservation at the noncoding region.

## Nucleotide Substitution Rate

The synonymous and nonsynonymous substitution rates of the coding regions were determined by method 1 of Ina (1995) (Table 3). Similar results were obtained by Ina's method 2 (data not shown). On average, the synonymous substitution rate was approximately six times higher than the nonsynonymous substitution rate. As expected, the synonymous substitution rate exhibits less variation (~1.6-fold range of variation). The gene-specific differences in synonymous substitutions suggest either purifying selection or local differences in mutation rate (Wolfe et al. 1989). The rate of nonsynonymous substitution, on the other hand, is extremely variable among the genes, ranging from $0.01 \times 10^{-9}$ substitution per nonsynonymous site per year for *BAP* to $2.26 \times 10^{-9}$ for *CD4* (~225-fold difference in variation). This high variation in nonsynonymous substitution rates could be due to adaptive selection and/or region-specific variation in mutation rate (Li and Graur 1991), although the contribution from the latter would likely be minimal, considering the close physical proximity of the genes in this region.

In mammals, synonymous substitution frequencies are gene-specific and they correlate with nonsynonymous substitution frequencies, implying that homologous mammalian genes evolve at similar rates (Mouchiroud et al. 1995). In addition, tissue-specific gene expression could determine the rate of evolution of genes, likely attributable to global functional constraints (Kuma et al. 1995). In vertebrates, proteins that are expressed in the immune system appear to evolve more rapidly than those expressed in nervous system, as has been suggested for members of protein kinase and immunoglobulin

Table 3. Synonymous and Nonsynonymous Substitution Rates

| | Synonymous ($\times 10^9$) | Non-synonymous ($\times 10^9$) | $K_a/K_s$ |
|---|---|---|---|
| *CD4* | 2.93 ± 0.33 | 2.26 ± 0.15 | 0.77 |
| *A-2* | 2.41 ± 0.22 | 0.14 ± 0.03 | 0.06 |
| *B* | 2.57 ± 0.25 | 0.41 ± 0.05 | 0.16 |
| *GNB3* | 2.56 ± 0.32 | 0.11 ± 0.03 | 0.04 |
| *C8* | 2.26 ± 0.32 | 1.03 ± 0.12 | 0.46 |
| *ISOT-1* | 1.98 ± 0.16 | 0.06 ± 0.01 | 0.03 |
| *TPI* | 2.77 ± 0.40 | 0.16 ± 0.04 | 0.06 |
| *C9* | 2.45 ± 0.34 | 0.32 ± 0.06 | 0.13 |
| *ENO-2* | 2.01 ± 0.23 | 0.06 ± 0.02 | 0.03 |
| *DRPLA* | 2.33 ± 0.16 | 0.19 ± 0.02 | 0.08 |
| *C10* | 3.07 ± 0.61 | 0.07 ± 0.04 | 0.02 |
| *PTPN6 (hem)* | 2.49 ± 0.24 | 0.18 ± 0.03 | 0.07 |
| *BAP* | 2.16 ± 0.29 | 0.01 ± 0.01 | 0.005 |
| *C2f* | 3.04 ± 0.44 | 0.43 ± 0.08 | 0.14 |
| *C3f* | 2.03 ± 0.25 | 0.35 ± 0.05 | 0.17 |
| *B7* | 3.18 ± 0.40 | 0.88 ± 0.09 | 0.28 |
| Average | 2.52 ± 0.31 | 0.42 ± 0.05 | 0.16 |

The numbers represent average substitution rate per nucleotide ± S.D., calculated by method 1 of Ina (1995). The estimate of human and rodent divergence of 80 million years was used in the calculation of rates (Li and Graur 1991). ($K_s$) Substitution rate per synonymous site; ($K_a$) substitution rate per nonsynonymous site.

gene families (Kuma et al. 1995). The rate of evolution of C2 domains of members of the immunoglobulin superfamily, including *CD4,* have been compared for human and murine rodents (Hughes 1997). There is a strong positive association between expression in immune system and nonsynonymous substitution rate at C2 domains (Hughes 1997). The major histocompatibility complex (MHC) proteins and T-cell receptors, all lacking C2 domains, also exhibit high nonsynonymous substitution rates, which are believed to represent adaptive selection (Hughes 1997). Because CD4 interacts with class II MHC molecules, due to amino acid diversity in the class II MHC molecules, CD4 might be under weaker structural constraints and exhibit a faster rate of evolution (Hughes 1997).

