Comparison of cytochrome P450 (*CYP*) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants

David R. Nelson[a], Darryl C. Zeldin[b], Susan M.G. Hoffman[c], Lois J. Maltais[d], Hester M. Wain[e] and Daniel W. Nebert[f]

Objectives **Completion of both the mouse and human genome sequences in the private and public sectors has prompted comparison between the two species at multiple levels. This review summarizes the cytochrome P450 (*CYP*) gene superfamily. For the first time, we have the ability to compare complete sets of *CYP* genes from two mammals. Use of the mouse as a model mammal, and as a surrogate for human biology, assumes reasonable similarity between the two. It is therefore of interest to catalog the genetic similarities and differences, and to clarify the limits of extrapolation from mouse to human.**

Methods **Data-mining methods have been used to find all the mouse and human *CYP* sequences; this includes 102 putatively functional genes and 88 pseudogenes in the mouse, and 57 putatively functional genes and 58 pseudogenes in the human. Comparison is made between all these genes, especially the seven main *CYP* gene clusters.**

Results and conclusions **The seven *CYP* clusters are greatly expanded in the mouse with 72 functional genes versus only 27 in the human, while many pseudogenes are present; presumably this phenomenon will be seen in many other gene superfamily clusters. Complete identification of all pseudogene sequences is likely to be clinically important, because some of these highly similar exons can interfere with PCR-based genotyping assays. A naming procedure for each of four categories of *CYP* pseudogenes is proposed, and we encourage various gene nomenclature committees to consider seriously the adoption and application of this pseudogene nomenclature system. *Pharmacogenetics* 14:1–18 © 2004 Lippincott Williams & Wilkins**

Pharmacogenetics 2004, **14**:1–18

Keywords: cytochrome P450, nomenclature, pseudogenes, mouse, human

[a]Department of Molecular Sciences, University of Tennessee, Memphis TN 38163, USA and The UT Center of Excellence in Genomics and Bioinformatics, [b]Division of Intramural Research, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC 27709, USA, [c]Department of Zoology, Miami University, Oxford, OH 45056, USA, [d]Mouse Genomic Nomenclature Committee (MGNC), The Jackson Laboratory, 600 Main St., Bar Harbor, ME 04609, USA, [e]HUGO Gene Nomenclature Committee, Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK and [f]Department of Environmental Health and Center for Environmental Genetics, University of Cincinnati Medical Center, P.O. Box 670056, Cincinnati OH 45267–0056, USA.

This work was supported in part by NIH Grant P30 ES06096 (D.W.N.) and the NIEHS Division of Intramural Research (D.C.Z.).

Correspondence to Dr David R. Nelson, Department of Molecular Sciences, University of Tennessee, Memphis TN 38163, USA.
Tel: +1 901–448–8303; fax: +1 901–448–7360; email dnelson@utmem.edu

Received 19 October 2003
Accepted 7 November 2003

Introduction

The mammalian cytochrome P450 (*CYP*) superfamily encodes enzymes involved in: the metabolism of pharmaceuticals, foreign chemicals and pollutants; arachidonic acid metabolism and eicosanoid biosynthesis; cholesterol, sterol and bile acid biosynthesis; steroid synthesis and catabolism; vitamin D_3 synthesis and catabolism; retinoic acid hydroxylation; biogenic amine and neuroamine metabolism; and orphan CYPs of unknown function [1]. What once had been described as predominantly a 'hepatic drug detoxication system' is now known to include a myriad of enzymatic reactions involved in critically important life processes. Consequently, mutations in a number of *CYP* genes are responsible for inborn errors of metabolism and contribute to several important clinically relevant diseases.

April 2003 brought the 50th anniversary of Watson and Crick's famous paper; the Human Genome Project (HGP) was declared complete, and a new assembly (build 33) of the genome was released. This assembly has 545 contigs, meaning that there are only about 500 gaps remaining. NCBI also released a new mouse genome assembly in February 2003 (build 30), with 37 998 contigs. Although the mouse genome is not as complete, the state of the mouse *Cyp* genes is very nearly finalized, with only a few unresolved regions. Therefore, a comprehensive comparison of the *CYP* genes between the two mammalian genomes is now possible for the first time. This review provides a detailed analysis of the *CYP* genes and pseudogenes in mouse and human, going beyond the brief treatment in the initial Mouse Genome Sequencing Consortium

0960-314X © 2004 Lippincott Williams & Wilkins

DOI: 10.1097/01.fpc.0000054151.92680.31

report [2]. Herein we have assigned names to all putatively functional *CYP* genes and pseudogenes, in the continuing effort to provide a complete annotation of the *CYP* gene superfamily [3]. In this review, we also propose a standardized nomenclature system for naming the four types of pseudogenes; this nomenclature is not officially sanctioned by any nomenclature body, but we hope our views will contribute to the discussions of pseudogene nomenclature that are now evolving within these groups.

Before complete genomes were available, gene nomenclature was simpler because detailed information about gene position on the chromosome did not affect the gene names. In the past, names were assigned based only on sequence similarity and evolutionary divergence [4,5]. Now there are several types of nomenclature issues that have arisen due to positional information. Genes often occur in clusters, with several related genes, pseudogenes and detritus exons aligned in tandem. Genes in subfamilies are sometimes clustered with genes of other subfamilies. Now that the clusters are known in detail, there is some value to naming all these *CYP* sequences in the order in which they appear in each cluster. With the mouse and human genomes, this is not really possible, because most *CYP* genes have previously been named from cDNAs, thereby resulting in the gene names within a cluster being out of sequence. It is possible to name the remaining genes and pseudogenes in other species, however, in order of their location; this positional naming has been done, for example, with most of the 105 mosquito (*Anopheles gambiae*) *CYP* genes [6]. For a compilation of the numbers of *CYP* genes in all eukaryotic genomes sequenced to date, see Table 1.

Methods

A data-mining strategy was devised to find all *CYP*-related sequences in the mouse and human genomes.

Table 1 Numbers of putatively functional full-length *CYP* genes in whole eukaryotic genomes[a]

Homo sapiens (human)	57	Apr. 2003[b]
Mus musculus (mouse)	102	Feb. 2003 build 30
Canis familiaris (dog)	54	Sep. 2003
Takifugu rubripes (pufferfish)	54	Aug. 2002 v.3.0
Caenorhabditis elegans (nematode)	74	Dec. 1998
Drosophila melanogaster (fruit fly)	84	Mar. 2000
Anopheles gambiae (mosquito)	105	Oct. 2002
Ciona intestinalis (sea squirt)	80	Dec. 2002
Ciona savignyi (sea squirt)	97	Apr. 2003 release 1
Dictyostelium discoideum (slime mold)	42	Apr. 2003
Arabidopsis thaliana (thale cress)	249	Dec. 2000
Oryza sativa (rice)	323	Apr. 2002
Neurospora crassa (fungus)	38	Apr. 2003
Saccharomyces cerevisiae (baker's yeast)	3	Oct. 1996
Schizosaccharomyces pombe (fission yeast)	2	Feb. 2002

[a]Includes partials that are expected to be functional (e.g. *CYP1A1* in *Takifugu rubripes*).
[b]Dates are for major genome publications or most recent genome assembly or data release.

This method used representative mouse and human *CYP* sequences to search each entire genome, as described in Nelson (2002) [7]. One *CYP* sequence from each of the 18 mammalian *CYP* gene families was used to perform BLAST searches against the NCBI genome assemblies for mouse and human. In cases where there were several distant subfamilies in a family, additional searches were performed so as not to miss any *CYP*-related sequences. These extra searches included members of the *CYP2D*, *CYP4F*, *CYP4V*, *CYP11B*, *CYP26B*, *CYP26C*, *CYP27B* and *CYP27C* subfamilies. Each BLAST 'hit' was searched against a collection of all known mouse or human *CYP* sequences in the blast server at http://132.192.64.52/p450.html. Any new sequences were added to the database file on this server, and the search was continued.

Gene clusters were mined in a more systematic way. Because some genes are very similar to other members in a cluster, care must be taken not to confuse nearly identical sequences (e.g. *CYP4A11* and *CYP4A22* genes show 94% identity). The DNA sequence from each gene cluster was retrieved from the Map Viewer at NCBI in 100 000-bp fragments. For example, the 31 Mb contig NT_011109.13 from human chromosome 19, was partially downloaded as five segments, covering the sequence from 3.4 Mb to 3.9 Mb; these were then placed in the Do-It-Yourself WU-Blast server at http://www.proweb.org/proweb/Tools/WU-blast.html and searched with each of the nine exons of a *CYP2* family gene. This method, which required 45 searches to cover the area, identified all *CYP* fragments, including pseudogene fragments (109 exons, in this case). These were then assembled, without being confused with nearly identical fragments from other related genes, in an order that correctly reconstructs the physical map of this region [8]. Searches of the mouse *Cyp2c* and *Cyp2j* subfamilies were also conducted, using the Celera Discovery System (assembly R26, D. Zeldin, unpublished data). Because some regions of both genomes are not completely sequenced, there are a few areas that might require slight revisions in the future, such as the mouse *Cyp3a* cluster and the human *CYP4X–CYP4Z* region.

Established rules for *CYP* gene nomenclature have been followed for all new mouse and human intact *CYP* genes that were discovered. Based on natural clusters of sequences on phylogenetic trees, values of 40% and 55% amino acid identity were chosen as cutoffs for membership in *CYP* families and subfamilies respectively [4,5]. The original rule of 40% identity or higher for membership in a *CYP* gene family is still a useful guideline; other considerations about how sequences are clustered on phylogenetic trees, however, have allowed this cut-off to fall well below 40% for some families, especially the *CYP4* family, as discussed

below. In this review, we propose a new method for the naming of pseudogenes and alternative transcripts, based in part on their alignment along a chromosome. This method will be used in the *CYP* database. It is our hope that this nomenclature system will be influential among other nomenclature committees, and be applied to other gene families.

Above the level of family names, a higher-order category for *CYP* genes has been in use for several years; clusters of related *CYP* families are called *clans* [9,10]. There are nine named clans in vertebrates: the *CYP2* clan, the *CYP3* clan, the *CYP4* clan, the *CYP*7 clan, the *CYP19* clan, the *CYP20* clan, the *CYP26* clan, the *CYP51* clan and the mitochondrial clan. Single sequences that do not cluster with other sequences in a reproducible way (e.g. *CYP39A1*, *CYP46A1*) are not yet placed in a named clan. This may be revised in the future, as it is desirable to place all *CYP* genes in a named clan. A phylogenetic tree showing clans for vertebrate (human and pufferfish (*Takifugu rubripes*)) *CYP* genes has been recently published [11].

A composite phylogenetic tree of the *CYP* superfamily genes in mouse and human (Fig. 1) was generated from an alignment, initially constructed using CLUSTAL W, and edited manually. The PHYLIP package was used to make a distance matrix from the alignment with PROTDIST, and a tree was computed with NEIGHBOR, set for unweighted-pair-group method of averaging. The tree was then drawn with NJ PLOT and edited in Adobe Illustrator 9.0.

It should be emphasized which genome build numbers were used for this comparison. The genome assemblies are based on freezes in the sequence data that have been used in the assembly process. The most recent assembly of the human genome is build 33 (April 10, 2003, with 545 contigs). This build represents the data that can be seen in the genome Map Viewer. The most recent assembly of the mouse genome is build 30 (Jan. 27, 2003, with 37 998 contigs). New builds often change the nucleotide numbering of gene locations in the assembly contigs; this may or may not result in any improvements in the sequences of the *CYP* genes.

The contigs illustrated in this report (Figs 2, 3, 4 and 5) originate from mouse build 30, and human builds 30 and 33. Even build 33 of human is not yet complete, meaning that there are still some gaps involving *CYP* genes. This is true for the *CYP4ABXZ* cluster in humans, affecting the *CYP4X1* and *CYP4Z1* genes. We had assembled the human *CYP4X1* gene, based originally on mRNA from the orthologous rat *CYP4X1* gene. The complete CYP4Z1 mRNA sequence (GenBank AY262056) from human was deposited in Genbank on April 9, 2003. The human CYP4X1 mRNA sequence (BC028102) was deposited on April 22, 2003, making it the last human *CYP* transcript or gene to be completely sequenced. The mouse genome is less complete than the human genome, as can be seen in the mouse *Cyp2abfgst* cluster, the *Cyp2c* cluster, and the *Cyp3a* cluster. Some revision of the mouse *Cyp* sequences may occur in these regions in the future.

