Genes from the ancient cytochrome P450 (CYP) superfamily encode a diverse group of monooxygenases that play important roles in both endogenous processes and in the metabolism of exogenous compounds, including many drugs. A cluster of Cyp2 genes on mouse chromosome 7 was mapped and analyzed in detail and compared with the homologous cluster on human chromosome 19. The mouse cluster includes 22 loci from the same six CYP2 subfamilies—Cyp2a, Cyp2b, Cyp2f, Cyp2g, Cyp2s, Cyp2t—that are found in the human cluster. Twelve of these loci are functional genes, and 10 are pseudogenes. Parts of the mouse and human gene clusters are similarly arranged, but the data indicate that a significantly different series of duplication events created the modern gene organizations in the two species. The comparison of the mouse and human clusters provides new insights into the evolution of gene families, whereas the detailed analysis provides background information that should be informative for future studies on the expression, regulation, and function of specific mouse Cyp2 genes. **Key words:** Cyp2a, Cyp2b, Cyp2f, Cyp2g, Cyp2s, Cyp2t, cytochrome P450, gene family, gene duplication. **Environ Health Perspect** 111:1835–1842 (2003). doi:10.1289/txg.6546 available via [http://dx.doi.org/](http://dx.doi.org/) (Online 24 September 2003)
techniques—because of high sequence similarity, a closely related gene or pseudogene can easily be mistaken for the target gene during Southern blotting or PCR amplifications. The general procedure discussed below was followed for all loci, but specific techniques were integrated in different combinations as necessary to localize and analyze each putative locus.

Mouse bacterial artificial chromosome (BAC) clones overlapping the Cyp2a-t cluster region were identified by library screening or database analysis, as described below. Specific exons and introns of CYP2 genes (see Supplemental Material) were then PCR-amplified from clone DNAs and sequenced. Some of the BAC clone DNAs were digested by multiple restriction enzymes, blotted onto nylon membranes, and hybridized with PCR-amplified products. Clones were assembled on the basis of shared gene sequences and on restriction mapping. Clone overlaps were confirmed by developing sequence tag sites from the ends of each clone and testing them against other cloned DNAs and against known sequence fragments from both the public and Celera (Celera Genomics, Rockville, MD) mouse genome projects.

Library Screening and Isolation of Bacterial Artificial Chromosome DNA

The RP22 mouse BAC library was screened (Invitrogen Corp., Carlsbad, CA) with PCR products amplified from mouse genomic DNA, using primers designed to match one exon each from the Cyp2a5, Cyp2b9, and Cyp2f2 cDNA sequences. The BAC library clones RP22-78A19, -44B20, -362C15, -127H7, -548M4, -160O14, and RP23-430G14, and RP23-174D7 were used only as PCR templates for specific primers to establish patterns of clone overlap. RP22-78A19, -44B20, -362C15, -127H7, -548M4, -160O14, and RP23-430G14, -174D7, and -113D13 were used for all other experiments in this study, including PCR amplifications, sequencing, restriction mapping, and Southern blotting. Unless otherwise noted, these nine clones are the BAC clones referred to throughout the rest of this article.

PCR Amplification

Polymerase chain reaction was performed on all samples using primers designed from mouse Cyp2 gene cDNA and genomic sequences in GenBank and from sequences available from Celera Genomics. All primer sequences and annealing temperatures can be found in Table 1 of the Supplemental Material or on the Cytochrome P450 Homepage (Nelson 2003). PCR amplification was carried out in a total volume of 25 µl consisting of 1X PCR buffer (10 mM Tris-HCl, pH 9.0; 50 mM KCl; 0.1% Triton X-100), 1 mM MgCl2, 200 µM each deoxynucleoside triphosphate, 0.5 µM each primer, 0.75 units of Taq polymerase (Promega Corp., Madison, WI), and 100–500 ng template DNA. Amplification reactions were performed with various annealing temperatures and cycles. PCR products were electrophoresed on 1.5% low-melt agarose gels, from which fragments were excised and purified using the QAquick Gel Extraction Kit (Qiagen).

Subcloning and Sequencing

Sequencing templates were isolated in either of two ways. For most of the Cyp2a-t loci, PCR amplifications were performed directly from BAC or genomic DNA preparations. In addition, some restriction fragments of BAC clones RP22-78A19, RP22-160O14, RP22-548M4, RP23-430G14, and RP23-174D7 were subcloned into plasmid vector pBlueScript KS- (Stratagene Inc, La Jolla, CA). Five to ten positive subclones were chosen by blue/white screening for each experiment, and plasmid DNAs were recovered using alkaline lysis minipreps, followed by direct sequencing using the T3 and T7 priming sites.

