The cytochrome P450 genesis locus: the origin and evolution of animal cytochrome P450s

David R. Nelson1, Jared V. Goldstone2 and John J. Stegeman2

1Department of Microbiology, Immunology and Biochemistry, University of Tennessee Health Science Center, 858 Madison Avenue Suite G01, Memphis, TN 38163, USA
2Department of Biology, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA

The neighbourhoods of cytochrome P450 (CYP) genes in deuterostome genomes, as well as those of the cnidarians Nematostella vectensis and Acropora digitifera and the placozoan Trichoplax adhaerens were examined to find clues concerning the evolution of CYP genes in animals. CYP genes created by the 2R whole genome duplications in chordates have been identified. Both microsynteny and macrosynteny were used to identify genes that coexisted near CYP genes in the animal ancestor. We show that all 11 CYP clans began in a common gene environment. The evidence implies the existence of a single locus, which we term the ‘cytochrome P450 genesis locus’, where one progenitor CYP gene duplicated to create a tandem set of genes that were precursors of the 11 animal CYP clans: CYP Clans 2, 3, 4, 7, 19, 20, 26, 46, 51, 74 and mitochondrial. These early CYP genes existed side by side before the origin of cnidarians, possibly with a few additional genes interspersed. The Hox gene cluster, WNT genes, an NK gene cluster and at least one ARF gene were close neighbours to this original CYP locus. According to this evolutionary scenario, the CYP74 clan originated from animals and not from land plants nor from a common ancestor of plants and animals. The CYP7 and CYP19 families that are chordate-specific belong to CYP clans that seem to have originated in the CYP genesis locus as well, even though this requires many gene losses to explain their current distribution. The approach to uncovering the CYP genesis locus overcomes confounding effects because of gene conversion, sequence divergence, gene birth and death, and opens the way to understanding the biodiversity of CYP genes, families and subfamilies, which in animals has been obscured by more than 600 Myr of evolution.

1. Introduction

(a) CYP clans, problems in understanding animal CYP evolution, and a way forward

The cytochrome P450s (CYPs) constitute one of the most diverse eukaryotic gene families, with a dizzying complexity within and between species. CYP enzymes use molecular oxygen to modify substrate structure, critical in a huge number of physiological, ecological and toxicological processes. Generally, anaerobes and some microaerophiles, such as the parasites Plasmodium or Giardia, lack CYP genes [1], but in aerobic organisms CYP numbers range from two genes in Schizosaccharomyces pombe to as many as 400 genes in some plants such as potato [2] or grapevine, cottonwood, soybean and rice [3]. Currently, in animals, the range is 35 CYP genes in the sponge Amphimedon queenslandica to approximately 235 in the cephalochordate Branchiostoma floridae (lancelet or amphioxus). CYP genes are classified into clans, families and subfamilies based on phylogenetics as well as sequence identity [4]. Orthologues, co-orthologues and paralogues of CYP genes are often difficult to properly classify as sequences can differ substantially between species, because of indels and gene conversion [5–7]. Likewise, the numbers of genes in a given
subfamily often differ even between closely related species, for example, differences in the mouse, rat and human CYP2D and CYP2J clusters, because of differences in deletion or expansion (or ‘blooms’) in the number of genes ([8,9] see also [10]). This has confounded discerning relationships and especially the path of CYP gene evolution in animals. We address these issues in this study.

Animal CYPs display a molecular phylogeny with 11 distinct clades, each containing one or more CYP gene families. In the late 1990s, reviews of metazoan CYP gene evolution [11,12] introduced the concept of clans to describe these deep branching clusters, although at that time no animal genome sequence had been completed, and several animal CYP families (and clans) were not yet known. By 2003, there were 18 CYP families known in vertebrates, although CYP39 was not yet identified in fish [13]. CYP16 in the CYP26 Clan was the last vertebrate family recognized (D. Nelson 2010, unpublished data). The 11th clan, CYP Clan 74, originally known only from land plants, was first observed in animals as an expressed sequence tag (EST) from lancelet.1 Lee et al. [14] reported the first crystal structure of a plant CYP74, and identified animal CYP74 Clan members in the cephalochordate B. fluviatae, the sea anemone Nematostella vectensis, the coral Acropora millepora and the placozoan Trichoplax adhaerens [14]. This discovery was the crowning touch that led to the 11 animal CYP clans currently known. A list of the 196 animal CYP families and their clan membership can be found in the electronic supplementary material, table S1.

One of the key objectives concerning animal CYP evolution is to understand the origin of these 11 clans. Did the ancestor of all animals have one or more than one CYP gene? How did the 11 clans evolve from the earliest CYP(s)? Comparison of the amino acid sequences does not answer these questions because the branches between the clans are so deep. Only CYP Clans 3 and 4 appear to be sisters, while the relationships among the other nine clans have been elusive. The problem is similar to defining the relationships among the eukaryotic meagroups, unikonts, bikonts and excavaotes2 [17]. Tree-building methods are currently not able to resolve the deepest branches with high confidence [15]. Yoshida et al. [18] showed that eukaryotic CYP51 (sterol 14a demethylase) was probably orthologous to prokaryotic CYP51, and therefore a strong candidate for the original eukaryotic CYP. An alternative view is that bacterial CYP51 resulted from a lateral transfer from plants [19]. This would place the first origin of CYP51 in early eukaryotes, with a subsequent divergence into a cycloartenol branch and a lanosterol branch. This idea is in conflict with Cavalier-Smith [20], who views actinobacteria as the probable precursors of the neomuran ancestor of eukaryotes in part because CYP51 is found in actinobacteria. The idea that plant, animal and fungal lineages all evolved separate CYP collections de novo apparently from a CYP51 was explicitly stated by Nelson [11]. This idea is independent of an actinobacterial or early eukaryotic origin for CYP51.