Importance of detecting pseudogenes

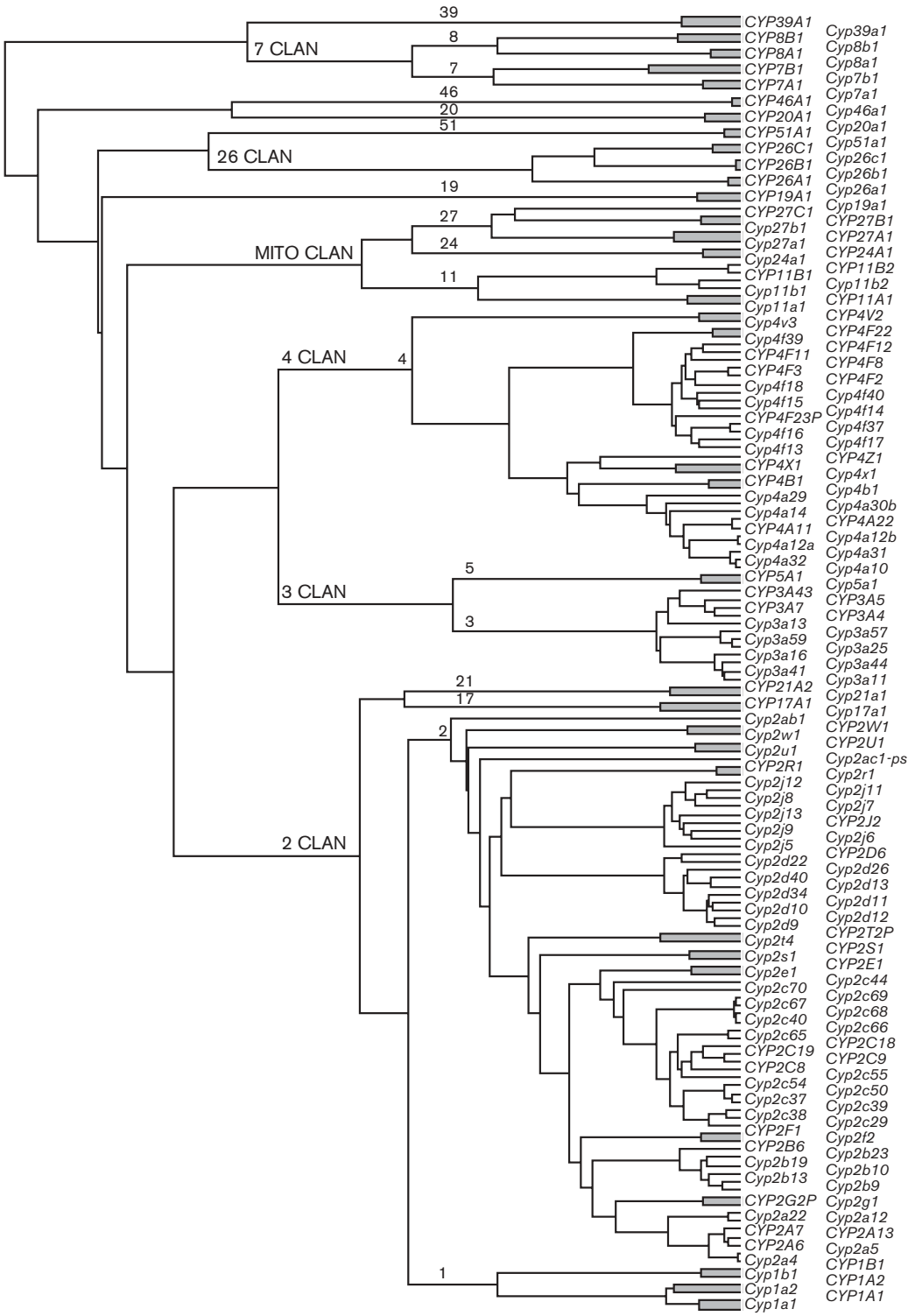
The absolute number of genes in the human genome, and mammalian genomes in general, has been debated extensively. Most recently, the number has been closing in on about 40 000 [12]. Though conservative estimates place it as low as 25 622 genes [13]. These genes will be annotated within the mouse and human genomes, and chromosomal location and intron–exon structures identified. The exon sequences only cover about 1.4% of the genome length [14]. The intergenic portions of the genome are largely without annotation, except to indicate the presence of numerous LINES, SINES, Alu sequences, and other repeats. Scattered within this mix are pseudogenes – genes that have suffered a defeat and are now fading away, or smaller fragments of gene sequences that have separated from their parent loci but never represented a functional gene.

Estimates predict about 20 000 pseudogenes in the human genome [15]. For example, the human genome appears to have more than 2000 ribosomal protein pseudogenes [16]. Since pseudogenes are not conserved by natural selection, they have a finite lifespan that is short, compared with that of their parent genes. Pseudogenes must be continuously generated; otherwise, we would not see them at all. Comparison of the human and chimpanzee genomes, which diverged at least 6 million years ago (MYA) [17], or the mouse and rat genomes, which diverged about 20 MYA [18], will soon give us better estimates as to how rapidly pseudogenes lose detectable similarity to a parent gene.

CYP genes are polymorphic, and variations can be relevant to drug metabolism and disease susceptibility [1,19]. Clinical diagnostic tests based on genotyping human *CYP* polymorphisms by PCR methods must take into account all the related genes, detritus exons, and pseudogenes, so that specific primer pairs can be synthesized. Lack of knowledge about highly similar pseudogenes and detritus exons can interfere with an accurate PCR-based clinical genotype assay. The difficulty of genotyping in the *CYP2D6* locus is an excellent example of such a problem [20]. Similar arguments will undoubtedly also apply to *CYP* genes in other species, as well as to many non-*CYP* genes.

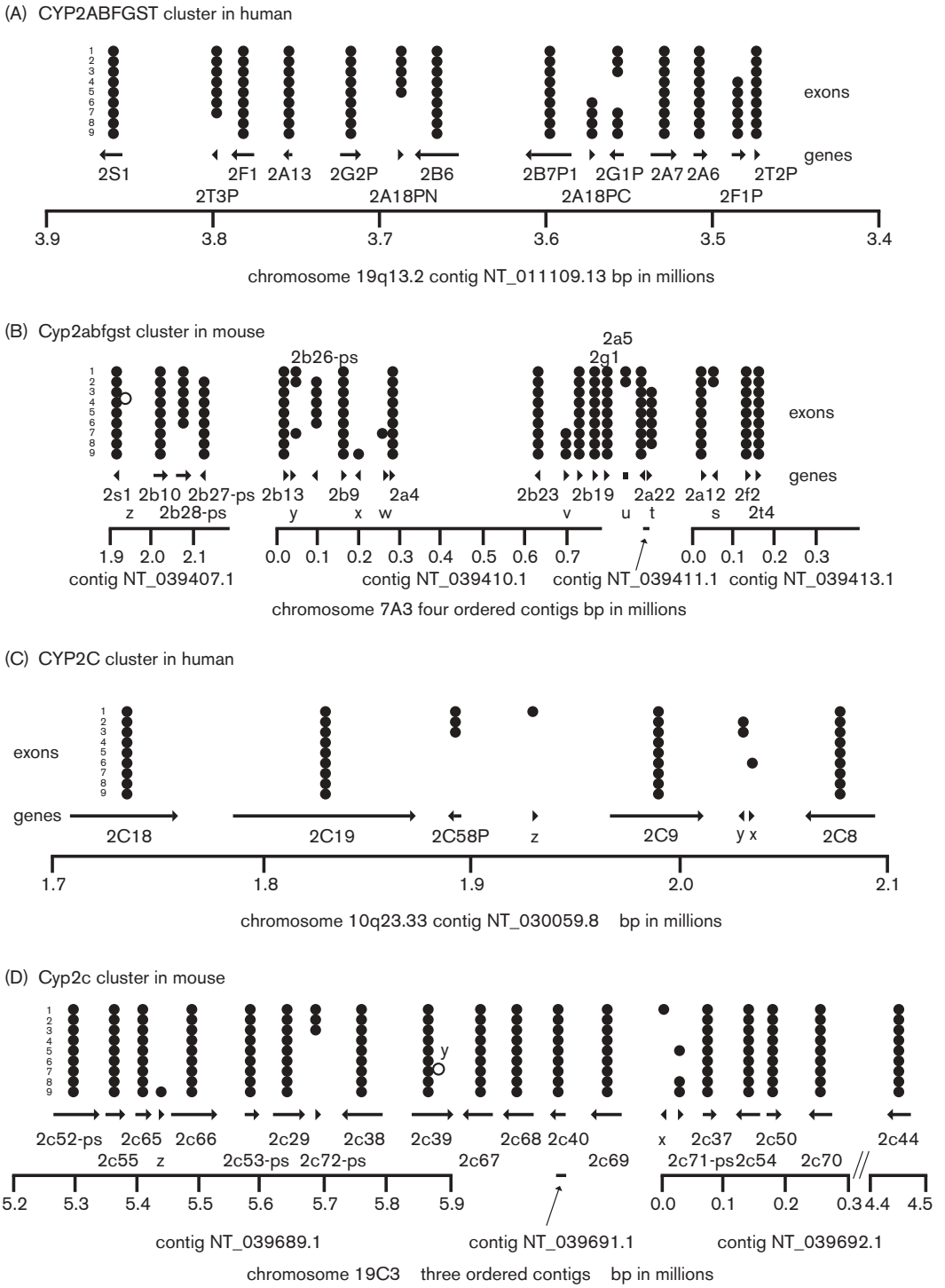
Pseudogenes provide rapid molecular clocks in nuclear DNA, as they are not conserved. These clocks can be

Fig. 1



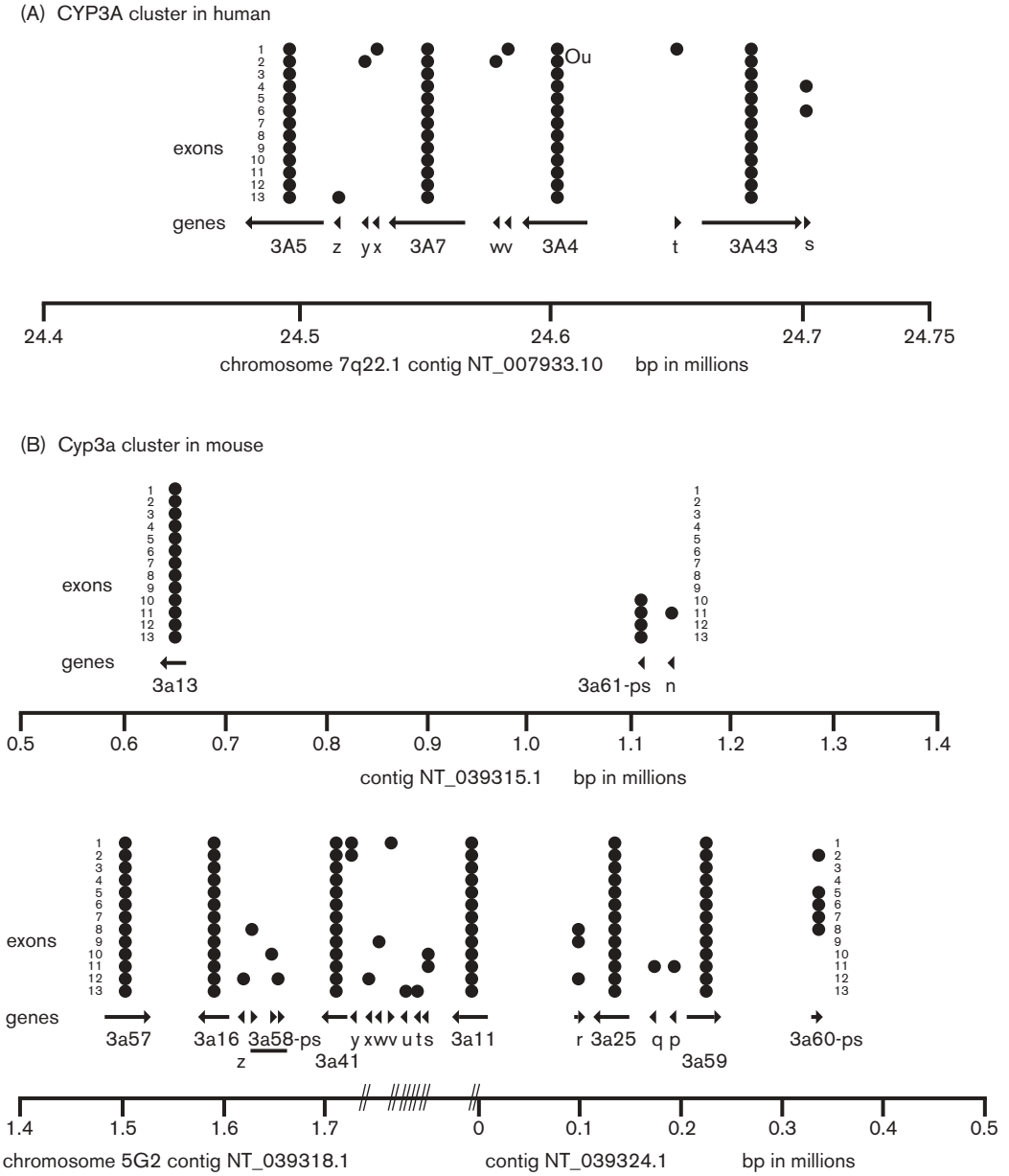
Unweighted-pair-group method of averaging tree of 102 mouse and 57 human *CYP* sequences, plus three human pseudogene sequences (*CYP2G2P, CYP2T2P, CYP4F23P*) and mouse *Cyp2ac1-ps*. The sequence alignment is posted at (http://drnelson.utmem.edu/mouse.human.aln2.html). The 34 orthologous pairs (discussed in text) are shaded *grey*. Clans, hierarchical clusters of related *CYP* gene families, are identified. Per nomenclature rules, human *CYP* gene names are in all capital letters, whereas mouse *Cyp* gene names are lower-case, except for the first letter.