## Distribution of Repetitive Elements

From the first exon of *CD4* to the putative penultimate exon of gene *C6f,* repetitive elements represent 33.36% and 26.39% of the human and mouse sequences, respectively. The number and the percentage of sequence occupied by different classes of repetitive elements is summarized in Table 4. These repeats can be mapped to the genomic sequence of the other species with some confidence.

Many of the mouse interspersed repeats could be accurately mapped onto human positions as follows. The alignment program was applied to the unmasked human and mouse sequences, and the four largest contiguous alignments, which covered human positions 1349–5688, 29117–85438, 116301–118786, and 124669–194806, totaling 59.8% of the human sequence and 60.9% of the mouse sequence, were extracted. The alignment of masked sequences, which was used to make the PIP, consists of 43 aligned segments totalling 64.8% of the human sequence. Thus, these four pieces contain the vast majority of the aligned positions. The distribution of repetitive elements in the unmasked aligned segments is summarized in Table 4.

As expected, there is a tendency for inserted elements to occur in unaligned regions, as their presence disrupts the alignment. In this respect, *Alu* repeats are more disruptive than common mouse SINEs, such as B1 elements, because of their greater length, a finding that may explain the lower frequency of *Alu* repeats in the aligned regions. MIRs (mammalian-wide interspersed repeats) are highly abundant tRNA-derived SINEs (Smit and Riggs 1995). Detected MIRs in the human sequence did not favor either aligned or unaligned regions. In mouse, many fewer MIR relics were found by RepeatMasker, and they also all occurred in aligned regions. Because of the higher rate of mutation in rodents (Li et al. 1990), one would expect that the MIRs detectable in mice will tend to occur in well-conserved regions of the mouse genome, in accordance with what we observed. MIRs have been found at orthologous sites in different mammals and, hence, are believed to have amplified, at least in part, before the mammalian radiation (Jurka et al. 1995; Smit and Riggs 1995). Available evidence suggests that MIR activity stopped before the divergence of humans and mice (Jurka et al. 1995). Table 5 lists the potential orthologous repeat

Table 4. Distribution of Repetitive Elements

| | Number of elements/ [percentage of sequence (human)] | Number of elements in aligned regions (human) | Number of elements/ [percentage of sequence (mouse)] | Number of elements in aligned regions (mouse) |
|---|---|---|---|---|
| *Alu* | 183 (21.77) | 55 | — | — |
| B1 | — | — | 156 (18.32) | 69 |
| B2 | — | — | 85 (8.25) | 32 |
| B4 | — | — | 36 (6.51) | 18 |
| ID | — | — | 30 (2.32) | 16 |
| MIR | 29 (1.86) | 18 | 8 (0.34) | 8 |
| LINE1 | 9 (2.15) | 2 | 9 (0.85%) | 3 |
| LINE2 | 14 (1.52) | 5 | 5 (0.21) | 3 |
| MaLR | 2 (0.39) | 0 | 24 (1.61) | 7 |
| Retroviruses | 2 (0.9) | 0 | 4 (0.36) | 2 |
| MER4_group | 1 (0.2) | 0 | 0 | 0 |
| MER1 type | 6 (0.37) | 3 | 3 (0.18) | 2 |
| MER2 type | 5 (0.23) | 2 | 3 (0.27) | 2 |
| Small RNA | 2 (0.04) | 1 | 7 (0.22) | 6 |
| Simple repeats | 62 (2.44) | 38 | 107 (3.35) | 64 |