Although the idea that CYP51 was the starting point has been around for some time, there has not been any consideration of the pathway(s) from CYP51 to the contemporary CYP clans, families and subfamilies. The accumulation of high-quality animal genome assemblies offers a new approach to CYP origins based on synteny. Gene neighbourhoods and gene structure can give additional evolutionary information not contained in coding sequence alone [21–23]. This is illustrated in several studies of genes in specific CYP families [24]. In this study, we examine CYP genes in their syntenic context, to discern the path by which the existing clans, including their families and subfamilies, can be linked to the original locus.

(b) The 2R hypothesis and its implications for CYP evolution

The concept of whole genome duplication (WGD) during animal evolution is an essential background for considering CYP gene evolution [25]. Several gene families provide evidence for two rounds (2R) of WGD in chordates, the most famous example being the clustered Hox genes. Each round of WGD results in the formation of duplicate genes, known as ohnologues [25,26]. Many ohnologues are lost as a tetraploid genome reduces the number of genes back down close to the original diploid number, while some are retained and acquire different functions (subfunctionalization of the original pre-duplicated gene, or neofunctionalization [25,27,28]).

Analysis of the Ciona genome confirmed the 2R hypothesis [29], as did the recent sequencing of the amphioxus (lancelet) genome [23]. Each region in the lancelet genome typically can be mapped in mammalian genomes to four paralogons, large paralogous regions originally derived from duplicated chromosomes ostensibly created in the two rounds of WGD. The ancestral karyotype has been reconstructed with 17 chordate linkage groups (CLGs). The human genome contains 135 segments that map 90 per cent of the genome onto these CLGs [23]. In addition to the 2R WGD, a third round (3R) has occurred in the vertebrate line leading to ray-finned (actinopterygian) fishes including zebrafish and medaka [30,31], but not lobe-finned (sarcopetygian) fishes such as the coelacanth [30].

Evolutionarily, the 2R events have been bracketed in time between the divergence of tunicates from vertebrates and the origin of the gnathostomes (jawed vertebrates). Recent analyses argue that both 2R events preceded the cyclosome–gnathostome split [32,33] (figure 1). Absolute geological dates for the 2R events are not known owing to insufficient fossil evidence. Fossils of chordates (Haikouella lanceolata) and the placozoan Trichoplax adhaerens, hemichordates (Yiannozoon lividum) and vertebrates (Myllkeninna fengjiana and Haikouichthys eraciumensis, now considered a single species) are found in the Chengjiang biota, dated to 525 Ma [37,38]. Therefore, the oldest divergence among deuterostomes (Xenambulacraria3 versus chordates) is older than 525 Ma. The vertebrate in this collection is thought to post-date the origin of the hagfish [37]. If the placement of the 2R events in figure 1 is correct, then the 2R duplications would have to be older than 525 Ma. A firm younger boundary for 2R is the minimum date for the split between the actinopterygians (ray-finned fish) and the sarcopterygians (lobe-finned fish and tetrapods), which was no later than 419 Ma [39]. This date probably should move back to the divergence of sharks from the bony fishes, but sharks do not leave good fossils. The oldest putative shark scales are from the Ordovician Harding Sandstone (approx. 450 Ma [40]). Therefore, the 2R events are at least 450 Ma old and are probably older than 525 Ma. Molecular sequence analysis places the origin of deuterostomes earlier than the fossil evidence, perhaps as much as 643–845 Ma. These estimates require some assumptions about rates of change and choices of calibration points so they are not as firm as fossil dates.
This time frame corresponds to the snowball Earth hypothesis that may have triggered key events in animal evolution [43,44]. It is now possible to determine which genes in vertebrate genomes that are mapped to the CLGs arose by WGD from a common ancestor in the original non-duplicated genome. After each tetraploidization, approximately half of new genes are randomly lost to restore a diploid condition [23], meaning that both copies have been retained for relatively few of the duplicated genes. However, a few CYP genes that were duplicated diverged to create new CYP families and subfamilies. Duplicated genes arising from WGD events are called ohnologues in recognition of Ohno [25–27], who articulated this concept. Large chromosomal segments that contain ohnologues are called paralogons [45]. Vertebrate genomes appear to be composed of four paralogons for each ancestral segment that is preserved from the last common ancestor before 2R.

Deuterostomes and protostomes together form Bilateria, one of the major extant lineages of animals (Bilateria, Cnidaria, Ctenophora, Placozoa and Porifera), although Porifera is possibly paraphyletic [46]. The relative order of animal radiation (phylogenetic topology and rooting) is still a matter of scientific dispute, hindered in part by incomplete sampling, but also complicated by a mixture of short and long branch lengths in the extant lineages [36,47]. There is currently no consensus as to which of the extant basal metazoan lineages (Porifera, Ctenophora or Placozoa) is the earliest diverging metazoan [34,36,48]. The recent results from the Mnemiopsis leidyi genome support an early divergence of Porifera and Ctenophora before the remaining groups (Placozoa, Cnidaria and Bilateria, together called Parahoxozoa; [46]). Continuing genome sequencing from the critical species will probably resolve the question in the near future. However, the current understanding is sufficient for the analysis here.