Fig. 2



The *CYP2ABFGST* and *CYP2C* gene clusters in human and mouse. Each *filled circle* represents an exon or a part of an exon. *Open circles* denote extra internal exons. In (A), the contig has been drawn in reverse orientation, to allow direct comparison to the mouse cluster in (B). The mouse *Cyp2c44* gene is 4.1 Mb downstream of the main mouse *Cyp2c* cluster. Some pseudogenes have been labeled with letters. In accordance with the nomenclature system described in the text: in (B), **z** = *Cyp2s1-ie4b*, **y** = *Cyp2b13-de1b2b7b*, **x** = *Cyp2b9-de9b*, **w** = *Cyp2a4-de7b*, **v** = *Cyp2b19-de7b8b9b*, **u** = *Cyp2a22-de1b2b*, **t** = *Cyp2a21-ps*, **s** = *Cyp2a12-de1b2b*. *Cyp2b26-ps* is not found on NT_039410.1, but other evidence supports its location here (Wang *et al.* [35]). In (C), **z** = *CYP2C9-de1b*, **y** = *CYP2C9-de2c3c*, **x** = *CYP2C8-de6b*. In (D), **z** = *Cyp2c65-de9b*, **y** = *Cyp2c39-ie6b*, **x** = *Cyp2c71-de1b*. In this and the following figures, these pseudogenes are designated with single lower-case letters to avoid clutter in the diagram. The reverse alphabet was used to avoid possible confusion with *CYP* names. Mouse chromosome bands (7A3, 19C3) are taken from comparing the NCBI build 30 Map View genes_seq map to the ideogram map; these bands do not always agree with the MGI map location.

Fig. 3

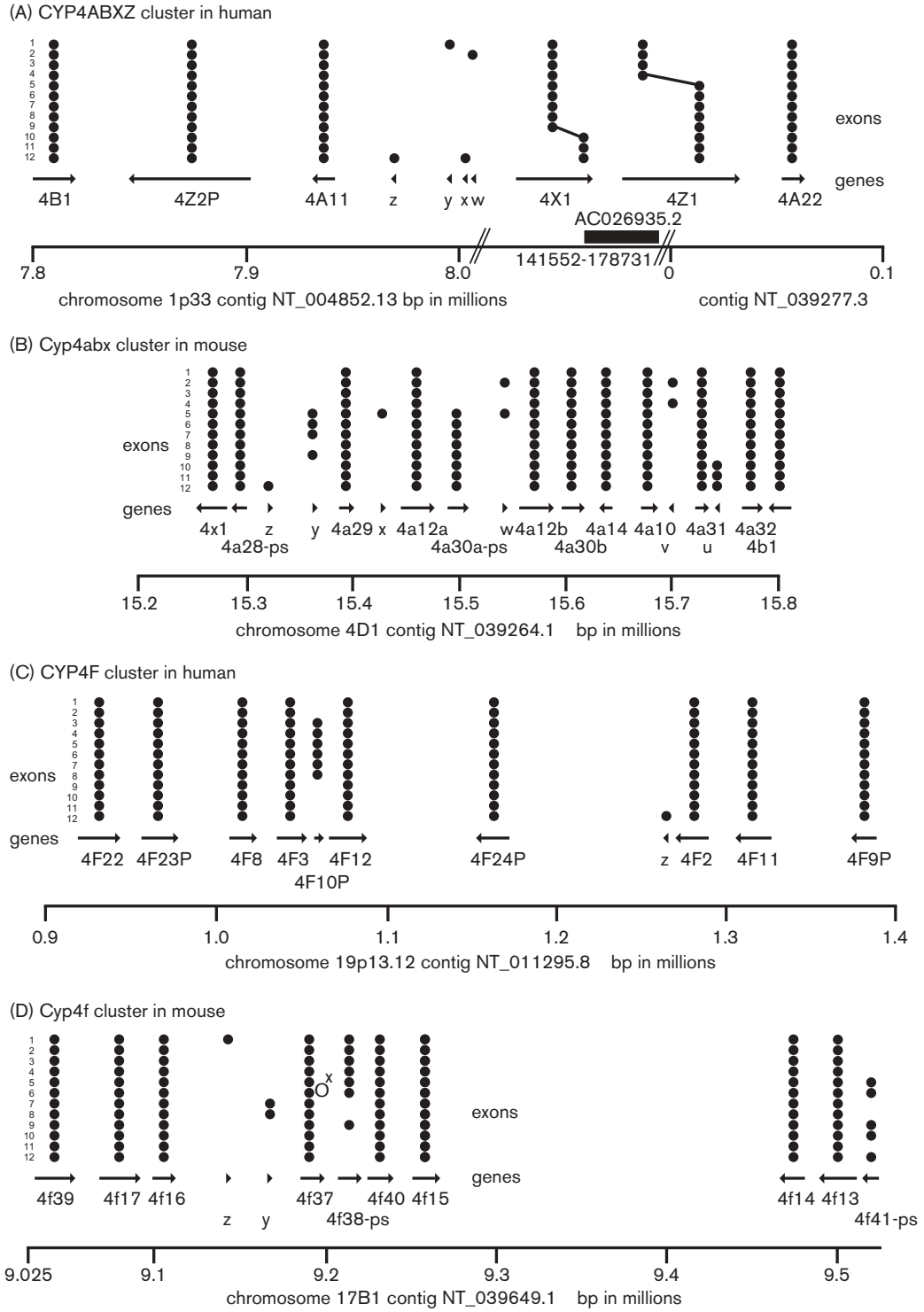


The *CYP3A* gene cluster in human and mouse. In (A), **z** = *CYP3A5-de13c*, **y** and **x** = *CYP3A5-de1b2b*, **w** and **v** = *CYP3A7-de1b2b*, **u** = *CYP3A4-ie1b*, **t** = *CYP3A43-de1b*, **s** = *CYP3A43-de4c6c*. In (B), the mouse *Cyp3a* cluster locus is still incomplete. *Cyp3a44* most likely lies between *Cyp3a41* and *Cyp3a11*. **z** = *Cyp3a16-de12b*, **y** = *Cyp3a41-de1b2b*, **x** = *Cyp3a41-de12c*, **w** = NT_039319.1 exon 9, **v** = NT_039319.1 exon 1, **u** = NT_039320.1 exon 13 (same as *Cyp3a41*), **t** = NT_039321.1 exon 13 (same as *Cyp3a41*), **s** = NT_039322.1 exon 10,11 (same as *Cyp3a41*), fragments **u**, **t** and **s** are likely to be incorrectly-mapped pieces of *Cyp3a41*. Fragments **w-s** are not named, since this region is incomplete and may need to be revised. **r** = *Cyp3a25-de8b9b12b*, **q** = *Cyp3a25-de11c*, **p** = *Cyp3a59-de11b*, **n** = *Cyp3a61-de11b*.

useful in understanding the dynamics of genome evolution on short time-scales. Comparison of pseudogenes from closely related genomes, such as those of the mouse and rat, may reveal more about mechanisms of evolution than comparisons of the much more slowly diverging functional genes. It is also likely that, over evolutionary time, pseudogenes will contain more single-nucleotide polymorphisms (SNPs), insertions and deletions than functional genes.

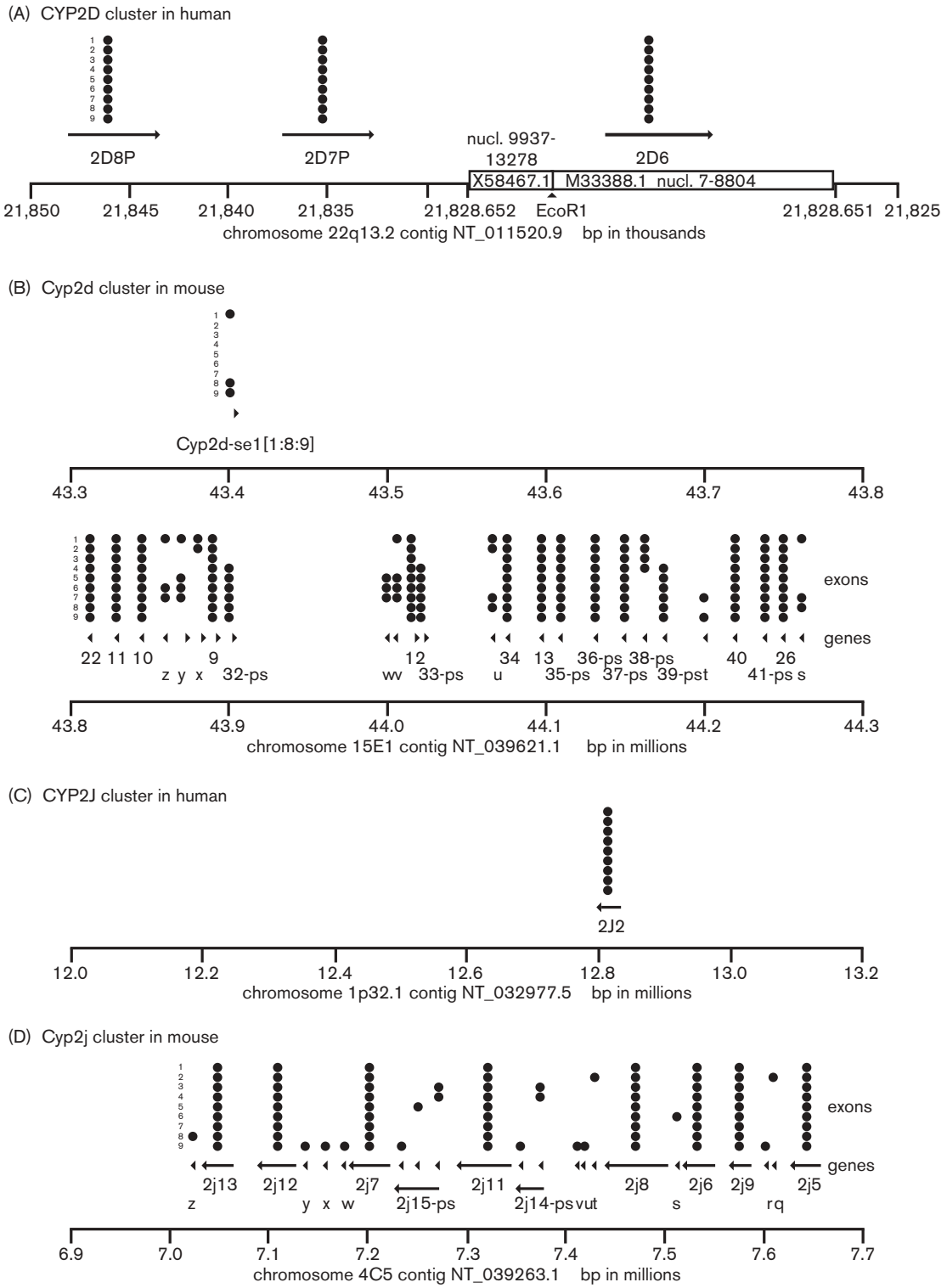
Pseudogenes are usually not functional, although an unusual exception was recently reported [21]. Pseudogenes, however, can play an important role in gene conversion and recombination events with a nearby