The distribution of human and mouse sequences between the first exon of *CD4* and the penultimate exon of *C6* (columns 1 and 3), and the four unmasked aligned fragments (columns 2 and 4) were determined by RepeatMasker (''.tbl'' output), with slight modifications.

segments, as suggested by alignments and confirmed by inspection. Several other MIRs were detected by RepeatMasker in one of the sequences but either postdate the rodent–primate split or had mutated beyond recognition in the other species. A MIR fragment at human position 47150–47297 is 68% identical between human and mouse, which suggests that it may have been incorporated into a functional element. Makalowski et al. (1996) report a MIR insertion that retains 65.4% identity. Additionally, MIRs have been identified as parts of coding and 3′-untranslated segments of a few mammalian genes (Murnane and Morales 1995).

More than half of all LINE-1 elements are believed to be inserted into the genome before the mammalian radiation, based on their presence at orthologous sites in different mammalian genomes (Smit et al. 1995). RepeatMasker identified nine human and nine mouse LINE1 fragments in the region studied here. Only two elements from human and three from mouse could be aligned with any confidence. Each of the human L1s was in roughly the same position as a mouse L1. Furthermore, the orientations, and in one case the LINE1 subclassification, were consistent with orthology. However, extensive divergence of these elements, par-

Table 5. Orthologous Repeat Elements

| | | Position | |
|---|---|---|---|
| **Repeat type** | **Orientation** | **human** | **mouse** |
| LINE2 | + | 3532–3574 | 10318–10358(a) |
| MIR | – | 31956–32106 | 31513–31600(a) |
| MIR | + | 47150–47297 | 49040–49163(a) |
| MER58a | – | 77576–77751 | 79993–80124(a) |
| MIR | + | 142076–142154 | 53214–53268(b) |
| MIR | + | 162765–162982 | 73486–73598(b) |
| MIR | + | 170795–170907 | 82766–82810(b) |
| MIR | – | 173648–173801 | 86906–87038(b) |

The human and mouse nucleotide sequence positions of likely orthologous repeat segments and their orientations are shown. *a* and *b* correspond to GenBank accession nos. AC002393 and AC002397, respectively.

ticularly in mouse, and proximity to other repeat elements ruled out a definitive conclusion as to orthology.

In this study, ~223 kb of human sequence on chromosome 12p13 was compared with 227 kb of syntenic mouse sequence on chromosome 6. Except for two pseudogenes that appear to be absent in mouse, the number, organization, and orientation of the genes are maintained between the two species. Conservation in the noncoding regions was observed more frequently in the areas flanking the first exon of each gene, representing the likely positions of regulatory elements. The genes in this region appear to be evolving at different rates, as inferred from great variation in their nonsynonymous substitution rates. Several orthologous MIRs were identified, supporting the insertion of MIRs before mammalian radiation. However, for the majority of the repeat elements, no direct correlation between their distributions and positions between the two species was found.

Cross-species sequence comparison permits variations in the rate of sequence conservation to be measured across a genomic region. That information is useful for identifying coding regions, regulatory signals, and other functional segments of the genome. Dynamic programming alignment methods permit precise identification of conserved segments, and recent algorithmic improvements allow 200-kb sequences to be aligned in a minute on a 200-MH workstation.

PIPs provide a useful visualization of sequence conservation. They could be added to interactive programs (Harris 1997; Zhang and Madden 1997) that view results from database searches and gene-prediction programs as an aid to annotating genomic sequence data. Alternatively, information from the alignment can be incorporated into an explicit gene-prediction algorithm, just as results from database searches are now incorporated (Xu and Uberbacher 1997). This will allow the benefits of genomic alignments as a source of information for gene prediction to be measured objectively.