(c) Use of synteny to track the origins of CYP clans
Orthologous regions of vertebrate genomes often retain the same gene order over long segments. A block of a few genes in the same order is called microsynteny, and shared microsynteny in different species assures a common genomic ancestral gene order is responsible [22,49–51]. In WGD events and segmental duplications, the gene order in both of the duplicate regions initially is identical to the parent sequence. After duplication, many paralogy pairs may lose one copy. When both are retained, over time the syntenic relationships may change as genes move when chromosomes undergo various rearrangements. This makes detection of the duplication difficult. However, the signal is much easier to detect when a duplicated genome is compared with an
unduplicated genome, such as that of the lancelet. Up to four paralogous segments from the 2R WGD events map back to one original locus, and the original gene neighbourhood is revealed. Invertebrate genomes such as *Trichoplax*, sea urchin, sponge and *Nematoscelis* are also unduplicated when compared with vertebrate genomes, so comparison of vertebrate genome regions with these unduplicated genomes can help one to reveal the original gene neighbourhoods for genes of interest. (Note, however, that few insect genes follow the 4:1 rule [52,53]). A recent study on microsynteny in metazoans did identify 795 groups of genes conserved across multiple animal taxa, so the occurrence of these groups is perhaps more common than previously thought [51]. Identification of syntenic relationships can be used to infer evolutionary history of the genes. It is this synteny approach that we apply to analysis of the evolutionary paths of the animal CYP genes, with the goal of defining the gene content of the original loci for each CYP clan. As we argue below, these separate loci can be traced to one original CYP locus.

2. Material and methods

The UCSC browser [54–56] was used to BLAT search [57] for CYP genes and their neighbours in animal genomes including most of the deuterostome genomes on the browser. Special emphasis was given to human, chicken (*Gallus gallus*), anole (*Anolis carolinensis*), frog (*Xenopus tropicalis*), medaka (*Oryzias latipes*), lancelet (*B. floridana*), tunicate (*Ciona intestinalis*) and sea urchin (*Strongylocentrotus purpuratus*). In addition, the JGI genome browsers were used for *Trichoplax adhaerens* [58] and *N. vectensis* [59], and the OIST browser for the coral *Acropora digitifera* [60]. Labelled screenshots of many of the results are provided as electronic supplemental material (figures S1–S37). The genomes used in the analysis of the Bilateria are from deuterostomes; the protostome genomes have very little conserved synteny remaining so they are not useful in this analysis. Data from the ctenophore comb jelly (or sea walnut, *Mnemiopsis leidyi*) genome were provided by A. Baxevanis and C. Schnitzler.

The analysis by Putnam et al. [23] was invaluable for identifying the human paralogs of CYP containing genes or their neighbours. Their supplemental figures S19 and S20 and table S14 were crucial for finding paralogous regions derived from the ancestral CLGs. Nucleotide coordinates from the UCSC Genome Browser were obtained for each human CYP. The segment ID was looked up in table S14 of Putnam, following which figures S19 or S20 were used to find the CLG that matched that segment ID and the other paralogons that matched that CLG [23].

We considered genes on the same chromosome/scaffold with less than five intervening genes to be syntenic, an observation in line with the findings of Irimia et al. [51] that three to 10 intervening genes produced a false discovery rate (less than 0.0002 per gene pair) similar to the less than or equal to four gene distance that they used [51].

Phylogenetic analyses were conducted on datasets consisting of previously published CYP gene sets [61–65], those genomes identified above, and additional complete CYP gene complements from numerous different genomes obtained from JGI (*Helobdella robusta*, *Capitella teleta*, *Lottia gigantea*) and other genome projects (*Amphimedon queenslandica*, *M. leidyi*). CYP sequences were predicted using a combination of hidden Markov model (HMM) searches using HMMer (v. 2.3.2 and v. 3.0b3 [66–68]). Sequences were aligned using *Muscle* (v. 3.8) [69] and masked using custom-written Perl scripts based on the alignment quality scores. Maximum-likelihood phylogenies were constructed using RAXML (v. 7.2.6) [70–73] and presented using *FigTree* (v. 1.3.1). Multiple inferences of the ML tree resulted in the same overall topology; although the short branches (tips) are not stable to multiple ML inferences. This type of dataset is not robust to the bootstrap assumptions; it has characters compatible with, but not informative for a node, large numbers of taxa and potential co-evolving sites [74].

3. Results

(a) Distribution and phylogeny of metazoan CYPs

The numbers of CYPs observed in metazoan genomes ranges widely, from 35 (*A. queenslandica*) to 2367 (*B. floridana*; figure 1). Insect genomes separately display a similar range of CYP numbers (approx. 40–170; R. Feyereisen 2012, personal communication). There is no obvious correlation of relative divergence time with the total number of CYPs, or of CYP families, in a particular genome. Indeed, the size of a gene family undergoing birth–death evolution, as is the case for the CYP superfamily [75,76] see also [10], does not appear to be related to divergence time. Instead, gene family size and distribution appear to be describable by a power-law function (see [10] and references therein). Further work will be necessary to characterize these distributions.

Phylogenetic analyses of large collections of metazoan CYPs support the previous designation of the clan structure (figure 2), and further demonstrate the existence of large ‘blooms’ of CYP families and subfamilies [75]. Certain genomes contain large numbers of CYPs in distinct families and subfamilies that are closely related phylogenetically, appearing as ‘blooms’ of related CYP genes (figure 2). These CYP genes are often the product of tandem duplication and are thus located in close proximity to one another in the genome. Detailed analyses of these regions for several genomes have been previously published (e.g. *Mus musculus* and *Homo sapiens* [9], *S. purpuratus* [61] and *Danio rerio* [63]).