All sequencing was done for both strands, using either BigDye or ET terminator chemistry on an ABI 310 automated DNA sequencer (Applied Biosystems, Foster City, CA). All samples were prepared for sequencing according to manufacturer instructions. The forward and reverse primers were the same primers used for PCR. DNA sequences were analyzed using the NCBI BLAST similarity search tool (http://www.ncbi.nlm.nih.gov/BLAST/).

Table 1. Table of Cyp2a-t subfamily cluster genes in mouse.

| Locus | Orientation | Location | Exons | mRNA | Gap |
|-------|-------------|----------|-------|------|-----|
| 2b4   | Cen → Tel   | 310 157707-161230 | 1–9 | NM_007817 | |
| 212   | Cen → Tel   | 310 124085-136041 | 1–9 | NM_007817 | |
| 2a20-20 | Tel → Cen | 310 52619-53192 | 1–2 | NM_133657 | |
| 2a12   | Tel → Cen   | 310 33243-40784 | 1–9 | NM_133657 | |
| 2a21-21 | Tel → Cen | 308 17546-20076 | 3–9 | | Exons 3, 4, 8, 9 |
| 2a22   | Tel → Cen   | 308 1879-9066 | 1–9 | | |
| 2a23-23 | Tel → Cen | NW_011833 5827-6402 | 1–2 | | |
| 2a5    | Cen → Tel   | — | 1–9 | NM_007812 | All exons |
| 2a1    | Cen → Tel   | 307 783765-775666 | 1–9 | NM_013899 | Exons 6, 9 |
| 2b19   | Cen → Tel   | 307 711178-725201 | 1–9 | NM_007814 | |
| 2b24-24 | Tel → Cen | 307 692575-699876 | 7–9 | | |
| 2b23   | Tel → Cen   | 307 618973-640139 | 1–9 | | |
| 2a4    | Cen → Tel   | 307 266450-274118 | 1–9 | NM_009997 | |
| 2g1-1  | Tel → Cen   | — | 7–9 | | Exons 7–9 |
| 2b25-25 | Tel → Cen | 307 195792-195980 | 9 | | |
| 2b9    | Cen → Tel   | 307 144000-171613 | 1–9 | NM_010000 | Exon 1 |
| 2b26-26 | Tel → Cen | AC157 22100-22600 | 2–6 | | Exons 2, 3, 6 |
| 2b13   | Cen → Tel   | 307 1-32260 | 1–9 | NM_007813 | Exon 1 |
| 2b27-27 | Tel → Cen | 303 2127292-2130037 | 2–9 | | Exons 2–4, 9 |
| 2b28-28 | Tel → Cen | 303 2064442-2094900 | 1–6 | | |
| 2b10   | Cen → Tel   | 303 2012844-2040458 | 1–10 | NM_00998 | |
| 2s1    | Cen → Tel   | 303 1917585-1931046 | 1–9 | NM_028775 | |

Abbreviations: Cen, centromere; Tel, telomere.

*GenBank supercontigs NW_000303, NW_000307, and NW_000310 are abbreviated 303, 307, and 310. Contig AC087157 is abbreviated AC157. *Locations are base pair numbers within supercontigs that span the first to last codon of a locus. Several loci are found only in BAC clone sequence (GenBank no. AC087157) and not in supercontigs. *Gap column shows parts of the coding sequences not included in the GenBank sequence assemblies.
Southern Blot Analysis
Each BAC clone DNA was separately digested with the restriction enzymes \textit{Cla} I, \textit{Eco} RI, \textit{Hind} III, \textit{Pvu} II, \textit{Sac} I, and \textit{Xba} I at 37°C for 12–36 hr. The digests were electrophoresed on 1% agarose gels, and the DNA fragments were blotted onto Hybond (Boehringer Mannheim, Indianapolis, IN) nylon membranes. PCR amplification products were made into fluorescent probes using the Genius DIG-labeling system (Boehringer Mannheim) and hybridized to the blots at 55°C overnight, followed by rinsing and labeling according to the manufacturer’s protocol.