## METHODS

### Isolation and Sequencing of the Mouse BAC

The initial screen of the mouse BAC library (Research Genetics, Inc.) was performed by PCR using four primer pairs specific for mouse *CD4* (R2633, 5′-CCATCTCTCTTAGGCGCTTG-3′ and R2634; 5′-GAACTTCCAGGTGAAGACTG-3′), *TPI* (R2635, 5′-CTTGGTTTGCTCGAACACGAC-3′; and R2636, 5′-CTGGCATGATCAAAGACTTAG-3′), *ENO-2* (R2673, 5′-GATCAATGGTGGCTCTCATG-3′; and R2674, 5′-CTGTTCTCCAGGATATTGGG-3′), and *BAP* (R2671, 5′-CCTGCACAATCTTCTGTCGC-3′; and R2672, 5′-CAAGGACTTCAGCCTCATCC-3′) genes. The five positive subpools were arrayed on filters and screened using radioactively labeled probes specific for mouse *CD4*, *TPI*, and *BAP* genes. Labeling and hybridization were performed as described (Ansari-Lari et al. 1996). A similar restriction pattern for all overlapping BACs was observed. Only the clone 284H12 was positive for all three probes. The shotgun M13 sequencing library and M13 sequencing templates were generated as described (Andersson et al. 1996a,b). Sequence assembly was performed by GAP (Bonfield et al. 1995) and Phrap software (Phil Green, http://genome.wustl.edu/gsc/finishing/PHRAP-INTRO.html).

### Computer Software

The repeat elements were masked using the new release of RepeatMasker [A.F.A. Smit and P. Green (1995–1997) http://ftp.genome.washington.edu/cgi-bin/RepeatMasker]. Initial analysis of the sequence was performed by PowerBLAST (Zhang and Madden 1997) searching databases for nonredundant sequences, ESTs and STSs, using default parameters. CpG islands were found by GRAIL (http://avalon.epm.ornl.gov/Grail-1.3). Dot plot analysis was performed using DOTTER (Sonnhammer and Durbin 1995). cDNA and protein sequence comparison was performed by the GCG BestFit program. Synonymous and nonsynonymous substitution rates were determined using DISTS1 and DISTS2 methods from Ina (1995). GRAIL-2 (Xu et al. 1994), FGENEH (Solovyev et al. 1995), and Genie (Kulp et al. 1996) were employed for exon prediction/gene modeling programs.

Alignments of the genomic sequences were computed using the parameters specified in Oeltjen et al. (1997), though a new alignment program was written to give a 100-fold speed-up. Alignment analysis (by the CONSERVED program) and display (Fig. 2) were performed as described (Oeltjen et al. 1997). Programs were written to automatically process output from RepeatMasker, Grail-2, and BLAST (Altschul et al. 1990), where EST hits were required to have length of at least 60% and at least 90% identity.

## ACKNOWLEDGMENTS

We thank Dr. Arian Smit for helpful discussions and Michael Chiu for help in submission of the data. This work was supported in part by the grant RO1 HG01459 from the National Center for Human Genome Research. S.S., Z.Z., and W.M. were supported by grant LM05110 from the National Library of Medicine.

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

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215: 403–410.

Andersson, B., J. Lu, K.E. Edwards, D.M. Muzny, and R.A. Gibbs. 1996a. Method for 96-well M13 DNA template

preparations for large-scale sequencing. *BioTechniques* 20: 1022–1027.

Andersson, B., M.A. Wentland, J.Y. Ricafrente, W. Liu, and R.A. Gibbs. 1996b. A ''double adaptor'' method for improved shotgun library construction. *Anal. Biochem.* 236: 107–113.

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

Ansari-Lari, M.A., Y. Shen, D.M. Muzny, W. Lee, and R.A. Gibbs. 1997. Large-scale sequencing in human chromosome 12p13: Experimental and computational gene structure determination. *Genome Res.* 7: 268–280.