The clan structure in figure 2 is robust to different alignments and protein sequence sets. However, the relative position of specific smaller clans and individual families is sensitive to alignment and does not display high bootstrap values (not shown). Hence, determining the evolutionary origin of CYP clans requires additional information, supplied by the analysis of conserved micro- and macrosynteny.

(b) Ohnologue CYPs

Table 1 lists the CYP gene families arising from the two WGD in the vertebrate line (with human as example), and those appearing in fish following the teleost-specific 3R WGD event [30]. These genes were identified from the location of CYP genes on the human–cephalochordate paralogs that were defined by Putnam et al. [23]. Duplicated genes resulting from these WGD events are in the same clan, and usually within the same family. Therefore, the 2R and 3R events are too recent to account for the origin of animal CYP clans. As detailed below, the CYP clans predate the origin of Cnidaria, and indeed may predate the origin of Metazoa.

(c) Gene neighbours and CYP clans

CYP genes in the 19 vertebrate CYP families were mapped back to their paralogous regions in lancelet using the...
often, these paralogons do not contain any CYP sequences, but they can be used to identify genes that were in the neighbourhood of an ancestral CYP. Over several 100 Myr, a human (or other vertebrate) CYP may not have a recognizable orthologue in a phylogenetically distant genome such as Trichoplax or Nematostella, as the sequences may have diverged too far to be classified even to the same CYP family. This is not necessarily true for genes closely linked to a CYP gene, as they may include highly conserved sequences serving essential functions. Thus, some genes may have highly conserved and clearly recognizable 1 : 1 orthologues, from human to Trichoplax. If a gene linked to a CYP gene in human matches strongly to a single gene in a distant genome, and that gene also is in the vicinity of a CYP gene, then one may conclude that those two CYP genes shared the same ancestral neighbourhood. This is the argument from macrosynteny. If the two CYP genes also happen to be in the same CYP clan, one may conclude that those CYP genes probably share a common ancestor. The probability increases when more than one homologous gene maps to the same region as a CYP gene. The synteny mapping approach did reveal unexpected linkages between many CYP clans as described in detail in the following sections.

Figure 3 shows an example of synteny analysis. The gene ZDHHC12 is found in all 10 species included in that figure. In the cnidarians N. vectensis and A. digitifera, the ZDHHC12 gene is adjacent to Clan 74 CYPs. In the placazoan T. adhaerens ZDHHC12 is one gene from a mito Clan CYP and eight genes from four Clan 74 CYP genes (see the electronic supplementary material, figure S10). The gene pair CRAT/DOLPP1 is adjacent to ZDHHC12 in Acropora and a similar gene association is found in lancelet, chicken and human. These associations suggest that this syntenic region has been conserved over more than 600 Myr of evolution. Other genes in the same neighbourhood provide indirect linkages to other genomic segments. Thus, in Trichoplax and in lancelet, RPL28 is adjacent to FAM20B. FAM20B is 1.2 Mb from a mito Clan CYP in lancelet, and is adjacent to the HOOK1/CYP2J/CYP2N/CYP2P locus in the fish medaka (figure 3).
(CYP440A2) and a CYP2J-like gene in lancelet. JUN/JUNB and HOOK1/2 are also close to CYP2J2 in human and CYP11B2 (a mito Clan gene) in opossum. Nematostella has CRSP8 and BAT2 in this region. These genes also appear in human and chicken. We conclude that these syntenic genes in different species, from mammals to coral, define an ancestral genomic region. This single figure shows three CYP clans (Clans 2, 74 and mito) that we predict were present in this ancestral animal locus.

(d) Extreme age of animal CYP clans

The 11 animal CYP clans (Clans 2, 3, 4, 7, 19, 20, 26, 46, 51, 74 and mitochondrial or mito) are shown in figure 4. Not all animals have all 11. Thus, the Ecdysozoa (insects, crustaceans, nematodes) only have Clans 2, 3, 4 and mito. However, other protostomes may have additional clans; thus, four more clans (7, 20, 26 and 51) appear in molluscs and annelids. Clan 74 was originally described as a land plant CYP clan, but now has been found in cnidarians, Trichoplax and lancelet. This raises the question of Clan 74 origin. CYP74 sequences exist in both land plants and some animals, but not in green algae or fungi. In fact, as discussed below, CYP74 is likely to have originated in marine animals and only later transferred horizontally to land plants. The CYP7 and CYP19 (aromatase) families are chordate-specific, but they are extremely sequence divergent from other CYP clans, which suggests that they are either rapidly evolving or that they may be much older than the chordate line. Gene duplication within chordates would be unlikely to produce new clans, as suggested in the discussion of ohnologues above. CYP39 is considered part of Clan 7, and CYP39-like sequences are known from Trichoplax, sponges and Monosiga brevicollis (a choanoflagellate). (CYP Clan 7 genes are
Figure 4. Distribution of CYP clans in animals and fungi. Presence of a clan is indicated by a closed circle. Absence is indicated by an open circle. The asterisk denotes a probable lateral gene transfer from animals to the filamentous fungi. Some clan losses are evident, particularly in insects and Crustacea (Ecdysozoa), which are known to lack CYP51 and thus require dietary cholesterol. Note also that CYP Clan 19 does not appear until cephalochordates.

observed in some filamentous fungi (electronic supplementary material, figure S23), but the limited distribution may indicate a lateral transfer.) Trichoplax, which is thought to have diverged between cnidarians and bilaterians, has eight clans (it is missing Clans 19, 20 and 46). Clan 46 is found in Nematostella and Clan 20 is found in Nematostella and sponges, which are thought to predate Trichoplax. Therefore, 10 of 11 CYP clans were in existence at least by the time of cnidian origins. The 2R events, which were discussed above, failed to create any new CYP clans, so the clans are older than the 2R duplications. Because two cnidarians are described in the Chengjiang biota including Xiangangia sinica, a primitive sea anemone [38], the CYP clans are presumably older than 525 Ma. Once again, molecular evidence places the origin of Cnidaria much earlier than the Chengjiang biota [41,42].