Fig. 4



The *CYP4ABXZ* gene cluster in human, the *Cyp4abx* gene cluster in mouse, and the *CYP4F* gene cluster in human and mouse. In (A), the human *CYP4ABXZ* cluster is incomplete, with a single fragment of AC026935.2 bridging the *CYP4X1* and *CYP4Z1* genes. Because the *CYP4X1* and *CYP4Z1* genes do not occur on a single genomic sequence, they are joined by an internal line to indicate this. **z** = *CYP4A-se1[12]*, **y** = *CYP4A-se2[1]*, **x** = *CYP4A-se3[12]*, **w** = *CYP4A-se4[2]*. (B) The *Cyp4abx* locus in mouse, because mouse does not have a *Cyp4z* gene. The region from fragment **x** to *Cyp4a30-ps* is a duplicate of the region from fragment **w** to *Cyp4a30b*; this results in two copies of the *Cyp4a12* gene called *Cyp4a12a* and *Cyp4a12b*. In (B), **z** = *Cyp4a29-de12b*, **y** = *Cyp4a29-de5b6b7b9b*, **x** = *Cyp4a12a-de5b*, **w** = *Cyp4a12b-de2b5b*, **v** = *Cyp4b1-de2b4*b, **u** = *Cyp4b1-de10c11c12c*. In (C), **z** = *CYP4F2-de12b*. In (D), **z** = *Cyp4f37-de1b*, **y** = *Cyp4f37-de7b8b*, **x** = *Cyp4f37-ie6b*.

Fig. 5



The *Cyp2d* and *Cyp2j* gene clusters in mouse and human. In (A), the deleted part of the human genome assembly has been restored as described in the text. In (B), **z** = *Cyp2d9-de1d6d7d*, **y** = *Cyp2d9-de1c5c6c7c*, **x** = *Cyp2d9-de1b2b*, **w** = *Cyp2d12-de5c6c7c*, **v** = *Cyp2d12-de1b5b6b7b*, **u** = *Cyp2d34-de1b2b7b8b*, **t** = *Cyp2d40-de7b9b*, **s** = *Cyp2d26-de1b7b8b*. In (C), the only human CYP2J is shown. In (D), **z** = *Cyp2j13-de8b*, **y** = *Cyp2j7-de9d*, **x** = *Cyp2j7-de9c*, **w** = *Cyp2j7-de9b*, **v** = *Cyp2j8-de9c*, **u** = *Cyp2j8-de9b*, **t** = *Cyp2j8-de2b*, **s** = *Cyp2j6-de6b*, **r** = *Cyp2j5-de9b*, **q** = *Cyp2j5-de2b*.

functional gene. Pseudogenes also provide points of reference in the genome, and their quantity and quality can suggest how actively a particular subfamily or gene cluster has been evolving. Pseudogenes might also testify to the dynamic nature of functional genes and genomes, which seem to be constantly throwing off pseudogenes like sparks sputtering from a fuse.

Why annotate and name all pseudogenes?

Pseudogenes pose a challenge for annotation and nomenclature, but they are too abundant to be ignored. The glyceraldehyde 3-phosphate dehydrogenase gene (*GAPDH*), for example, exists as a single functional gene in the human, mouse, rat and chicken [22]. In the mouse and rat, however, there are more than 400 retroviral-processed *GAPDH* pseudogenes [23,24]. Pseudogenes therefore need identification tags as much as or more than, any single functional gene. Fortunately, the *CYP* superfamily does not have as large a pseudogene faction as *GAPDH* (only 88 mouse and 58 human *CYP* pseudogenes), but a systematic nomenclature system still needs to be established for their proper identification. Such a system could be applied generally to other gene families. Users of the UCSC genome browsers for mouse or human have probably noticed the lines denoting GenScan predictions of genes. The output from GenScan is an automated attempt to find genes in the genome [25], and this program tries very hard to make a gene out of anything that looks like a gene, even if it is a pseudogene. The resulting constructs contain exons, with frameshifts and/or stop-codons, which are either truncated or missed by the program; new best-guess N- and C-terminal sequences are also added in order to provide plausible start-codons and final exons, respectively. These GenScan predictions are used to make estimates of the number of genes in a genome. Correct annotation of all pseudogenes in the genome would eliminate spurious predictions of this type from being counted as true genes.

Some pseudogenes are very nearly intact, for example, human *CYP2G2P*. This gene has all its intron-exon boundaries preserved and contains only two stop-codons in the coding sequence – one in exon 1, and one in exon 3. If the gene were to have polymorphisms at both of these sites, then functional alleles of *CYP2G2P* may exist. Mouse *Cyp2g1* [26] and rabbit *CYP2G1* [27] are functional genes. Another example is human *CYP2D7AP*, which has only one frameshift in exon 1 and one aberrant GT donor splice site at the exon 2–intron 2 boundary. Because the *CYP2D6* gene is highly polymorphic, with more than 70 allelic variants [19], it is possible that gene conversion or alternative splicing might occur in the similar *CYP2D7AP* pseudogene, resulting in a functional gene.

In the combined mouse and human *CYP* gene families, there are 146 pseudogenes and 159 putative functional genes, indicating that the number of pseudogenes is almost equal to the number of actual genes. One study has suggested there exist about 20 000 pseudogenes in the human genome [15], and although this group's website identifies about 12 000 pseudogenes [28], they have annotated only one of the 58 human *CYP* pseudogenes (*CYP51P1*). We suspect that their global approach is not based on a careful search for all members of a gene family, and therefore the prediction of 20 000 pseudogenes is very likely to be an undercount.

Proposed nomenclature system for pseudogenes

The data in Figs 2, 3, 4 and 5 show that there are many *CYP* pseudogenes in mouse and human. At present, the precise definition of 'pseudogene' is still under discussion by nomenclature committees. Our observations suggest that there are at least four categories of pseudogenes: (a) full-length, or nearly full-length, pseudogenes; (b) solo exons, or small groups of exons, away from a gene cluster; (c) escaped exons near a parent gene or gene cluster – which we propose to call detritus exons – representing fragments caused by the common processes of gene disintegration and partial gene duplication; and (d) internal exons, duplicated intact exons, or partial exons found inside genes. We propose a systematic nomenclature system, based on extensions of the usual *CYP* gene name, and designed to distinguish between these four types of pseudogenes. This system does not identify the pseudogenes as 'processed' or 'non-processed.' The mouse and human pseudogenes are listed using this new system, within the mouse and human sections of the Cytochrome P450 Homepage [3].

Full-length, or nearly full-length, pseudogenes

Previously, only one designation was given for all pseudogenes; this designation includes a '*P*' in most species (in mouse, '*-ps*' [29]) appended to the name – as in human *CYP2T2P* or mouse *Cyp2c52-ps*. If two or more pseudogenes related to a given parent gene were characterized, then they were given additional numbers, as in *CYP51P1*, *CYP51P2* and *CYP51P3*. This designation will be kept for the full-length, or nearly full-length, pseudogenes. These 'almost complete' pseudogenes may be chromosomally secluded from other members of their family or subfamily, as free-standing pseudogenes (e.g. the three human *CYP51* processed pseudogenes located on three different chromosomes), or they may be found within their gene cluster (e.g. the mouse *Cyp2d* cluster, Fig. 5B). *CYP* pseudogenes are predominantly duplicated genes rather than retrotransposed genes.

Solo exons

Solo-exon pseudogenes will have the extension '*-sen*[x_1:x_2:x_i]', in which '*-se*' denotes 'solo exon,' and '*n*'

represents a unique chronological number (1, 2, 3, etc.) to specify multiple pseudogenes from the same parent gene or gene subfamily. The '*[x_1:x_2:x_i]*' values in brackets refer to which exon(s) has(have) been duplicated (e.g. exon 2, exon 7, etc.) in the pseudogene. For example, there are 13 similar human *CYP4F* pseudogenes scattered about the genome, whose old nomenclature included: *CYP4F25P*, *CYP4F26P*, *CYP4F27P*, *CYP4F29P*, *CYP4F32P*, *CYP4F33P*, *CYP4F34P* and *CYP4F35P*. These small pseudogenes represent solo exons, are very similar to one another, and probably arose by duplication of a single exon-6,7,8 fragment. In this new pseudogene nomenclature system, these will be named *CYP4F-se1[6:8]*, *CYP4F-se2[6]*, *CYP4F-se3[6:7:8]*, *CYP4F-se4[6:7:8]*, *CYP4F-se5[6:8]*, *CYP4F-se6[6]*, *CYP4F-se7[6:7:8]*, *CYP4F-se8[6:7:8]*, *CYP4F-se9[6:7:8]*, *CYP4F-se10[6:7:8]*, *CYP4F-se11[6:7:8]*, *CYP4F-se12[6:8]* and *CYP4F-se13[6:8]*, respectively. Because the species is not included in the name, numbering will be sequential across species.

Solo exons, by definition, do not occur inside gene clusters, unless they are well separated from neighboring genes. There may exist a continuum from full-length pseudogene to solo exons. A transition in the nomenclature between solo exons and longer, free-standing pseudogenes is required. An arbitrary breakpoint of four exons is recommended. If a pseudogene has more than four exons present, it makes sense to name it as a free-standing pseudogene; otherwise, the name becomes too long and cumbersome. If three or four exons are present, and the pseudogene is inside a gene cluster, it may be given an autonomous pseudogene name (e.g. human *CYP2C58P*), or it could be treated as a detritus exon pseudogene.

Solo exons are rare in the mouse genome; the only exception is that two *Cyp2c* pseudogenes are found on other chromosomes away from the *Cyp2c* gene cluster (see below). All other mouse *Cyp* pseudogenes remain close to the cluster of origin.

Detritus exons

Detritus exons will have the extension '*-dex_1jx_2j...x_ij*', where '*de*' denotes detritus exon, 'x_1, x_2, x_i' represents the number(s) of the exon that has(have) been duplicated, and '*j*' refers to a lower-case letter (*b*, *c*, *d*, etc.) to designate the unique pseudogene. The letter '*a*' is reserved for the parent-gene normal exons; these letters are necessary because there may be two or more pseudogenes associated with one parent gene. For example, the mouse *Cyp2j* gene cluster (Fig. 5D) contains: *Cyp2j7-de9b*, *Cyp2j7-de9c*, *Cyp2j7-de9d*. These represent the three exon-9 pseudogene fragments downstream of the *Cyp2j7* functional gene. *Cyp2j7-de9b* is so named, because it is the closest detritus exon to the *Cyp2j7* gene. The curious case of the human *CYP3A7* gene – which has 13 exons (Fig. 3A) – is also instructive. Here, three detritus exons – *CYP3A7-de1b*, *CYP3A7-de2b* and *CYP3A7-de13c* – are downstream of the human *CYP3A7* gene. Transcripts have been found that include both *CYP3A7-de2b* and *CYP3A7-de13c* in an out-of-frame fusion to the end of the normal CYP3A7 mRNA [30]. In this case, not only are these detritus exons, but they are also used to make a non-CYP-related sequence extension of the CYP3A7 mRNA, such that they become the 14th and 15th exons in an alternative transcript.