Antequera, F. and A. Bird. 1993. Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci.* 90: 11995–11999.

Bonfield, J.K., K.F. Smith, and R. Staden. 1995. A new DNA sequence assembly program. *Nucleic Acids Res.* 25: 4992–4999.

Collins, F. and S.M. Weissman. 1984. The molecular genetics of human hemoglobin. *Prog. Nucleic Acid Res. Mol. Biol.* 31: 315–462.

den Dunnen, J.T., J.W. van Neck, F.P. Cremers, N.H. Lubsen, and J.G. Schoenmakers. 1989. Nucleotide sequence of the rat gamma-crystallin gene region and comparison with an orthologous human region. *Gene* 78: 201–213.

Epp, T., R. Wang, M. Sole, and C.C. Liew. 1995. Concerted evolution of mammalian cardiac heavy chain genes. *J. Mol. Evol.* 41: 284–292.

Hardison, R.C., J. Oeltjen, and W. Miller. 1997. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.* 7: 959–966.

Harris, N.L. 1997. Genotator: A workbench for sequence annotation. *Genome Res.* 7: 754–762.

Haung, X., R. Hardison, and W. Miller. 1990. A space-efficient algorithm for local similarities. *Comput. Appl. Biosci.* 6: 373–381.

Hillier, L., G. Lennon, M. Becker, M.F. Bonaldo, B. Chiapelli, S. Chissoe, N. Dietrich, T. DuBuque, A. Favello, W. Gish et al. 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* 6: 807–828.

Hughes, A.L. 1997. Rapid evolution of immunoglobulin superfamily C2 domains expressed in immune system cells. *Mol. Biol. Evol.* 14: 1–5.

Ina, Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *J. Mol. Evol.* 40: 190–226.

Jurka, J., E. Zietkiewicz, and D. Labuda. 1995. Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era. *Nucleic Acids Res.* 23: 170–175.

Koop, B.F. 1995. Human and rodent DNA sequence comparisons: A mosaic model of genomic evolution. *Trends Genet.* 11: 367–371.

Koop, B.F. and L. Hood. 1994. Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nature Genet.* 7: 48–53.

Koop, B.F., J.E. Richards, T.D. Durfee, J. Bangsberg, J. Wells, A.C. Gilliam, H.L. Chen, A. Clausell, P.W. Tucker, and F.R. Blattner. 1996. Analysis and comparison of mouse and human immunoglobulin heavy chain JH-Cmu-Cdelta locus. *Mol. Phylogenet. Evol.* 5: 33–49.

Kulp, D., D. Haussler, M.G. Reese, and F.H. Eeckman. 1996. A generalized hidden Markov model for the recognition of human genes in DNA. In *Proceeding Conference on Intelligent Systems in Molecular Biology '96* (ed. D.J. States, P. Agarwal, T. Gaasterland, L. Hunter, and R. Smith), pp. 134–142. AAAI/MIT Press, Menlo Park, CA.

Kuma, K.-I., N. Iwabe, and T. Miyata. 1995. Functional constraints against variations on molecules from the tissue level: Slowly evolving brain-specific genes demonstrated by protein kinase and immunoglobulin supergene families. *Mol. Biol. Evol.* 12: 123–130.

Lamerdin, J.E., M.A. Montgomery, S.A. Stilwagen, L.K. Scheidecker, R.S. Tebbs, K.W. Brookman, L.H. Thompson, and A.V. Carrano. 1995. Genomic sequence comparison of the human and mouse XRCC1 DNA repair gene regions. *Genomics* 25: 547–554.

Lamerdin, J.E., S.A. Stilwagen, M.H. Ramirez, L. Stubbs, and A.V. Carrano. 1996. Sequence analysis of the ERCC2 gene regions in human, mouse, and hamster reveals three linked genes. *Genomics* 34: 399–409.

Li, W.-H. and D. Graur. 1991. *Fundamentals of molecular evolution.* Sinauer Associates, Sunderland, MA.