These facts argue for a relatively rapid origin of the animal CYP clans followed by a much slower divergence within the clans.

The following sections present evidence for the linkage of the CYP clans to each other in a common neighbourhood near the Hox genes before the advent of sponges (Clans 51, 3, 4, 7, 20, 26, mito) or cnidarians (Clans 2, 46, 74). CYP19 does not appear until chordates, but linkage evidence suggests that it was part of the original locus. These sections detail the paths connecting the CYP clans (via families and subfamilies).

(e) The CYP4V and CYP2C connection

Attempts to trace the origins of CYP genes backward in time by mapping the locations of neighbouring genes turned up some unexpected connections between the CYP4s (Clan 4) and CYP2s (Clan 2). The CYP4V2 orthologue in medaka (O. latipes) lies next to TLR3, SORBS2 and PDLIM3. In human, a similar syntenic region contains CYP4V2, FAM149A, TLR3, SORBS2 and PDLIM3 (figure 5, lines H, I). There are three SORBS genes in human and each is adjacent to a PDLIM gene. SORBS1 and SORBS3 are on paralogons as defined in Putnam et al. [23]. SORBS2 is not mapped to a CLG in that paper but it is presumably on a third human paralogon. The human CYP2C gene cluster is next to ACSM6, PDLIM1 and SORBS1 (figure 5, line J). A similar gene arrangement is seen in chicken and lizard (figure 5, lines K, L), though in the lizard the contig ends at HELLs, just before where a CYP2C gene would be expected. Nematostella has a SORBS1- and PDLIM7-like gene pair on scaffold_597 illustrating an ancient association between SORBS and PDLIM genes, but the scaffold is too small to see whether there might be an adjacent CYP gene. Zebrfish have PDLIM1, PDLIM3A and PDLIM3B. PDLIM3A is next to TLR3 and CYP4V2, similar to the medaka. Zebrfish PDLIM1 is 11 genes from ATOH7 and 20 genes (400 kb) from CHUK. Line L in figure 5 shows that in the lizard CHUK is two genes from the SORBS1 PDLIM1 pair, and line M shows that CHUK is 380 kb from the HoxB gene cluster in medaka, and 3.5 Mb from CYP26A1. The close proximity of these CYP genes to PDLIMs and SORBSs indicates the ancestors of CYP4V and CYP2C were close neighbours. The possibility that CYP2C is an ohnologue of CYP4V can be ruled out because the CYP Clans 2 and 4 are older than the duplication of the SORBS1 PDLIM1 gene pair, which probably occurred during the 2R WGD events. This implies that CYP4V and a CYP2C precursor were neighbours before the SORBS1 PDLIM1 gene pair duplicated. Following duplication, one of the CYP genes was lost from each of the descendant regions, separating CYP2C from CYP4V. We will discuss the linkage of these genes with CYP26A1 later.

(f) CYP2D and CYP2K/CYP2W are ohnologues whose ancestor was near the CYP Clan 3 ancestor in early animals

The gene neighbourhoods around CYP2D6 and CYP2W1 in human and some CYP2K genes in fish are descended from a common ancestral region duplicated in a 2R WGD event. Figure 6 shows a map of some of the genes in this region, with ohnologues in the two human regions a and c connected
Figure 5. All 11 CYP clans can be linked directly or indirectly to Hox gene clusters. P450 genes are boxed. Hox genes are aligned down the centre with chr and Mb locations given. Non-Hox regions are labelled as non-Hox. Orthologues and ohnologues are aligned as much as possible to highlight syntenic relationships as seen with EXOC6 in lines M and N. Not all genes in the region are shown because of space limitations.

by lines. Figure 7 shows the paralogons for this region, with connections between ohnologues (electronic supplementary material, figures S1 and S2 show the expanded gene neighbourhood depicted in abbreviated form in figures 6a, c and 7; electronic supplementary material, figures S3 and S4 show these regions in mouse and rat, respectively). These genes will be central to defining the original CYP2 gene neighbourhood in very early animals. Many of these genes appear also in figure 5. The chicken CYP2W1 region is abbreviated in figure 5 as line A and the human CYP2W1 region is part of line B.

The CYP2W1 region is closely associated with the original CYP3A gene location. The evolution of CYP genes in 16 vertebrates has been studied by Qiu et al. [77]. Note that in chicken the CYP3A genes lie between SDK1 and FOXX1. This is the condition seen in opossum, platypus, chicken and lizard. This location is called CYP3HR1 in Qiu et al. [77]. The evolution of CYP3 genes in 16 vertebrates has been studied by Qiu et al. [77]. Note that in chicken the CYP3A genes lie between SDK1 and FOXX1. This is the condition seen in opossum, platypus, chicken and lizard. This location is called CYP3HR1 in Qiu et al. [77]. The evolution of CYP3 genes in 16 vertebrates has been studied by Qiu et al. [77]. Note that in chicken the CYP3A genes lie between SDK1 and FOXX1. This is the condition seen in opossum, platypus, chicken and lizard. This location is called CYP3HR1 in Qiu et al. [77]. The evolution of CYP3 genes in 16 vertebrates has been studied by Qiu et al. [77]. Note that in chicken the CYP3A genes lie between SDK1 and FOXX1. This is the condition seen in opossum, platypus, chicken and lizard. This location is called CYP3HR1 in Qiu et al. [77].