Internal, duplicated-intact, or partial exons

Internal exons will carry the extension '*-iex_ij*', in which '*ie*' denotes internal exons, 'x_i' refers to the number of the exon(s) that has(have) been duplicated, and '*j*' represents a lower-case letter (*b*, *c*, *d*, etc.) to distinguish the extra exons from the parent exon, always called '*a*.' Internal exons are rare in *CYP* genes, but may be more common in other gene superfamilies. Two *CYP* examples include human *CYP3A4-ie1b*, a partial duplication of exon 1 of the *CYP3A4* gene, and mouse *Cyp2c39-ie6b*, an intact duplication of exon 6 of the *Cyp2c39* gene. In the latter case, it is possible that this exon might be used for alternative splicing of the mouse *Cyp2c39* gene.

Unidentified or unaffiliated pseudogenes

Some pseudogenes have decayed to the point where their subfamily relationship is no longer detectable. There are several nomenclature options in this case. The pseudogene could be given a new subfamily name; this seems undesirable, however, because it will be counted as a new subfamily, even when no functional sequence exists in that subfamily. Another option is to give the pseudogene a name that denotes the family only, as in *CYP2P1*; this option runs into problems with the naming system already in place, where *CYP2P* is a legitimate gene subfamily of its own (in fish). To avoid this, the '*un*' suffix (for unassigned) could be used, such as human '*CYP2-un1*' or mouse '*Cyp2-un1*,' meaning pseudogene 1 of the human *CYP2* or mouse *Cyp2* family (not assigned to a specific subfamily). There are no examples of these in the mouse or human P450s, though *CYP1A8P* is close. The '*un*' suffix would only be used for pseudogenes, and not for full-length functional genes.

If a pseudogene is so decayed that it cannot be reliably assigned to any specific family (for example, equally identical to two families), the name human '*CYP-un1*' or mouse '*Cyp-un1*' can be used to indicate 'a *CYP* pseudogene 1 that is not assigned to any specific family.' Because of the potential number of pseudogenes of this type across various phyla, we suggest that it might be beneficial to include the species in the name (e.g. *Cyp-un1mouse* for mouse, *CYP-un1human*

for human, *CYP-un1anoga* for *Anopheles gambiae*, *CYP-un1takru* for *Takifugu rubripes*, etc.; except for *mouse*, *human*, *rat*, *pig*, *rabit*, *bovin* and *horse*, usually by using the first three letters of the genus, plus the first two letters of the species, those five letters will cover unequivocally all species on this planet (http://www.expasy.org/cgi-bin/speclist). The shorter 'Hsa' for *Homo sapiens* is sometimes used, but does not distinguish this species from *Halobacterium salinarium*, *Hypsugo savii* or *Herpesvirus saimiri*).

Alternative splicing and suggested nomenclature of transcripts

The presence of internal exons raises the additional problem of how to name alternative-splice variants. In the simplest case, such as mouse *Cyp2c39* described above, the name for the transcript needs to include the exon used when there is an alternative option. Assuming that the CYP2C39 mRNA can exist in two alternative forms, we suggest the corresponding transcript names *Cyp2c39_v6a* and *Cyp2c39_v6b*, since all exons other than exon six are the same.

The _v symbol for alternative transcripts is in keeping with the Guidelines for Human Gene Nomenclature (2002) [31]. The _v symbol is not to be confused with existing *CYP* nomenclature using v without an underscore to indicate alleles or sequence variants – as in *CYP74B4v1* and *CYP74B4v2*. In humans, the allele nomenclature has adopted the * symbol for alleles (see the human *CYP* alleles nomenclature website http://www.imm.ki.se/CYPalleles/).

Alternative-splice variants that are more complex could be specified by including all of the optional exons in the name. This can become difficult and long, however, as shown in the case of the previously named *CYP3A5P1* (Genbank L26985) and *CYP3A5P2* (X90579). These both encode non-functional transcripts of the human *CYP3A5* gene; they had initially been named as distinct pseudogenes, assuming separate loci [32], but this was shown not to be the case. These two transcripts are derived by alternative splicing from the *CYP3A5* gene in a rather complex fashion. If the normal gene transcript is represented by a contiguous list of all 13 exons, as CYP3A5–1a2a3a4a5a6a7a8a9a10a11a12a 13a, then the *CYP3A5P1* transcript would be CYP3A5_v1a2a3a14n4a15n5a16nΔ6a7a8a9a10axΔ11aΔ12aΔ13a. In this name, '14n,' '15n' and '16n' are new cryptic exons not related to any of the normal *CYP3A5* exons. The 'Δ' preceding a normal exon (e.g. Δ6a) denotes that the exon has been skipped. The '10ax' signifies that normal exon 10 has been *extended* downstream into the intron sequence. A 'y' as in 10ay could indicate exon extension upstream. The additional symbols q and r could be used to indicate a shift of intron–exon boundaries into the exon from the 5′ or 3′ end. *CYP3A5P2* is the same as *CYP3A5P1*, except that it is missing the cryptic exon 15n. Clearly there are benefits to using the shorthand names, although the short names imply that these transcripts come from a pseudogene, when in this case they do not. The solution is to treat the normal transcript as *CYP3A5_v1* and rename the *CYP3A5P1* transcript as *CYP3A5_v2* and the *CYP3A5P2* transcript as *CYP3A5_v3*, as shorthand for the long descriptive names. The long descriptive names could be kept on a website or database for reference, whereas the shorthand names could be used in publications. We hope that this extremely complex relationship – between a gene and several reasonably abundant transcripts that lead to a non-functional protein – is rare.

There also needs to be conventional nomenclature when an alternative exon 1 exists. The first exon 1 may be far upstream of the gene, and it may be used infrequently or not at all (see *CYP3A5* and *CYP3A7*). If the *CYP3A5* and *CYP3A7* exons are named 1a and 1b in order on the chromosome, then the most-often-used exon would be exon 1b. In the case of first exons, we suggest that it may be better to name the alternative exon 1 that is closer to exon 2 as 'exon 1a,' and the more distant alternative exon 1 as 'exon 1b.' In the case of *CYP3A5* and *CYP3A7*, there are two upstream exons, named 1 and 2; these have been named as detritus exons *CYP3A5-de1b2b* and *CYP3A7-de1b2b* (Fig. 3A). A transcript using these exons would be named 'CYP3A7_v1b2b mRNA'.

Human *CYP19A1* is another case of alternative first exons [33], but this gene is far more complex. The human *CYP19A1* gene spans more than 130 kb, with the coding region in exons 2–10 spanning about 30 kb, and perhaps as many as six non-coding first exons that determine tissue-specific expression. The expressed gene product (aromatase) is the same in all tissues. Again, we propose the alternative exon 1 closest to exon 2 be called '1a,' and the sixth alternative exon 1 upstream would be called '1f.'

Comparison of *CYP* genes from mouse and human

Figure 1 shows a phylogenetic tree containing 163 *CYP* sequences (102 full-length putatively functional mouse *Cyp* genes, plus 57 full-length putatively functional human *CYP* genes). We have also included three human pseudogenes: *CYP4F23P* is very nearly intact, whereas *CYP2G2P* and *CYP2T2P* are orthologs of the mouse *Cyp2g1* and *Cyp2t4* functional genes. The mouse *Cyp2ac1-ps* pseudogene is an ortholog of the functional rat *CYP2AC1* gene. A direct comparison in list form of all functional mouse and human *CYP* genes is given in Table 2. This tree clearly shows that, compared with human, the mouse has undergone significant expansion of *Cyp* genes in seven clusters: the *Cyp2abfgst* cluster,

Table 2 Human and mouse putatively functional full-length *CYP* genes

Human	Mouse	Human	Mouse	Human	Mouse
CYP1A1	Cyp1a1	CYP2E1	Cyp2e1	CYP4F2	Cyp4f13
CYP1A2	Cyp1a2	CYP2F1	Cyp2f2	CYP4F3	Cyp4f14
CYP1B1	Cyp1b1		Cyp2g1[b]	CYP4F8	Cyp4f15
CYP2A6	Cyp2a4	CYP2J2	Cyp2j5	CYP4F11	Cyp4f16
CYP2A7	Cyp2a5		Cyp2j6	CYP4F12	Cyp4f17
CYP2A13	Cyp2a12		Cyp2j7	CYP4F22	Cyp4f18
	Cyp2a22		Cyp2j8		Cyp4f37
CYP2B6	Cyp2b9		Cyp2j9		Cyp4f39
	Cyp2b10		Cyp2j11		Cyp4f40
	Cyp2b13		Cyp2j12	CYP4V2	Cyp4v3
	Cyp2b19		Cyp2j13	CYP4X1	Cyp4x1
	Cyp2b23	CYP2R1	Cyp2r1	CYP4Z1	a
CYP2C8	Cyp2c29	CYP2S1	Cyp2s1	CYP5A1	Cyp5a1
CYP2C9	Cyp2c37		Cyp2t4[b]	CYP7A1	Cyp7a1
CYP2C18	Cyp2c38	CYP2U1	Cyp2u1	CYP7B1	Cyp7b1
CYP2C19	Cyp2c39	CYP2W1	Cyp2w1	CYP8A1	Cyp8a1
	Cyp2c40		Cyp2ab1[b]	CYP8B1	Cyp8b1
	Cyp2c44		Cyp2ac1-ps[b]	CYP11A1	Cyp11a1
	Cyp2c50	CYP3A4	Cyp3a11	CYP11B1	Cyp11b1
	Cyp2c54	CYP3A5	Cyp3a13	CYP11B2	Cyp11b2
	Cyp2c55	CYP3A7	Cyp3a16	CYP17A1	Cyp17a1
	Cyp2c65	CYP3A43	Cyp3a25	CYP19A1	Cyp19a1
	Cyp2c66		Cyp3a41	CYP20A1	Cyp20a1
	Cyp2c67		Cyp3a44	CYP21A2	Cyp21a1
	Cyp2c68		Cyp3a57	CYP24A1	Cyp24a1
	Cyp2c69		Cyp3a59	CYP26A1	Cyp26a1
	Cyp2c70	CYP4A11	Cyp4a10	CYP26B1	Cyp26b1
CYP2D6	Cyp2d9	CYP4A22	Cyp4a12a	CYP26C1	Cyp26c1
	Cyp2d10		Cyp4a12b	CYP27A1	Cyp27a1
	Cyp2d11		Cyp4a14	CYP27B1	Cyp27b1
	Cyp2d12		Cyp4a29	CYP27C1	a
	Cyp2d13		Cyp4a30b	CYP39A1	Cyp39a1
	Cyp2d22		Cyp4a31	CYP46A1	Cyp46a1
	Cyp2d26		Cyp4a32	CYP51A1	Cyp51a1
	Cyp2d34	CYP4B1	Cyp4b1		
	Cyp2d40				

[a]Subfamilies absent in mouse.
[b]Subfamilies in mouse with only pseudogene orthologs in human.
CYP2AC1P and *Cyp2ac1-ps* are recently discovered pseudogenes in human and mouse, respectively, but a functional gene in rat. The gene names are listed numerically and alphabetically by subfamilies, and pairing does not necessarily denote orthologous genes, e.g. it is not known whether human *CYP2A6* is the ortholog of mouse *Cyp2a4*. Genes having the identical combination of numbers and letters are orthologs between the two species (see text).

the *Cyp2c* cluster, the *Cyp2d* cluster, the *Cyp2j* cluster, the *Cyp3a* cluster, the *Cyp4abx* cluster, and the *Cyp4f* cluster. Note in Fig. 1 that all the *CYP* gene clusters occur at about the same depth on the tree; this might indicate a common time for beginning the expansion in all the clusters.

The initial analysis of the mouse genome noted some gene families are expanded in the mouse [2]. Even though 80% of mouse and human genes have a 1:1 orthologous relationship, there are 147 gene clusters found in mouse with four or more related genes; these represent recent gene duplications in the mouse. The olfactory receptor family is expanded about three-fold in the mouse, and includes 47 of these 147 gene clusters. Many other gene clusters are related to reproduction and immunity.

It is important to discuss the time-scale of *CYP* evolution in mammals. The mouse-human split is estimated to have occurred 75 MYA [2]. Thus, the mouse and human have had about 75 MY to accumulate differences in their genomes. Mouse and rat diverged approximately 20 MYA [18], whereas human and chimpanzee diverged at least 6 MYA. Seventy-five million years is sufficient time for major changes in gene-cluster size and organization, but not enough time for new *CYP* families to evolve. When the human and *Takifugu rubripes* genomes are compared [11], for example, only one *CYP* family differs between mammal and fish; *CYP39A1* is found in mammals but not in fish or other non-mammalian vertebrates. *CYP39A1* appears to be specific for mammals and has presumably evolved since the mammal-bird divergence about 310 MYA.