Restriction Mapping
Restriction maps of BAC clones RP22-78A19, -44B20, -362C15, -127H7, -548M4, -160O14, and RP23-430G14, -174D7, and -113D13 were constructed using \textit{Eco} RI and \textit{Hind} III (data not shown). These restriction maps were compared with the draft genomic sequences and with the sizes of the fragments that hybridized with gene-specific and BAC end probes to confirm the overall assembly. Independent \textit{Bst} 1107 I restriction maps of clones RP23-430G14, -174D7, -113D13, and -353B5 from the study of Kim et al. (2001) are available as supplemental data through the Lawrence Livermore National Laboratory (Livermore, CA) website (http://greengenes.llnl.gov/mouse/html/syn_table.htm). These maps cover the region from \textit{Cyp2a4} through \textit{Cyp2b10} (Figure 1) that was most poorly resolved by \textit{Eco} RI–\textit{Hind} III mapping and genomic sequencing. Though the online assembled maps (Kim et al. 2001) have large unresolved regions and a number of mistakes, the actual digest data are accurate. We have extracted and reassembled the digest data to form a new map that is consistent with our \textit{Eco} RI–\textit{Hind} III restriction maps and with the draft genomic sequences. This composite restriction map (Table 2 of Supplemental Material) is the basis for the distances between the \textit{Cyp2a4} through \textit{Cyp2b10} loci shown in Figure 1.

Analysis of Draft Genomic Sequences
Partial draft sequences of some BAC clones from the relevant region of mouse chromosome 7 are available on GenBank as of May 2003 (build 30). Some of the sequences from these clones are incorporated into the supercontig NT_039407, but this assembly contains many errors and should be disregarded. The older supercontigs NW_000303, NW_000307, and NW_000310 (build 29) are incomplete, but for the most part are correctly assembled. We created a more complete and accurate assembly of the draft sequences by

---

Figure 1. Organization of the \textit{Cyp2a-t} gene cluster on mouse chromosome 7. The total map has been broken in half to fit the page. The public sequences (GenBank nos. NW_000303, NW_000307, and NW_000310, build of 15 November 2002), Celera sequences (assembly GA_x8K6ZTzPUB9, as of 1 June 2002), and individual BAC clones from the RP22 and RP22 libraries (labeled by GenBank clone number) are aligned to the composite map produced by this study. More recent assemblies of the public data are not included because they are significantly less accurate. The two sequence assemblies are labeled with their own numbering systems in kilobases to clarify the comparisons. The locations of the \textit{Cyp2a-t} genes and their directions of transcription are indicated with broad arrows on the composite map. A total of 22 loci from six different subfamilies are identified; the distance spanned by each locus is exaggerated relative to the distances between loci for clarity. Large gaps within the assembled sequences are shown by dashed lines, whereas many smaller gaps are unmarked. A region of 300 kb at the break point in the middle of the cluster contains no \textit{Cyp} genes and is therefore deleted from the map. The \textit{Egln2} and \textit{Axl} genes flanking the cluster are shown by solid arrows on the composite map.
rearranging the sequence contigs inside each clone to conform to our maps and by comparing small overlapping portions of contigs from different clones. Predicted restriction enzyme cut sites in the database sequences were compared with actual restriction mapping fragment sizes and with the data from Southern blotting to determine the order of sequence contigs. An independent sequence assembly for this region obtained from Celera Genomics (GA_x6k02T2P9UB, as of 1 June 2002) was also used to fill in some gaps. Sequence comparisons were done using the NCBI BLAST and Jellyfish (LabVelocity Inc., San Francisco, CA) software.

Results

The mouse Cyp2a-t gene cluster forms a small part of a 30-Mb region of chromosome 7 that is syntenic to most of the q arm of human chromosome 19 (Kim et al. 2001). This entire region of synteny in the two species is in the opposite orientation relative to the centromere; the mouse map in Figure 1 is shown with the telomere to the left, the reverse of the conventional mode of display, so that it aligns with the established human map (Hoffman et al. 2001). The mouse Cyp2a-t cluster spans about 1.4 Mb, as measured by restriction mapping. It is delimited by the Egl2 gene on the telomeric side and the Axl gene on the centromeric side (Figure 1).

These genes are orthologs of the EGLN2 and AXL genes that bracket the human cluster. A total of 22 CYP loci from six subfamilies were found in the mouse gene cluster; 10 of the loci match previously sequenced mRNAs (Hoffman et al. 2001) and are therefore functional genes. Information on individual loci is compiled in Table 1, and the specific evidence used to localize and identify each locus is organized below by subfamily. The complete map of the region is shown in Figure 1. The relevant parts of the public and private sequence assemblies of mouse chromosome 7 (as of December 2002) are compared in Figure 1 with the composite map generated by this study. Both of these previous sequence assemblies are mostly accurate but incomplete across this region of the chromosome; a more recent assembly of the public data (build 30, February 2003) is markedly less accurate. Significant gaps occur in both assemblies. Gaps in the public contigs are quite accurately sized, but the sizes of several large gaps in the Celera assembly are seriously underestimated (Figure 1). Some apparent gaps in the public assembly can be filled, in fact, by integrating sequences from the draft versions of BAC clones RP23-430G14, -174D7, and -113D13 (GenBank accession nos. AC087157, AC087137, and AC087130) and from the small assembled contigs NW_000304, NW_000305, NW_000306, NW_000308, NW_000309, and NW_011833. Two distinct regions of about 50 kb each, which contain the 2b26-ps and 2b27-ps pseudogenes, respectively (Figure 1), are incorrectly merged by both assemblies because of the very high level of sequence similarity between them. The presence of both of these regions on the chromosome was confirmed by restriction mapping and by specific PCR amplifications of fragments that bridge small gaps in the draft sequences. Additional details of experimental methods and results, including tables of the primers used, exact exon/intron boundaries, and restriction map data, are available in Tables 1–4 of the Supplemental Material and through the Cytochrome P450 Homepage (Nelson 2003).