Acropora digitifera (reef-building coral) has a CYP Clan 3 gene next to NPTXR, a gene in human only two genes before SDK1 and FOXX1. This is the condition seen in opossum, platypus, chicken and lizard. This location is called CYP3HR1 in Qiu et al. [77]. The evolution of CYP3 genes in 16 vertebrates has been studied by Qiu et al. [77]. Note that in chicken the CYP3A genes lie between SDK1 and FOXX1. This is the condition seen in opossum, platypus, chicken and lizard. This location is called CYP3HR1 in Qiu et al. [77].

The sea anemone has 19 Clan 3 members. One, in particular, is useful for this analysis because it is adjacent to a small cluster of at least three Clan 2 genes. Trichoplax has 39 Clan 3 members. One 450 kb region has 22 Clan 3 CYPs in several blocks and four adjacent 4CL genes (scaffold_5:880272–1300000), showing that an association between CYP Clan 3 and 4CL has been preserved from Trichoplax to lancelet. The four 4CL genes are bracketed by three CYPs before and nine CYPs after. The last CYP in this large cluster is a Clan 4 member, 15 genes away from the CYP Clan 3 genes. There is a 4CL gene in Nematostella, but it is on the edge of the scaffold and adjacent genes downstream are not in view. The data presented here unite members of the Clans 2, 3 and 4 into a single locus in very early animals. Even today, in animals such as chicken, lancelet and Nematostella, Clan 2 and Clan 3 members are still close neighbours. Direct linkages between CYP genes from different clans near one another on the same chromosome are summarized in figure 8.
(g) Other CYP Clan 4 members

The oldest CYP4 genes in chordates seem to be in the subfamilies CYP4V and CYP4F. A detailed evolutionary study of Clan 4 genes from 28 species has been conducted, noting a major bifurcation of CYP4 subfamilies into CYP4V and related sequences (figure 1, node F, [35]), and all other vertebrate CYP4 subfamilies CYP4A, B, F, T, X and Z [35]. Synteny involving the CYP4V subfamily was discussed above. Here, we give attention to CYP4A, B, F, T, X and Z. The study of Kirischian and Wilson [35] splits this set into CYP4F (node D) and CYP4A, B, T, X and Z (node G). The CYP4F cluster on human chr 19 is on a paralogon with the CYP4ABXZ cluster. Therefore, the CYP4ABXZ and CYP4F gene clusters are possible descendants of ohnologues that were created in the 2R WGDs from a CYP4 gene in the chordate ancestor. This would explain the bifurcation into node G (CYP4A, B, T, X, Z) and node D (CYP4F) in fig. 1 of Kirischian and Wilson [35]. The Kirischian and Wilson study described difficulty in determining the relationships between the CYP4T and CYP4B subfamilies and suggested additional studies would be needed. Examination of the macrosynteny surrounding the CYP4T genes of fish and X. tropicalis supports a syntenic relationship with the mammalian CYP4ABXZ locus, although this is hard to demonstrate conclusively because of much expansion and chromosomal rearrangement in tetrapods. The genes near the fish CYP4T genes are very spread out in human. A comparison of the fugu CYP4T5 genomic region (300 kb) with the human genome shows that every gene in the fugu region matches to a human gene on chr 1, though the human segment is 21.4 Mb (see the electronic supplementary material, figure S36). The size difference of these regions is partly because of the genome size difference, as the fugu genome is 365 Mb, whereas the human genome is 10 times larger, at about 3 Gb [78]. The genes in lancelet surrounding a CYP4T homologue (KIF2A CYP4T PTCH1 MKNK1) also match the human region but in a much narrower 2.4 Mb window. This is consistent with CYP4T being in node G in fig. 1 of Kirischian and Wilson [35] and being an ohnologue derived from a CYP4F ancestor.

The only CYP4F-like gene in Nemastomella is adjacent to CHD7. In lancelet, the CHD7 gene is 375 kb (four genes) from a small CYP2 gene cluster, whereas in chicken, lizard and human CHD7 is near CYP7A1 (as shown for human in the electronic supplementary material, figure S22). The lancelet genome has Clan 2 genes that are less than 100 kb (i.e. two genes SOX5, BCAT1) away from Clan 4 genes (see the electronic supplementary material, figure S57). These neighbour relationships with CHD7 link Clan 4 with Clan 2 and Clan 7 ancestors in early animals.

The ctenophore M. leidyi (comb jelly or sea walnut) has two Clan 4 genes adjacent to one another on AGC01004322.1. Mnemiopsis is a very early branching animal, probably diverging before cnidarians, so it may provide valuable clues to the early neighbours of Clan 4. A Mnemiopsis gene most like CRIP2 is only 3.5 kb upstream of the CYP Clan 4 genes. In medaka, CRIP2 is seven genes from DICER1 and 11–12 genes from AK7 and VRK1. These genes have tight linkages to members of CYP Clans 26 and 46 (figure 9). These observations link Clan 4
Figure 7. Human paralogons covering the CYP2D6 and CYP2W1 regions, including two CYP3 pseudogenes (arrows mark P450s). Probable ohnologues are connected by lines. The regions shown are boxed in red in the ideograms.