Subfamilies have changed somewhat in mouse and human, with the loss of four subfamilies in humans relative to mouse or rat: *CYP2G*, *CYP2T*, *CYP2AB* and *CYP2AC* are present only as pseudogenes in the human. Perhaps the products of these genes carry out functions

in rodents that are no longer needed in the human. On the other hand, orthologs of the human *CYP27C1* and *CYP4Z1* genes are absent in mice.

CYP27C1 is found in humans, fish, birds and frogs, whereas the mouse (and probably the rat) appears to have lost *Cyp27c1*, due to a chromosomal rearrangement in the precise region between the *Bin1* and *Ercc3* genes. We speculate that the *CYP27C1* gene in the ancestral rodent was broken and lost during this event. In *Takifugu rubripes*, the *CYP27C1* gene is on the minus strand of scaffold 106, from nucleotides 31 916–36 680. In fish, the nearest gene, 2624 bp away, is *ERCC3*, at 39 305–43 553 on the minus strand. The *ERCC3* gene is also found in humans, 39 kb away from *CYP27C1*. Therefore, the *CYP27C-ERCC3* linkage is very old, at least 420 MY. The next human gene in the *ERCC3–CYP27C1* series is *BIN1*. In mouse, chromosome 1 is syntenic to human chromosome 2 at the *ERCC3* locus, but *Bin1* is on mouse chromosome 18, indicating that a chromosome break has occurred in this region.

CYP4Z1 has been seen only in humans so far, suggesting that this may be a new subfamily in humans, rather than a loss of *Cyp4z1* in mouse or rat. It will be interesting to see if *CYP4Z1* is present as a functional gene in a non-human primate – chimpanzee – when this genome draft assembly is released in late 2003.

The *CYP2ABFGST* cluster

The mammalian *CYP2ABFGST* cluster is the only example so far, in the *CYP* gene superfamily, in which there is intermingling of members from six subfamilies in the same chromosomal region. Even after 75 MY, there is some detectable similarity of order in the *CYP2ABFGST* clusters of mouse and human (Fig. 2A, B). The clusters both begin with *CYP2S1* orthologs and end with *CYP2T* orthologs, although the latter has become a pseudogene in humans, *CYP2T2P*. The gene order is similar between the mouse and human clusters on the *CYP2T* end, with *CYP2G*, *CYP2A* and *CYP2F* loci in a series in both genomes, though the mouse has more *Cyp2a* genes. In the human, the region from *CYP2G1P* to *CYP2T2P* appears to have duplicated in mirror symmetry, inside the *CYP2S1*-end of the cluster [34]. Between these two symmetric segments of the human cluster, a single *CYP2A* gene appears to have been split (into *CYP2A18PN* and *CYP2A18PC*, reflecting the N- and C-termini of the protein, respectively) by insertion of *CYP2B6* and *CYP2B7P1*. The mouse cluster has neither this mirror symmetry nor a split *Cyp2a* pseudogene. Between the *Cyp2s1* and *Cyp2g1* genes, there has been a considerable expansion of *Cyp2b* genes in the mouse but not in the human [8,35].

The close proximity of so many related genes has given rise to some gene recombinations. The CYP2A7 cDNA U22029 does not match the *CYP2A7* genomic sequence at exon 8. The U22029 exon 8 sequence matches instead *CYP2A18PC*, suggesting a possible conversion event between these two genes. The *CYP2A6*12* allele is a hybrid of exons 1–2 from *CYP2A7* and exons 3–9 of *CYP2A6* [36]. Another *2A6/2A7* hybrid is described by Fernandez-Salguero *et al.* [37]. This gene has exons 3, 6 and 8 derived from the *CYP2A7* gene, probably by gene conversion, while the remainder matches *CYP2A6*. The potential for recombining genes is enhanced many times in the large mouse clusters of very similar genes such as the *Cyp2c*, *Cyp2d* and *Cyp2j* clusters. Alternative splicing adds another layer on top of this complexity. A *CYP2A7* transcript is known that extends 10 bp into intron 1 and is deleted for exon 2 (NM_030589, CYP2A7_v1axΔ2a).

The *CYP2C* cluster

The *CYP2C* cluster in humans is small, with only four genes, whereas in mouse it has expanded to 15 genes (Figs. 3C & 3D). The human *CYP2C* genes have a strong potential to recombine, due to many L1 LINE repetitive DNA sequences, which are located principally in intron 5. *CYP2C9* and *CYP2C19* share *L1PA7*, *L1M4*, *L1MB5* and *L1PA16* repeats in this intron. *CYP2C18* and *CYP2C19* share *L1PA5* repeats. *CYP2C8* and *CYP2C19* share an *L1P* repeat, although the two genes are on opposite strands. Recombination at these repeat sites could produce novel human *CYP2C* transcripts containing hybrid sequences that correspond to exons 1–5 from one gene and exons 6–9 from the other; other types of recombination or gene conversion are also possible.

The mouse *Cyp2c* cluster on chromosome 19 (Fig. 2D) is a prime example of the expansion that has taken place in some *Cyp* subfamilies in this species. Nearly all the duplications have occurred inside the cluster, except for *Cyp2c44*, which is about 4 Mb downstream. There are also two solo exon-9 *Cyp2c* pseudogenes on chromosomes 14 and 16. Note that the *Cyp2c44* gene is the most distant branch on the *Cyp2c* cluster (Fig. 1). *Cyp2c70*, which is also outside the main group of *Cyp2c* genes in the tree (Fig. 1), is found on the edge of the main cluster (Fig. 2D). These outer locations seem to be more resistant to recombination events (see the *CYP4F* cluster below). It should be possible to estimate the rapidity of change in this mouse cluster, by comparison with the rat genome that will soon be available.

In Fig. 1, the *Cyp2c* cluster joins with *Cyp2e*, and this group then joins with the *Cyp2abfgst* sequences. These clusters do not have orthologous clusters in fish. The most similar sequences in fish are the *Fugu CYP2Y1* and *CYP2Y2* and the trout *CYP2M1*. Since there is a *CYP2C45* gene in chicken, evolution of the *CYP2C* subfamily must have occurred before the bird-mammal

divergence, about 310 MYA. A *CYP2S* expressed-sequence tag (EST, BM491346) is also found in chicken, meaning that by 310 MYA the *CYP2C* cluster was apparently distinct from the *CYP2ABFGST* cluster. These two clusters, now quite expanded in mammals, probably arose from a single *CYP2Y*-like gene present in the tetrapod ancestor. Because there is more diversity in the *CYP2ABFGST* cluster, it is probably older than the *CYP2C* cluster.

The *CYP2D* cluster

The *Cyp2d* cluster is greatly expanded in the mouse (Fig. 5B), having nine full-length putatively functional genes, compared with only one in humans (*CYP2D6*). Human *CYP2D6* allelic variants include one (**5* allele) in which the gene is deleted and another (**1X13* allele) in which 13 copies of the duplicated functional gene are located in tandem. *CYP2D6* encodes a drug-metabolizing enzyme that has more than seventy drugs as substrates [38]; patients who lack *CYP2D6* are subject to adverse drug reactions, whereas patients with multiple copies of *CYP2D6* commonly exhibit therapeutic failure due to overly-rapid drug metabolism [39]. The recent demonstration that CYP2D6 endogenous substrate activities include 5-methoxyindolethylamine *O*-demethylase [40] and serotonin formation [41] causes us to query whether non-primate CYP2D enzymes would carry out similar endogenous functions.

The copy of chromosome 22 sequenced by The Human Genome Project has a deletion allele of *CYP2D6*; this has been named the *CYP2D6*5* allele, with an allelic frequency of about 0.04 in Caucasian populations [19]. Figure 5A attempts to restore the deleted sequence to NT_011520.9 by adding portions of two independent sequences (X58467.1 covering the end of *CYP2D7AP*, and M33388.1 covering the *CYP2D6* gene). These two sequences are joined at an *Eco*R1 site at the end of X58467 and at the beginning of M33388. The break in the genome assembly is shown as occurring at nucleotide 21 828 652. This is the site of an extra G base in *CYP2D7AP* that is not seen in *CYP2D6*. The assembly resumes at the equivalent position in the *CYP2D6* gene. Therefore, an estimated 12 142 bp have been deleted. About 2400 bp of the 3′ sequences are 99% identical between *CYP2D7AP* and *CYP2D6*, allowing an unequal crossover event to have deleted the *CYP2D6* gene [42].

It is not clear why the mouse has expanded the *Cyp2d* cluster from one to nine genes during the past 75 MY. Nor is it clear why the *Cyp2c* and *Cyp2j* clusters have been expanded in mouse, as compared with that in human. However, because *CYP2D6* is noted for its exogenous drug metabolism capabilities, it is more likely that most – if not all – of the duplicated *Cyp2d* genes are acting on foreign substrates such as plant and dietary components, rather than on new endogenous substrates. The biochemical pathways involving steroids, sterols, fatty acids, bile acids, biogenic amines and other CYP substrates [1] are unlikely to have changed during the past 75 MY, because these pathways seem important for signaling needs in mammals. In rodents, however, there might have occurred additional selective pressure from the diet, or from pheromone production, to increase the P450 repertoire. Perhaps the tissue- or cell type-specific location of *Cyp2d* gene expression in the mouse will hold clues to the functions of some of these genes.

The *CYP2J* cluster

The mouse *Cyp2j* cluster (Fig. 5D) has eight genes, compared with the single *CYP2J2* gene in human [43–46]. This cluster has the unusual property that all the genes and pseudogene fragments are oriented in the same direction, which is not the case for the other six *CYP* gene clusters. Also, there are no mouse *Cyp2j* pseudogenes outside the cluster. These data suggest a fairly recent amplification, with insufficient time for inversions to randomize the gene order. The proteins made from these mouse genes have similar substrate preferences, but the product profiles and the cell and tissue distribution are unique for each isoform; similar results have been found with the *Cyp2c* gene cluster (D. Zeldin, unpublished data).

The *CYP2J* and *CYP2D* clusters are phylogenetically close (Fig. 1). This is also true when fish sequences are included in the analysis [11]; in fish, however, there are no *CYP2J* or *CYP2D* orthologs. *CYP2J* and *CYP2D* probably had a single common ancestor in fish. Other loci derived from this ancestral gene probably include the members of the *CYP2K*, *CYP2N*, *CYP2P*, *CYP2V*, and *CYP2Z* subfamilies. Interestingly, CYP2N and CYP2P subfamily enzymes in fish have tissue distributions, substrate specificities and product profiles similar to Cyp2j enzymes in mammals [47,48].

The *CYP3A* cluster

The *Cyp3a* cluster in mouse has only twice as many P450 genes as in human (Fig. 3). This region on mouse chromosome 5 is not completely assembled; consequently, the *Cyp3a44* gene is not included in Fig. 3, although it is likely to be adjacent to *Cyp3a11*, based on data from the incomplete HTGS sequence AC111090.3. The human *CYP3A* genes all have potential alternative first and second exons, allowing for possible alternative transcripts. As mentioned above, alternative transcripts for the *CYP3A7* gene have been reported [30] that extend the C-terminus with out-of-frame fusions to a downstream exon 2 and exon 13 (x and z in Fig. 3A).

There are no orthologous *CYP3A* pairs between mouse

and human, suggesting that their common ancestor had a single *CYP3A* gene that has been expanding independently during the last 75 MY. *Takifugu rubripes* has a *CYP3B* sequence not seen in mammals, which may be a relic of the whole-genome duplication that was postulated to have taken place in ray-finned fish after they diverged from the tetrapod ancestor [11,49,50].

The *CYP4ABXZ* cluster

This cluster is the only one, other than the *CYP2ABFGST* cluster, that contains a mixture of loci from distinct subfamilies. The *CYP4B* and *CYP4X* genes have orthologs between mouse and human (Fig. 1), and therefore they must have existed as separate subfamilies more than 75 MYA. This gene cluster contains recent gene duplications – as shown in Fig. 1 by the very short branch-lengths between human *CYP4A11* and *CYP4A22*, between mouse *Cyp4a12a* and *Cyp4a12b* (found inside a 100-kb tandem duplication; see Fig. 4B), and between the adjacent triplet of *Cyp4a10*, *Cyp4a31* and *Cyp4a32* in the mouse cluster.