Descriptions of Loci by Subfamily

Cyp2a. Three functional mouse Cyp2a genes, 2a4, 2a5, and 2a12, were previously identified from mRNAs (Iwasaki et al. 1993). Our analysis has discovered a new full-length 2a locus, located between the 2a5 and 2a12 genes, that corresponds to a single mouse expressed sequence tag (EST) in the database (GenBank accession no. BB667610) and is therefore likely to be functional. It has been given the name Cyp2a22. There are also three partial 2a pseudogenes—Cyp2a20-ps and Cyp2a23-ps, each of which consists only of exons 1 and 2, and Cyp2a21-ps, which has part of exon 3 and all of exons 4–9. To search for additional 2a loci, PCR-amplified fragments from exons 2, 6, and 9 of 2a5 were used to probe Southern blots of the BAC clone DNAs. They hybridized to the expected fragments for all genes, which collectively accounted for all positive signals. Specific primers were used to amplify and sequence the sixth exons of 2a4, 2a5, and 2a12 and the third exons of 2a12 and 2a22 from the appropriate BAC clone DNAs (Figure 1) to confirm the locations and identities of these genes. The 2a20-ps pseudogene and the 2a12 gene were identified only from the genomic sequence assemblies, as they were not included in any BAC clones used in this study.

The 2a4 and 2a5 genes are extremely similar (98% exons/96% introns), even though they are not physically close together (Figure 2). The 2a12 locus is strongly related by sequence to 2a22 but is quite different from the other genes, with only 75 and 76% exon identities to 2a4 and 2a5, respectively (Table 2). The 2a20-ps and 2a23-ps pseudogenes are very similar to each other and are slightly more similar to the 2a5 locus than to 2a12. The Cyp2a21-ps pseudogene is most similar to 2a5 (Table 2), Cyp2b. The 2b subfamily is the most diverse group within the gene cluster. Four genes previously known to be functional (Nelson et al. 1996)—2b9, 2b10, 2b13, and 2b19—have been identified and localized. A fifth gene previously identified as functional, 2b20 (Damon et al. 1996), and its putative pseudogene 2b20-ps (Marc et al. 1999) are not found in the chromosome 7 gene cluster. Our repeated attempts to amplify a 2b20-specific fragment with the PCR primers designed by Damon et al. (1996), using both BAC clone and mouse genomic templates, were unsuccessful. We therefore conclude that the 2b20 and 2b20-ps transcripts are artifacts of the very similar 2b10 gene. In addition, the 2b10 gene in both sequence assemblies differs markedly (19 base pair substitutions) from the originally reported 2b10 mRNA (Noshiro et al. 1988); this may be because of interstrain heterogeneity or to mistakes in sequencing. We now consider all 2b10, 2b20, and 2b20-ps mRNAs in the database to be products of the single gene identified as 2b10 in Figure 1.

To determine the location of each 2b locus, PCR products generated from exons 1, 7, and 9 of 2b9 were used to probe the BAC clone Southern blots. They hybridized to the appropriate DNA fragments from each of the 2b genes and pseudogenes. Specific probes were also made for exon 4 of 2b13 and exon 3 of 2b19 that hybridized uniquely to fragments of those genes on blots. Primers specific for exon 2 of 2b10 were used to amplify and sequence a fragment of BAC clone RP23-113D13 to confirm the identity of this locus. Specific primers were also used to amplify and sequence fragments of intron 2 and intron 3 from the nearly identical 2b26-ps and 2b27-ps pseudogenes. These fragments were then used as probes on the blots to prove the separate existence of the two pseudogenes. PCR products encoding intron 2 of 2b26-ps/2b27-ps hybridized to Eco RI fragments of 10.6 and 9.5 kb, respectively, in the BAC clones RP23-430G14 and -113D13, and to both fragments in the BAC clone RP23-174D7, which overlaps both pseudogenes (Figure 1).