CYP Clans

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>mito</th>
<th>74</th>
<th>26</th>
<th>46</th>
<th>7</th>
<th>19</th>
<th>20</th>
<th>51</th>
<th>Hox</th>
</tr>
</thead>
<tbody>
<tr>
<td>chicken</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lancelet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nematostella</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trichoplax</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lancelet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lancelet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>human</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trichoplax</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lancelet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lancelet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>human</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>human</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sea urchin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>human</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>human</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>medaka</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lancelet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>medaka</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chicken</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chicken</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 8. Direct linkage between different CYP clans including Hox clusters (A–D). Each bar indicates a neighbour relationship between the clan members listed on the top of the figure, in the species listed on the left.
to Clans 2, 7, 26 and 46. Section 3h links CYP Clan 4 to CYP Clan 74.

(h) **FAM20, PDRX and CYP Clan 74**

The synteny relationships depicted in figures 3, 5 and 6, and electronic supplementary material, figures S8–S15 support a neighbourhood in the past that included members of Clans 2, 3, 4, 7, mito, 51 and Clan 74. In figure 6, the neighbourhood in the past that included members of Clans 2, 7, 26 and 46. Section 3h links CYP Clan 4 to CYP Clan 74 via IDE and Dicer1 and to the CYP7 and mito Clans via HexB. P450 genes are in shaded boxes. CYP16B1 has a direct linkage to CYP20 in lancelet.

Figure 9. Synteny mapping of CYP16 family members showing linkages to CYP26 family members, the CYP Clan 46 and the CYP Clan 2 via IDE and Dicer1 and to the CYP7 and mito Clans via HexB. P450 genes are in shaded boxes. CYP16B1 has a direct linkage to CYP20 in lancelet.

to Clans 2, 7, 26 and 46. Section 3h links CYP Clan 4 to CYP Clan 74.

(i) **Mitochondrial clan CYPs**

The mitochondrial CYP clan in vertebrates includes the CYP11, CYP24 and CYP27 families. Additional mito CYP families are found in invertebrates, including CYP12 and CYP49 in insects. The mito Clan appears as a monophyletic group descended from one common ancestor in the animals, as no other eukaryotic organisms have mitochondrial CYPs. Mito Clan genes are nuclear-encoded, and contain mitochondrial-targeting sequences and not hydrophobic N-terminal anchor domains. The original founder mitochondrial CYP probably acquired a mitochondrial-targeting sequence early in the history of animals. Of note, however, is that in some species, a few CYPs of more recent origin have bimodal
endoplasmic reticulum and mitochondrial targeting, including CYP1A [79,80] and CYP2E1 [79,81].

Vertebrate CYP11A and CYP11B/11C are orthologues that duplicated during 2R (table 1). These genes are on paralogous blocks defined by the CLK2 and CLK3 genes. CLK2 is adjacent to the CYP11C gene in fish (fugu, zebrfish, medaka, stickleback and coelacanth; see electronic supplementary material, table S2), whereas CLK3 is between CYP11A1 and the CYP11A1–CYP11A2 pair in tetrapods. CYP11B1 evolved from CYP11C1 of fish as shown by the indirect linkage to BAI1: stickleback has two CLK2 genes, CLK2b adjacent to CYP11C1 and CLK2a adjacent to BAI1, while BAI1 is eight genes from CYP11B1 in humans (see electronic supplementary material, table S2). Additional evidence for tetrapod CYP11B being orthologous to fish CYP11C1 comes from the genes DGA1 and SCAMP3. DGA1 is adjacent to CYP11B1 in turtle and frog, and two genes from SCAMP3 in fugu, linking SCAMP3 to CYP11B1. In tetrapods, SCAMP3 is next to CLK2, and, as shown earlier, CLK2 is adjacent to CYP11C1 in fish.

The CYP11-related gene seen in modern lancelet (39% identical to catfish CYP11A1) is approximately eight genes from TNP02. The opossum has a gene block TNP02 ASNA1 HOOK2 JUNB PRDX2 RANSEH2A ATP6VOD1 CYP11B1 that exhibits shared synteny with a similar gene set in lancelet. Therefore, the lancelet ancestor had a CYP11-like gene that is orthologous to vertebrate ohnologues CYP11A and CYP11C/CYP11B. The new synteny evidence thus counters the argument of Markov & Laudet [82] that CYP11 is strictly vertebrate, though the enzymatic activity of this lancelet P450 is not yet known and it may not have side-chain cleavage activity.

The evolution of CYP11 is of critical importance because it is an essential enzyme in the steroidogenesis pathway. Baker [83] discussed the evolution of steroids and their nuclear receptors. CYP11 sequences are rare prior to the emergence of teleost fish, although a fragment of CYP11A is known from little skate (WGS AESE01265227.1 Leucosynia erinacea). Searches for CYP11 in lamprey or hagfish have so far failed to find a candidate. Therefore, the lancelet CYP11 sequence is a key sequence and its function needs to be determined.

Linking various other CYPs to each of these CYP11 paralogues ultimately links lancelet CYP Clans 11, 2, 4, 7 and 74 together. Thus, CYP11A1 is only seven genes from CYP11A1 and CYP11A2 (in Clan 2) on human chr15 and on chicken chr10. In medaka, CYP11A1 is only three genes from HEXB (see the electronic supplementary material, figure S16); HEXB is adjacent to two CYP7 genes in the lancelet, linking CYP7 to CYP11. In opossum, a CYP11B orthologue is three genes from the PRDX2 JUNB HOOK2 locus (figure 3 and electronic supplementary material, figure S17) that is a paralogon of the JUN HOOK1 CYP21 locus (based on data in Putnam et al. [23]).