The *CYP4Z1* putatively functional gene and its pseudogene are seen only in humans. We can say this with certainty because the mouse *Cyp4abx* cluster has been completely sequenced. The *CYP4Z1* gene is presumably an innovation in the line leading to humans.

The fact that the human has only two *CYP4A* genes, and that these have duplicated only recently, suggests that the ancestor of mouse and human had a single *CYP4A* gene. There is one *CYP4T* gene in fish, which is the likely ancestor of the *CYP4ABX* and *CYP4Z* genes in mammals. The transition from the *CYP4T* lineage to the *CYP4ABXZ* cluster will be better understood when frog and chicken *CYP4* clusters are sequenced.

The human *CYP4ABXZ* cluster is not completely assembled (Fig. 4A). In the mouse *Cyp4abx* cluster, note that *Cyp4b1* and *Cyp4x1* are on the outside (Fig. 4B). In human, *CYP4B1* is outside, whereas *CYP4X1* is inside the cluster (Fig. 4A); this may indicate that the human *CYP4ABXZ* cluster in build 33 is not assembled correctly and that the last three genes need to be inverted. Such an inversion would create a more perfect symmetry in the whole cluster and would place *CYP4X1* in a syntenic position, compared with *Cyp4x1* in mouse.

The *CYP4F* cluster and the *CYP4* clan

Mice have nine *CYP4F* genes and humans have six; in both species, these genes are in a coherent physical cluster with a few pseudogenes. In the human, an additional 13 *CYP4F* pseudogenes (most of them copies of exons 6,7,8) have migrated to many locations close to six centromeres in the genome, perhaps reflecting a special mechanism for their dispersal.

The *CYP4F* cluster is unusual in that there is one clear orthologous pair of genes, human *CYP4F22* and mouse *Cyp4f39*, with all other *CYP4F* genes being in a neighboring cluster on the tree (Fig. 1). Aside from this one pair, the human *CYP4F* sequences cluster together (except pseudogene *CYP4F23*), as do the mouse *Cyp4f* sequences. One interpretation of these data is that a single ancestor existed for all these genes, with independent duplications in the two species. *Takifugu rubripes* has a single *CYP4F28* gene, supporting this hypothesis. The *CYP4F22* and *Cyp4f39* genes may have the same function as *CYP4F28* in fish, with new functions having evolved for the other mammalian *CYP4F* genes. The *CYP4F* cluster and the *4ABXZ* cluster (Fig. 1) are joined rather deeply by the two *CYP4V* genes, which are clear orthologs in the mouse and human. A single *CYP4V* is also found in *Takifugu rubripes*, along with one *CYP4F* and one *CYP4T* gene; these three genes predate the ray-finned fish divergence from the tetrapod ancestor, i.e. about 420 MYA.

The depth of the *CYP4V* branch in Fig. 1 plus its distinct intron–exon structure, indicate that the *CYP4V* subfamily should have been given family status, resulting in 19 *CYP* families in mammals rather than the current 18. This is an anomaly of the nomenclature system, caused by keeping insect sequences, such as cockroach *CYP4C1*, inside the *CYP4* family. This lack of stringency in the rules for a family boundary, and a reluctance to create new families have led to 'family creep,' stretching the definition of a *CYP* family beyond the original 40% cut-off. This warping occurred before the creation of *CYP* 'clans', higher-order clusters of related families [3]. If the clan nomenclature had been proposed earlier, then the *CYP4* family might have been limited to vertebrates. Instead, the *CYP4* family has grown to be one of the largest P450 families, so that the *CYP4* family is almost synonymous with the *CYP4* clan.

Gene conversion events can erase history in gene clusters

The tree in Fig. 1 is shaded along some branches, to indicate pairs of orthologs between mouse and human. The ortholog assignment in Fig. 1 is based on sequence relatedness and not function. Note that the *CYP11B1* and *CYP11B2* branch does not contain any shaded pairs. In this case, the *CYP11B1* gene of each species is more similar to the paralogous *CYP11B2* gene of the same species than to the orthologous gene of the other species. These sequences have been double-checked to make sure that no mislabeling has occurred, and the functions (and names) have been assigned correctly to these sequences. These genes are adjacent on both the mouse (9.4 kb apart) and human (32 kb apart) chromosomes. This greater similarity between paralogs than between orthologs can be explained by gene conversion

acting on these closely linked genes. Exchanges between the paralogs have homogenized the sequences, so that they are more similar than their true orthologs, (as defined by function). In fact, glucocorticoid-remediable aldosteronism and hypertension is a clinical disease manifested by a chimeric fusion gene formed between the normal human *CYP11B1* and *CYP11B2* genes [1,51].

Gene conversion has also been noted at the human *CYP2D* locus [42] and between the human *CYP2A6* and *CYP2A7* genes [37,52,53]. At the *CYP11B* locus, there are only two genes, so the possibilities for recombination and gene conversion are limited. If one considers the mouse *Cyp2d* cluster, with 13 full-length and 12 partial sequences that could undergo recombination or conversion events, it is not surprising that assignment of orthologs is not usually possible between genes in mouse and human clusters. The one exception from a multilocus subfamily appears to be mouse *Cyp4f39* and human *CYP4F22*, as mentioned above. *Cyp4f39* is on the end of the mouse *Cyp4f* gene cluster (Fig. 4D) and *CYP4F22* is on the end of the human *CYP4F* cluster (Fig. 4C). Somehow, these two genes have escaped gene conversion and remain as unequivocal orthologs. Perhaps these two genes might have a unique structure/function relationship that will not tolerate conversion events, or their chromosomal positions on the ends of the clusters might protect them.

Duplication signatures visible in some clusters

The mouse *Cyp4a* cluster has had a recent duplication of about 100 000 bp; this region duplicated the *Cyp4a12* and *Cyp4a30* genes, which are now named *Cyp4a12a*, *Cyp4a12b*, *Cyp4a30b* and *Cyp4a30a-ps* (the latter became non-functional after the duplication; Fig. 4B). This region also includes two pseudogene fragments, *Cyp4a12a-de5b* and *Cyp4a12b-de2b5b*, which are 100% identical to each other in exon 5 (Fig. 4B). The human *CYP2ABFGST* cluster exhibits mirror symmetry, indicating an inverse duplication [8]. The pseudogene *CYP2A18P* was at the center of this symmetric cluster (Fig. 2A), until it was split in intron 5 by insertion of the *CYP2B6* and *CYP2B7P* genes. This mirror inversion in human does not exist in mouse; instead, there has been a multilocus tandem duplication [35].

CYP genes outside the seven gene clusters

Mouse and human each have 30 *CYP* genes that lie outside the seven gene clusters. (For maps of all human P450 gene locations see the ideograms at http://drnelson.utmem.edu/hum.html) Surprisingly, 28 and 27 of these 30 mouse and human sequences, respectively, have clear orthologs between the two species (27 of these are shown in Fig. 1). The two mouse *Cyp* genes not shown as having orthologs on the tree are *Cyp11b1* and *Cyp11b2*, but as discussed above, this is an artifact of gene conversion, since they do in fact have functional orthologs, *CYP11B1* and *CYP11B2*. Besides *CYP11B1* and *CYP11B2*, the only non-clustered human *CYP* gene without an ortholog is *CYP27C1*, due to gene loss during the restructuring of chromosomes in mice and rats, as described above. There is one more *CYP* gene with an ortholog in the mouse than in humans (28 vs. 27) because mouse *Cyp2ab1* has only a pseudogene ortholog in humans, and this human pseudogene was not included in the tree.

Inside the seven clusters, there are only seven clearly orthologous pairs: *Cyp2f1*/*CYP2F2*, *Cyp2g1*/*CYP2G2P*, *Cyp2s1*/*CYP2S1*, *Cyp2t4*/*CYP2T2P*, *Cyp4b1*/*CYP4B1*, *Cyp4f39*/*CYP4F22*, and *Cyp4x1*/*CYP4X1*. Interestingly, of these seven, five are located at one end of a gene cluster. Except for the *CYP4F* orthologous pair, these genes are all the sole members of their respective subfamilies. The fact that these genes have been conserved more than other non-orthologous members of the clusters suggests functional constraints. The sharp dichotomy between orthologous pairs and non-orthologous genes in clusters might be related to substrate specificity: the orthologous genes are predicted to be highly specific for endogenous substrates, whereas the non-orthologous genes are more likely to act on foreign substrates.

Similar to the *CYP4* family, the *CYP2* family has undergone 'family creep.' *Cyp2ab1*, *CYP2W1*, *CYP2U1*, *Cyp2ac1-ps* and *CYP2R1* are all deep branches in this family (Fig. 1), meaning that these genes are evolutionarily older, or they have mutated more rapidly, than the rest of the *CYP2* genes. Should these all be *CYP2* family members? *CYP2* genes typically have nine exons, and the intron–exon boundaries are preserved in location and phase; *Cyp2ab1* and *CYP2W1* have this same intron–exon structure. *Cyp2ac1-ps* is a pseudogene in mouse and human (*CYP2AC1P*), but functional in rat, also with the typical 9-exon structure.

On the other hand, *CYP2U1* and *CYP2R1* have five exons each – not the usual nine. Both *CYP2U1* and *CYP2R1* have introns that are equivalent to introns 2, 6 and 8 of all other *CYP2* genes and therefore must have shared a common ancestral gene that had been partially processed to remove introns 1, 3, 4, 5, and 7, and was then reinserted into the genome. Since then, both *CYP2U1* and *CYP2R1* have independently acquired one more intron. The *CYP2U1* intron 3 has a unique GC donor–splice-site boundary in both mouse and human, which is supported by ESTs from both species [BX354123, BX498753]. Both genes have *Takifugu rubripes* orthologs, and are thus more than 400 MY old. The *Takifugu rubripes CYP2U1* and *CYP2R1* introns are in the same places for both genes, so the loss of introns

occurred before tetrapods diverged from ray-finned fish. Although *CYP2U1* and *CYP2R1* have a different intron–exon structure from other *CYP2* genes, they fall within the *CYP2* cluster on phylogenetic trees, and they clearly are derived from the *CYP2* family. Based on these considerations, it seems best to keep them in the *CYP2* family.

Conclusions

Nearly complete sets of CYP gene sequences now exist for mouse, human, *Anopheles gambiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Takifugu rubripes*, *Ciona intestinalis*, *Arabidopsis thaliana*, *Oryza sativa* (rice), and several fungi. The raw data are available, or will soon be available, to trace the history of this enzyme family in animals back to our urbilaterian ancestor. Presumably, the more ancient the CYP gene, the more fundamental will be its role. The flip side of the coin will be to understand the present-day collection of CYP genes, especially in mammals. This includes determining the functions of the orphan CYP enzymes not yet assigned a biochemical role. Having access to mouse, zebrafish, *Drosophila*, *C. elegans*, and even tunicates as experimental models will greatly aid in this process. For example, morpholino antisense experiments in zebrafish may offer the best model system in which to establish the developmental and tissue-specific roles of some CYP genes, since the developing tissues in transparent embryos of this vertebrate can be observed directly. Gene knockouts and knock-ins, and 'humanization' of mice by introducing whole clusters of human CYP genes into a mouse line, will create experimental models that can be manipulated in ways that are not possible in humans.

This review defines the similarities and differences between the complete *CYP* gene sets in mouse and human. This information will be critical in designing experiments and interpreting P450 data from mice, and in extrapolating the results to humans. The 36 orthologous pairs of *CYP* genes (including *CYP11B1* and *CYP11B2* as functional orthologs) will all be suitable genes for study in mice, with direct relevance to human biology; presumably, these genes carry out similar or identical functions in both species. The genes in the seven gene clusters described herein, however, pose serious problems in interpretation, when extrapolating from the mouse to human. Interesting phenotypes have been observed for knockouts of mouse *Cyp* genes in gene clusters; for example, disruption of *Cyp4a12a* causes male-specific hypertension in the mouse (http://bret.mc.vanderbilt.edu/vpsd/cfm_files/view_facname.cfm?KeyNo=179). Even so, the lack of a one-to-one orthologous relationship between the two *CYP4A* clusters makes extrapolation difficult. One cannot predict with confidence the phenotype of a defect in human *CYP4A11*, based on the mouse results. The more radical approach of humanizing the mouse, by replacing the mouse gene clusters with complete human clusters, would at least partially eliminate this problem.