Five of the 2b loci can confidently be identified as pseudogenes because they consist of less than the nine exons common to functional Cyp2b genes (Table 1). The partial nature of the 2b24-ps, 2b25-ps, 2b26-ps, 2b27-ps, and 2b28-ps pseudogenes was confirmed by exon-specific PCRs from the appropriate BAC clones. The locus labeled 2b23 in Figure 1 has not been previously
identified as a functional gene, but it has all nine exons, includes a legitimate heme signature in the ninth exon, and has no premature stop codons or frameshift mutations. Differences between the public and Celera versions of this sequence yield a few alternative amino acid residues but do not affect the viability of any potential product. The 2b23 sequence does not match any mRNAs or ESTs currently listed in GenBank, so it must be listed as a potentially functional new 2b gene, pending a search through more tissue types for a matching mRNA. There are thus a total of five confirmed or potentially functional genes and five partial

Figure 2. (A) Postulated evolution of the Cyp2a-t/CYP2A-T clusters in mouse and human. The modern arrangements are shown in the middle of the figure, with the separate evolutionary paths of the two clusters converging, as indicated by the vertical arrows. The ends of the clusters are very similar, but the inverted duplication in the human CYP2A-T cluster (adapted from Hoffman et al. 2001) is not present in mouse. Instead, a tandem duplication of the central Cyp2a, 2g, and perhaps 2b loci, without an inversion, established the organization of the mouse cluster. The telomeric Cyp2a group and the centromeric Cyp2b group in mouse (gray boxes) were probably formed by series of smaller duplications, as shown in B and C. Straight dashed arrows indicate direct duplications, and bent arrows indicate inverted duplications. The large vertical arrows indicate evolutionary time. Mouse pseudogenes are labeled with the suffix “ps” rather than “p” because of space restrictions. (B) Detailed diagram of telomeric mouse 2a gene group showing the hypothesized tandem and inverted duplications that formed the three genes and three pseudogenes in this group. The open boxes show the extent of each duplicated block of DNA. Straight arrows indicate direct duplications, and bent arrows indicate inverted duplications. (C) Detailed diagram of centromeric mouse 2b gene group showing the hypothesized series of duplications that formed this group. The open boxes show the extent of each duplicated block of DNA. Straight arrows indicate direct duplications, and bent arrows indicate inverted duplications.
pseudogenes within the 2b subfamily in mouse.

The rule that functional genes in the Cyp2 family have a nine-exon structure is violated in the mouse by the 2b10 gene. Two cDNA sequences were originally described for this gene (Noshiro et al. 1988), one with a standard length of 1,476 base pairs, and a second rare form with a stretch of 27 extra nucleotides that were presumed to belong either to the end of exon 8 or the beginning of exon 9. However, our analysis of the genomic sequence of 2b10 makes it clear that these base pairs in fact make up a small additional portion of the eighth intron of the ancestral 2b gene. We have identified sequences in the eighth introns of 2b9 and 2b13 genes that are very similar to the minixeon, but both of these introns have critical differences that prevent the formation of splice sites (Figure 3). Because the nine additional amino acids would disrupt a critical motif in the enzyme, it is likely that the long form of 2b10 represents a nonfunctional splice variant.

Evolutionary relationships among the paralogous mouse 2b loci are far from clear. The 2b9 and 2b13 genes are more highly related to each other than to either 2b10 or 2b19, and the new 2b23 gene is somewhat more similar to 2b19 than to the other genes. Any other paralogous genes to the functional 2b genes gives the same average identity level of about 85% across exons (Table 2). The 2b pseudogenes are also not closely related to specific functional genes except for the 2b28-ps partial pseudogene, which is somewhat more similar to 2b13 and 2b9 than to the other 2b genes (Table 2). As noted above, the 2b26-ps and 2b27-ps pseudogenes are nearly identical, but they do not show a particular affiliation to any of the functional genes.

Cyp2a. There is only a single member of the 2fa subfamily in the mouse, which is located centromeric of and close to the 2fa gene (Figure 1) in a position exactly corresponding to that of the human 2F1P locus. Unlike the human and gorilla (Chen et al. 2002), the mouse does not have a second 2fa locus. This was established by the failure of intron 1 and exon 9 primers to amplify from any of the BAC clones and by the lack of hybridization to the Southern blots using a 2fa exon 9 probe.