Li, W.-H., M. Gouy, P.M. Sharp, C. O'hUigin, and Y.-W. Yang. 1990. Molecular phylogeny of rodentia, lagomorpha, primates, artiodactyla, and carnivora and molecular clocks. *Proc. Natl. Acad. Sci.* 87: 6703–6707.

Makalowski, W., J. Zhang, and M.S. Boguski. 1996. Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res.* 6: 846–857.

Mouchiroud, D., C. Gautier, and G. Bernardi. 1995. Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. *J. Mol. Evol.* 40: 107–113.

Murnane, J.P. and J.F. Morales. 1995. Use of a mammalian interspersed repetitive (MIR) element in the coding and

processing sequences of mammalian genes. *Nucleic Acids Res.* 23: 2837–2839.

Nellen, W. and C. Lichtenstein. 1993. What makes an mRNA anti-sense-itive? *Trends Biochem. Sci.* 18: 419–423.

Oeltjen, J.C., T.M. Malley, D.M. Muzny, W. Miller, R.A. Gibbs, and J.W. Belmont. 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.

Richards, S., D.M. Muzny, A.B. Civitello, F. Lu, and R.A. Gibbs. 1994. Sequence map gaps and directed reverse sequencing for the completion of large sequencing projects. In *Automated DNA sequencing and analysis* (ed. M.D. Adams, C. Fields, and J.C. Venter), pp. 191–198. Academic Press, San Diego, CA.

Schwartz, S., W. Miller, C.-M. Yang, and R.C. Hardison. 1991. Software tools for analyzing pairwise alignments of long sequences. *Nucleic Acids Res.* 19: 4663–4667.

Sheehee, W.R., D.D. Loeb, N.B. Adey, F.H. Burton, N.C. Casavant, P. Cole, C.J. Davies, R.A. McGraw, S.A. Schichman, D.M. Severynse, C.F. Voliva, F.W. Weyter, G.B. Wisely, M.H. Edgell, and C.A. Hutchinson. 1989. Nucleotide sequence of the BALB/c mouse β-globin complex. *J. Mol. Biol.* 205: 41–62.

Smit, A.F.A. and A.D. Riggs. 1995. MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Res.* 23: 98–102.

Smit, A.F.A., G. Toth, A.D. Riggs, and J. Jurka. 1995. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J. Mol. Biol.* 246: 401–417.

Solovyev, V.V., A.A. Salamov, and C.B. Lawrence. 1995. Identification of human gene structure using linear discriminant functions and dynamic programming. In *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology* (ed. C. Rawling, D. Clark, R. Altman, L. Hunter, T. Lengauer, and S. Wodak), pp. 367–375. AAAI Press, Cambridge, UK.

Sonigo, P., S. Wain-Hobson, L. Bougueleret, P. Tiollais, F. Jacob, and P. Brulet. 1987. Nucleotide sequence and evolution of ETn elements. *Proc. Natl. Acad. Sci.* 84: 3768–3771.

Sonnhammer, E.L.L. and R. Durbin. 1995. A dot matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 167: GC1–10.

Wolfe, K.H., P.M. Sharp, and W.H. Li. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* 337: 283–285.

Xu, Y., and E.C. Uberbacher. 1997. Automated gene identification in large-scale genomic sequences. *J. Comput. Biol.* 4: 325–338.

Xu, Y., R.J. Mural, M.B. Shah, and E.C. Uberbacher. 1994. Recognizing exons in genomic sequence using GRAIL II. In *Genetic engineering: Principles and methods* (ed. J. Setlow), Vol. 16, pp. 241–253. Plenum Press, New York, NY.

Zhang, J. and T.L. Madden. 1997. PowerBLAST: A new network BLAST application for interactive automated sequence analysis and annotation. *Genome Res.* 7: 649–656.

*Received September 8, 1997; accepted in revised form December 2, 1997.*