The proposed nomenclature system for pseudogenes addresses an area that is of significance for genome annotation in all species. Within the mammalian *CYP* gene superfamily, pseudogenes are about as abundant as functional genes. We anticipate that this is probably true for mammalian genomes in general, leading to an estimate of perhaps 35 000 to 40 000 pseudogenes in the human genome. Proper annotation of pseudogenes in databases and on genome browsers would greatly aid in the interpretation of BLAST searches, which often show pseudogenes incorrectly assembled as functional genes. We encourage various nomenclature committees to consider the nomenclature system proposed here. The present system of recognizing only one type of pseudogene seems inadequate, based on our experience. We anticipate that a more detailed nomenclature for additional genome features will need to be invented. Although the nomenclature presented here for pseudogenes is not official, we hope it might contribute to further discussions among nomenclature committees, and we look forward to the complete annotation of the mouse and human genomes.

Acknowledgments

We would like to thank Janan Eppig, Judith Blake, Carol Bult and Bob Sinclair (MGI sequencing group) for their helpful criticisms of the manuscript.

References

1 Nebert DW, Russell DW. Clinical importance of the cytochromes P450. *Lancet* 2002; **360**:1155–1162.
2 Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002; **420**:520–562.
3 Nelson DR. Cytochrome P450 gene superfamily. 2003. (web site) http://drnelson.utmem.edu/cytochromeP450.html/
4 Nebert DW, Adesnik M, Coon MJ, Estabrook RW, Gonzalez FJ, Guengerich FP, *et al.* The P450 gene superfamily. Recommended nomenclature. *DNA* 1987; **6**:1–11.
5 Nelson DR, Koymans L, Kamataki T, Stegeman JJ, Feyereisen R, Waxman DJ, *et al.* P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics* 1996; **6**:1–41.
6 Ranson R, Claudianos C, Ortelli F, Abgrall, C, Hemingway J, Sharakhova MV, *et al.* Evolution of supergene families associated with insecticide resistance. *Science* 2002; **298**: 179–181.
7 Nelson DR. Mining databases for cytochrome P450 genes. *Methods Enzymol* 2002; **357**:3–15.
8 Hoffman SMG, Nelson DR, Keeney DS. Organization, structure and evolution of the *CYP2* gene cluster on human chromosome 19. *Pharmacogenetics* 2001; **11**:687–698.
9 Nelson DR. Metazoan cytochrome P450 evolution. *Comp Biochem Physiol Part C* 1998; **121**:15–22.
10 Nelson DR. Cytochrome P450 and the individuality of species. *Arch Biochem Biophys* 1999; **369**:1–10.
11 Nelson DR. Comparison of P450s from Human and *Fugu*. 420 million years of vertebrate P450 evolution. *Arch Biochem Biophys* 2003; **409**:18–24.
12 Xuan Z, Wang J, Zhang MQ. Computational comparison of two mouse draft genomes and the human golden path. *Genome Biol* 2003; **4**:R1.
13 Flicek P, Keibler E, Hu P, Korf I, Brent MR. Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. *Genome Res* 2003; **13**:46–54.

14 Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, *et al*. The sequence of the human genome. *Science* 2001; **291**:1304–1351.
15 Harrison PM, Hegyi H, Balasubramanian S, Luscombe NM, Bertone P, Echols N, *et al*. Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res* 2002; **12**:272–280.
16 Zhang Z, Harrison P, Gerstein M. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res* 2002; **12**:1466–1482.
17 Glazko GV, Nei M. Estimation of divergence times for major lineages of primate species. *Mol Biol Evol* 2003; **20**:424–434.
18 Springer MS, Murphy WJ, Eizirik E, O'Brien SJ. Placental mammal diversification and the Cretaceous-Tertiary boundary. *Proc Natl Acad Sci U S A* 2003; **100**:1056–1061.
19 Oscarson M, Ingelman-Sundberg M, Daly AK, Nebert DW. Human cytochrome P450 (*CYP*) alleles. 2003. (web site) http://www.imm.ki.se/CYPalleles/
20 Nevilie M, Selzer R, Aizenstein B, Maguire M, Hogan K, Walton R, *et al*. Characterization of cytochrome P450 2D6 alleles using the Invader system. *Biotechniques* 2002; **(suppl 34–8)**:40–43.
21 Hirotsune S, Yoshida N, Chen A, Garrett L, Sugiyama F, Takahashi S, *et al*. An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature* 2003; **423**:91–96.
22 Piechaczyk M, Blanchard JM, Riaad-el Sabouty S, Dani C, Marty L, Jeanteur P. Unusual abundance of vertebrate 3-phosphate dehydrogenase pseudogenes. *Nature* 1984; **312**:469–471.
23 Riad-el Sabouty S, Blanchard JM, Marty L, Jeanteur P, Piechaczyk M. The muridae glyceraldehyde-3-phosphate dehydrogenase family. *J Mol Evol* 1989; **29**:212–222.
24 Garcia-Meunier P, Etienne-Julan M, Fort P, Piechaczyk M, Bonhomme F. Concerted evolution in the *GAPDH* family of retro-transposed pseudogenes. *Mamm Genome* 1993; **4**:695–703.
25 Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 1997; **268**:78–94.
26 Hua Z, Zhang QY, Su T, Lipinskas TW, Ding X. cDNA cloning, heterologous expression, and characterization of mouse CYP2G1, an olfactory-specific steroid hydroxylase. *Arch Biochem Biophys* 1997; **340**:208–214.
27 Ding XX, Porter TD, Peng HM, Coon MJ. cDNA and derived amino acid sequence of rabbit nasal cytochrome P450NMb (P450IIG1), a unique isozyme possibly involved in olfaction. *Arch Biochem Biophys* 1991; **285**:120–125.
28 Zhang Z. 2003. http://bioinfo.mbb.yale.edu/genome/pseudogene/human-all/index.html
29 Maltais LJ, Blake JA, Chu T, Lutz CM, Eppig JT, Jackson I. Rules and guidelines for mouse gene, allele, and mutation nomenclature: a condensed version. *Genomics* 2002; **79**:471–474.
30 Finta C, Zaphiropoulos PG. The human cytochrome P450 *3A* locus. Gene evolution by capture of downstream exons. *Gene* 2000; **260**: 13–23.
31 Wain HM, Bruford EA, Lovering RC, Lush MJ, Wright MW, Povey S. Guidelines for Human Gene Nomenclature (2002). *Genomics* 2002; **79**:464–470.
32 Schuetz JD, Guzelian PS. Isolation of CYP3A5P cDNA from human liver: a reflection of a novel cytochrome P-450 pseudogene. *Biochim Biophys Acta* 1995; **1261**:161–165.
33 Meinhardt U, Mullis PE. The essential role of the aromatase/p450arom. *Semin Reprod Med* 2002; **20**:277–284.
34 Hoffman S, Fernandez-Salguero P, Gonzalez F, Mohrenweiser H. Organization and evolution of the cytochrome P450 *CYP2A-2B-2F* subfamily gene cluster on human chromosome 19. *J Molec Evol* 1995; **41**: 894–900.
35 Wang H, Donley KM, Keeney DS, Hoffman SM. Organization and evolution of the *Cyp2* gene cluster on mouse chromosome 7, and comparison with the syntenic human cluster. 2003; *Environ Health Perspect* 2003; **111**:1835-1842.
36 Oscarson M, McLellan RA, Asp V, Ledesma M, Ruiz ML, Sinues B, *et al*. Characterization of a novel CYP2A7/CYP2A6 hybrid allele (CYP2A6*12) that causes reduced CYP2A6 activity. *Hum Mutat* 2002; **20**:275–283.
37 Fernandez-Salguero P, Hoffman SM, Cholerton S, Mohrenweiser H, Raunio H, Rautio A, *et al*. A genetic polymorphism in coumarin 7-hydroxylation: Sequence of the human CYP2A genes and identification of variant CYP2A6 alleles. *Amer Jour Human Genet* 1995; **57**:651–660.
38 Nebert DW, Jorge-Nebert LF. Pharmacogenetics and pharmacogenomics. In: Rimoin DL, Connor JM, Pyeritz RE, Korf BR, eds. *Emery & Rimoin's Principles and Practice of Medical Genetics*, 4th ed. London: Churchill Livingstone; 2002. pp. 590–631.
39 Ingelman-Sundberg M. Duplication, multi-duplication, and amplification of genes encoding drug-metabolizing enzymes: evolutionary, toxicological, and clinical pharmacological aspects. *Drug Metab Rev* 1999; **31**: 449–459.
40 Yu AM, Idle JR, Herraiz T, Kupfer A, Gonzalez FJ. Screening for endogenous substrates reveals that CYP2D6 is a 5-methoxyindolethylamine *O*-demethylase. *Pharmacogenetics* 2003; **13**:307–319.
41 Yu AM, Idle JR, Byrd LG, Krausz KW, Kupfer A, Gonzalez FJ. Regeneration of serotonin from 5-methoxytryptamine by polymorphic human CYP2D6. *Pharmacogenetics* 2003; **13**:173–181.
42 Heim MH, Meyer UA. Evolution of a highly polymorphic human cytochrome P450 gene cluster: *CYP2D6*. *Genomics* 1992; **14**:49–58.
43 Ma J, Bradbury JA, King L, Maronpot R, Davis LS, Breyer MD, *et al*. Molecular cloning and characterization of mouse CYP2J6, an unstable cytochrome P450 isoform. *Biochem Pharmacol* 2002; **64**:1447–1460.
44 King LM, Ma J, Srettabunjong S, Graves J, Bradbury JA, Li L, *et al*. Cloning of CYP2J2 gene and identification of functional polymorphisms. *Mol Pharmacol* 2002: **61**:840–852.
45 Qu W, Bradbury JA, Tsao CC, Maronpot R, Harry GJ, Parker CE, *et al*. Cytochrome P450 CYP2J9, a new mouse arachidonic acid omega-1 hydroxylase predominantly expressed in brain. *J Biol Chem* 2001; **276**:25467–25479.
46 Scarborough PE, Ma J, Qu W, Zeldin DC. P450 subfamily CYP2J and their role in the bioactivation of arachidonic acid in extrahepatic tissues. *Drug Metab Rev* 1999: **31**:205–234.
47 Oleksiak MF, Wu S, Parker C, Karchner SI, Stegeman JJ, Zeldin DC. Identification, functional characterization, and regulation of a new cytochrome P450 subfamily, the CYP2Ns. *J Biol Chem* 2000; **275**: 2312–2321.
48 Oleksiak MF, Wu S, Parker C, Qu W, Cox R, Zeldin DC, *et al*. Identification and regulation of a new vertebrate cytochrome P450 subfamily, the CYP2Ps, and functional characterization of CYP2P3, a conserved arachidonic acid epoxygenase/19-hydroxylase. *Arch Biochem Biophys* 2003; **411**:223–234.
49 Smith SF, Snell P, Gruetzner F, Bench AJ, Haaf T, Metcalfe JA, *et al*. Analyses of the extent of shared synteny and conserved gene orders between the genome of *Fugu rubripes* and human 20q. *Genome Res* 2002; **12**:776–784.
50 Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y. Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Res* 2003; **13**:382–390.
51 Lifton RP, Dluhy RG, Powers M, Rich GM, Cook S, Ulick S, *et al*. A chimaeric 11β-hydroxylase/aldosterone synthase gene causes glucocorticoid-remediable aldosteronism and human hypertension. *Nature* 1992; **355**:262–265.
52 Oscarson M, McLellan RA, Gullsten H, Agundez JA, Benitez J, Rautio A, *et al*. Identification and characterisation of novel polymorphisms in the *CYP2A* locus: implications for nicotine metabolism. *FEBS Lett* 1999; **460**:321–327.
53 Cok I, Aygun Kocabas N, Cholerton S, Karakaya AE, Sardas S. Determination of coumarin metabolism in a Turkish population. *Hum Exp Toxicol* 2001; **20**:179–184.