Cyp2g. The 2g subfamily consists of the gene responsible for the known CYP2G1 enzyme, located just centromeric of the 2a5 gene, and the partial pseudogene 2g1-ps, which lies just centromeric of the 2a4 gene (Figure 1). Cyp2g1-ps has only exons 7, 9, and half of 8, which collectively are about 96% identical to the corresponding portions of 2g1 (Table 2). As both sequence assemblies are very fragmented near 2g1-ps, the pseudogene sequence can currently be found only in the Celera assembly, and even there it is incomplete. Because 2g1 is also incomplete in the public assembly (Table 1), its identity was confirmed by PCR-amplifying and sequencing fragments of exons 1, 2, 6, and 9 from the BAC clone RP22-78A19. To prove the partial nature of the 2g1-ps pseudogene, the RP23-430G14 BAC clone was used as template for the same amplifications; only the exon 9 primers gave a product. In addition, the exon 1 and 6 products hybridized only to clone RP22-78A19 on the Southern blots and not to RP23-430G14.

Cyp2b. As is true for the human, the mouse has a single member of the 2b subfamily located at one end of the cluster, close to the Axl gene (Figure 1). Neither Southern blots nor PCR amplifications gave evidence of any additional 2b loci.

| Locus | 2a5 | 2a12 | 2a20-ps (Exons 1, 2) | CYP2A6<sup>+</sup> |
|-------|-----|------|---------------------|-------------------|
| 2a4   | 98% (1,485) | 75% (1,491) | 72% (345) | 83% (1,485) |
| 2a5   | — | 76% (1,485) | 73% (334) | 84% (1,485) |
| 2a12  | — | — | 72% (182) | 75% (1,485) |
| 2a22  | — | 96% (1,506) | — | — |
| 2a21-ps | 97% (1,066) | — | — | — |
| 2a22-ps | 97% (576) | — | — | — |

| Locus | 2b10 | 2b13 | 2b19 | 2b22 |
|-------|------|------|------|------|
| 2b9   | 96% (1,503) | 92% (1,473) | 84% (1,479) | 86% (1,476) |
| 2b10  | — | 85% (1,503) | 85% (1,506) | 87% (1,476) |
| 2b13  | — | — | 84% (1,479) | 86% (1,462) |
| 2b19  | — | — | — | 88% (1,425) |

| Locus | 2b24-ps (Exons 7–9) | 2b25-ps (Exon 9) | 2b26-ps (Exons 4, 5) | 2b27-ps (Exons 5–8) |
|-------|---------------------|-----------------|---------------------|---------------------|
| 2b9   | 87% (516) | 80% (188) | 84% (333) | 76% (651) |
| 2b10  | 86% (509) | 82% (184) | 85% (334) | 78% (656) |
| 2b13  | 86% (505) | 81% (186) | 80% (334) | 76% (655) |
| 2b19  | 85% (470) | 80% (183) | 84% (335) | 78% (652) |
| 2b23  | 85% (505) | 85% (182) | 82% (334) | 81% (505) |

| Locus | 2b28-ps (Exons 1–4) | CYP2B6<sup>+</sup> |
|-------|---------------------|-------------------|
| 2b9   | 92% (684) | 75% (1,476) |
| 2b10  | 83% (649) | 76% (1,503) |
| 2b13  | 94% (656) | 75% (1,476) |
| 2b19  | 82% (632) | 77% (1,479) |
| 2b23  | 82% (688) | 77% (1,470) |

*Numbers shown are percent identical nucleotides and total nucleotides compared between coding sequences (in parentheses). CYP2A6 and CYP2B6 are representative genes from the same subfamilies in humans.

Figure 3. The structure of the 2b10 gene. There is a minixeon of 27 nucleotides, labeled “10,” between the standard exons 8 and 9. The 2b10 minixeon and surrounding sequences are compared with the corresponding regions from the eighth introns of 2b9 and 2b13. Potential splice sites are underlined. Exon sizes are exaggerated relative to intron sizes for clarity.
cDNA or EST now in GenBank. The gene is located at the extreme telomeric end of the mouse cluster, only 8 kb from the Egln2 locus (Figure 1), in the same relative position as the human pseudogene CYP2F. The mouse 2r4 is slightly more related to human 2T2P than to human 2T3P (Table 2).

Discussion

The Cyp2 subfamilies in the mouse cluster are the same six present in the corresponding human cluster (CYP2A2, 2B, 2F, 2G, 2S, and 2T), but the total number of loci is significantly greater in mouse (22 vs. 13). This difference is due primarily to the expansion of the 2a and 2b subfamilies in mouse (Figure 2). The similarities between the CYP2A-T gene clusters in the mouse and human indicate that the six component subfamilies were already present in a common mammalian ancestor. The differences in organization indicate that most of the individual loci within the subfamilies developed after the primate and rodent lineages split. Some specific loci, however, may have developed in the common ancestor, and therefore may be truly orthologous. To trace the evolution of the gene clusters in any detail, it is necessary to distinguish these older orthologous loci from newer, species-specific loci. Defining orthologs between mouse and human also facilitates the creation of appropriate knockout animals.

It is often difficult or impossible to identify orthologs of CYP genes in all but very closely related species (Nelson et al. 1993), but when sequence similarity, physical location, and protein function all match, this can be done at least tentatively (Chen et al. 2002; Hoffman et al. 2001). In the case of the CYP2A-T clusters, some orthologous relationships can be reasonably deduced. The mouse Cyp2a5 locus may be a true ortholog of the human CYP2A6, as they have the most similar sequences (Table 2) and they are located at similar positions within the two clusters (Figure 2A). The single mouse 2J2 and the functional human 2F1 both express proteins with similar substrate ranges and the same limited tissue distribution and are thus considered orthologous (Chen et al. 2002). The mRNAs known from the single mouse 2J1 and human 2S1 genes are 81% identical, and these genes are similarly located on the Axl ends of their respective clusters, so they can also be considered orthologous (Hoffman et al. 2001). The mouse 2g1 is equally similar to the two human 2G pseudogenes (83%), which are probably degraded copies of an earlier functional gene that was orthologous to 2g1. Finally, the presumably functional mouse 2r4 has the same position on the EGLN2 end of the cluster as does the human 2T2P and is likely to be its ortholog (Nelson et al. 2003).

Though the two ends of the mouse and human clusters are very similar, with orthologous genes in corresponding positions, the distinct organization in the middle of each cluster indicates that some major rearrangements have occurred since the two species diverged from a common ancestor. In the human, a large inverted repeat was inferred to explain the mirror-image organization underlying the paired 2F, 2T, 2A, and 2G loci in the center of the cluster (Hoffman et al. 1995, 2001). In the mouse, there is no such mirror-image set of loci. Instead, there are single 2f and 2t genes, while the central loci are arranged a-g-b-a-g-b rather than a-g-b-b-g-a as in humans, suggesting a more limited tandem duplication without an inversion (Figure 2A). Additionally, in humans the 2B6 and 2B7P loci were apparently inserted into the middle of the 2A18P locus (Hoffman et al. 2001), whereas in mouse the many 2b subfamily genes occur in two separate groups, with no sign of a late insertion.

Figure 2A illustrates our hypothesis that the basic organization of the central loci in mouse is due to a large tandem duplication encompassing 2a, 2g, and perhaps 2b loci. This hypothesis is supported by the fact that the 2aS and 2g1 genes on the telomeric side of the cluster are transcribed in the same direction and are spaced a similar distance apart as are the 2aA and 2g1-ps loci on the centromeric side. It is also consistent with the suggestion of Aida et al. (1994) that because Mus musculus has both 2aA and 2aS genes, whereas its close relative Mus spretus appears to have two nearly identical 2aS-like loci, 2aA must have been formed recently by the alteration of a critical residue in a previously duplicated copy of the 2aS gene.

This large tandem duplication may or may not have included loci from the 2b subfamily. Though 2aV is very similar in sequence to 2aS, and 2g1 to 2g1-ps, there are no highly related 2b genes across the two groups, as would be expected if an a-g-b block of sequence had been recently duplicated. In addition, the 2b loci are not as clearly patterned as the 2a and 2g loci—multiple 2b genes are transcribed in different directions next to both the 2g1 and 2g1-ps loci. Conversely, the 2b loci do occur in two distinct groups that are in similar positions relative to the 2a and 2g loci. Although there is thus good evidence for a tandem duplication that included at least one 2a and one 2g locus, the timing and the extent of this duplication cannot be determined until the corresponding genes are examined in additional mammalian species.

Full or partial deletions of single loci may have occurred in both the primate and rodent lineages, but these cannot be detected. Gene losses by deletion should always be harder to distinguish than duplications, as they are unlikely to leave behind any characteristic pattern or signature sequences. In particular, partial pseudogenes in both clusters may have been created either by a duplication followed by deletion of some exons or by a duplication encompassing only part of a locus.

On a smaller scale, several interesting comparisons can be made among the numerous 2a and 2b subfamily genes. The 2a subfamily expanded in the mouse by a series of duplications involving single and multiple loci. The group of six 2a loci on the telomeric side of the cluster includes three highly related pairs (Table 2). The whole 35-kb block of DNA that includes Cyp2a22 and 2a23-ps is highly similar to the block containing Cyp2a12 and 2a20-ps, with more than 90% identity between the regions around the genes. The 20-kb region that encompasses Cyp2a5 shares more than 90% sequence identity with the region around Cyp2a21-ps. The simplest explanation for this arrangement, shown in Figure 2B, requires several rounds of duplication. The exact order in which the duplications happened is ambiguous, as there is no significant patterning of sequence in between the duplicated blocks.

The two strongest similarities within the 2b subfamily are between the 2b9 and 2b13 loci and between the 2b26-ps and 2b27-ps pseudogenes (Table 2). The relative positions and the directions of transcription of these gene pairs suggest that a second, smaller tandem duplication occurred within the centromeric 2b group to create the 2b9–2b26-ps and 2b13–2b27-ps regions, as shown in Figure 2C. Evidence for this duplication also comes from the extremely high level of identity (99%) found between short non-coding sequences in the introns of the 2b26-ps and 2b27-ps pseudogenes (data not shown).

The information provided by this study has allowed us to draw a complete and accurate picture of the Cyp2a-t gene cluster in the mouse and to understand the similarities and differences between its evolution and that of the human cluster. This comparison should enable researchers to better utilize the mouse as a model system for the study of these CYP genes in humans and in other mammals.
REFERENCES

Aida K, Moore R, Negishi M. 1994. Lack of the steroid 15α-hydroxylase gene (Cyp2a-4) in wild mouse strain Mus spretus: rapid evolution of the P450 gene superfamily. Genomics 19:564–566.

Chen N, Whitehead SE, Caillat AW, Gavit K, Ishphording DR, Kovacevic D, et al. 2002. Identification and cross-species comparisons of CYP2F subfamily genes in mammals. Mutat Res 499:151–161.

Damon M, Fautrel A, Marc N, Guillouzo A, Corcos L. 1996. Isolation of a new mouse cDNA clone: hybrid form of cytochrome p450 2b10 and NAPDH-cytochrome p450 oxidoreductase. Biochem Biophys Res Commun 226:900–905.

Gonzalez F, Nebert D. 1990. Evolution of the P450 gene superfamily: animal-plant “warfare,” molecular drive and human genetic differences in drug oxidation. Trend Genet 6:182–186.

Hoffman S, Fernandez-Salguero P, Gonzalez F, Mohrenweiser H. 1995. Organization and evolution of the cytochrome P450 CYP2A-2B-2F subfamily gene cluster on human chromosome 19. J Mol Evol 41:894–900.

Iwasaki M, Lindberg R, Juvonen R, Negishi M. 1993. Site-directed mutagenesis of mouse steroid 7α-hydroxylase (cytochrome P-450 7α): role of residue-209 in determining steroid-cytochrome P-450 interaction. Biochem J 291:569–573.

Iwatsuki M, Lindberg R, Juvonen R, Negishi M. 1993. Site-directed mutagenesis of mouse steroid 7α-hydroxylase (cytochrome P-450 7α): role of residue-209 in determining steroid-cytochrome P-450 interaction. Biochem J 291:569–573.

Lewis DF, Watson E, Lake BG. 1998. Evolution of the cytochrome P450 superfamily: sequence alignments and pharmacogenetics. Mutat Res 410:245–270.

Marc N, Damon M, Fautrel A, Guillouzo A, Corcos L. 1999. Isolation of a cyp2b10-like cDNA and of a clone derived from a cyp2b10-like pseudogene. Biochem Biophys Res Commun 258:11–16.

Nebert DW, Russell DW. 2002. Clinical importance of the cytochromes P450. Lancet 360:1155–1162.

Nelson D. 2003. Cytochrome P450 Homepage. Memphis, TN: University of Tennessee. Available: http://drnelson.utmem.edu/CytochromeP450.html [accessed 19 June 2003].

Nelson D, Kamataki T, Waxman D, Guengerich F, Estebrook R, Feyereisen R, et al. 1993. The P450 superfamily: update on new sequences, gene mapping, accession numbers, early trivial names of enzymes, and nomenclature. DNA Cell Biol 12:1–51.

Nelson D, Koymans L, Kamataki T, Stegean J, Feyereisen R, Waxman D, et al. 1996. Cytochrome P450 superfamily: update on new sequences, gene mapping, accession numbers, and nomenclature. Pharmacogenetics 6:1–42.

Noshiro M, Lakso M, Kawaji H, Negishi M. 1988. Rip locus: regulation of female-specific isozyme (I-P-450 16α) of testosterone 16α-hydroxylase in mouse liver, chromosome localization, and cloning of P-450 cDNA. Biochemistry 27:6434–6443.

Waterson RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.