A CYP11-like gene in the lancelet is four genes away from a Clan 4 member (electronic supplementary material, figure S18). CYP4T1 in zebrfish is adjacent to ARNT on chr16. In medaka, ARNT is three genes from CYP11B (figure 5, line E, chr16). The Nemastoma mito Clan member XM_001640379 is two genes from a SLIT1-like gene. SLIT1 is adjacent to the Clan 74 member CYP440A1v1 in lancelet (see the electronic supplementary material, figure S19). These close neighbours link CYP11 to the CYP genesis locus as well as to genes in the lancelet Clans 4, 7 and 74. CYP27 shows an indirect linkage to CYP20 via the gene PDE1B. CYP20 is six genes from PDE1B in chicken and PDE1B is 3.2 Mb from CYP27B1 in human (figure 5, lines T and W). The CYP51 section below also links CYP27 to CYP51.

(j) CYP51

CYP51 (lanosterol 14a-demethylase; Clan 51) is adjacent to lanosterol synthase (LSS) in the lancelet genome. These enzymes function sequentially in the steroid biosynthetic pathway and thus CYP51 and LSS constitute a functional gene pair. This arrangement is also evident in Trichoplax, which has two CYP51 genes flanking LSS (scaffold 3: 300–323 kb), and also in sponge (contig 13514, 61–65 kb). By using this apparently ancestral link of CYP51 to LSS, we can link other CYP genes through adjacent orthologues and paralogues. Even though CYP51 and LSS are not linked in vertebrates, the association of a CYP gene with LSS ties that gene back to the P450 genesis locus that contained both CYP51 and LSS. For example, CYP27A is 85 kb (seven intervening genes) from LSS in medaka (see the electronic supplementary material, figure S20). CYP27A is the only CYP in a 10 Mb window centred on LSS. This region also contains a Hox cluster (see §3b).

Additional genes linked to LSS or CYP51 provide evidence for their membership in the P450 genesis locus. In sea urchin, LSS is four genes from ZDHHC12, a gene that is adjacent to two Clan 74 genes in Nemastoma (figure 3 and electronic supplementary material, figure S8) and Trichoplax (figure 3 and electronic supplementary material, figure S10). A ZDHHC12-like gene is adjacent to RPL28, FAM20B in lancelet (see figure 4 and §3b). METTL6 is two genes from CYP51 in the lancelet. In Ciona, METTL6 is two genes from B4GALT4, and B4GALT5 is adjacent to PTGIS/CYP8A1 in human, opossum, lizard and platypus. The gene ATOH7 is found in lancelet two genes from CYP51. In medaka, the genes EIF3L, ARNT7, PKD2, PDLIM1 and PRDX3 are found in an approximately 30 gene 763 kb window. EIF3L is seen in figure 6c in the CYPD6 region. Lines F and G in figure 5 have ATOH7 and PKD2 near CYP51 and the HoxB cluster. PDLIM1 is seen in lines H to L, figure 5, near CYP4V and CYP2C genes. PRDX genes were discussed earlier in relation to Clans 74 and 2. The chicken has a direct linkage of CYP51 with the HoxA cluster (figure 5, line D, chr2). The presence of HDAc9 and TWIST1 in chicken also links the CYP51 gene to the HoxA cluster in medaka (figure 5, line E), which has a direct linkage to CYP11B. TWIST1 is one gene away from ATOH7 and three genes from CYP51 in the lancelet (figure 5, line F). These facts support a close neighbourhood of CYP51 with Clans 2, 3, 4, 7, mito and 74 in an early animal ancestor. The presence of PDE11A1 (figure 5, lines E and W) further links CYP20 to this neighbourhood.

(k) Clan 7 (CYP7, CYP8, CYP39) and Hox genes near the ancient CYP locus

CYP7 occurs twice in lancelet between a HEXB pair and VPS24 GOT1 (electronic supplementary material, figure S21). The two CYP7 genes are on opposite strands and both are incompletely known because of gaps in the genome sequence to date. As mentioned earlier, CYP11A1 in medaka is only three genes from a HEXB gene (see the
electronic supplementary material, figure S16). The gene sequence SLIT1 PRDX3 . . . GOT1 VP524 is found at another location in the lancelet. Note that SLIT1 and PRDX1 were associated with a CYP Clan 74 member (electronic supplementary material, figure S19), and many PRDX genes are linked to CYP Clan 2 genes.

Three CYP7 sequences are present in C. intestinalis. CYP7c is adjacent to PIP4K2C, as is CYP7B1 in zebrafish. In human, PIP4K2C is nine genes from CYP27B1 (figure 5, line T, chr 12). The PIP5K2C gene is only three genes from CYP27B1 in medaka. In opossum, a PIP4K2C homologue is next to ARM C3. ARM C1 is next to CYP7B1 in human, chicken, frog and fugu. Thus, CYP7 is linked to CYP27.

Further linkages bring in the Hox genes. In chicken, PIP4K2C is approximately 300 kb from a Hox gene cluster on Chr 27. A Clan 74 gene in lancelet (CYP442A2) is only four genes (300 kb) from the single Hox gene cluster (figure 5, line Y). PNPO is seven genes from the HoxB gene cluster in human and the WNT3, WNT9B pair is another 18 genes from PNPO (figure 5, line G, chr 17). WNT3, PDK1 and PNPO are near the HoxB cluster in medaka with five genes between PNPO and HoxB (figure 5, line M). PNPO is adjacent to CYP51 in the lancelet (figure 5, line F). WNT1 and ARF1 are near the HoxC cluster in medaka (figure 4, line R). (An association between WNT genes and Hox clusters was noted previously by Putnam et al. [23].) Medaka has a HoxA gene cluster on chr 16, 260 kb (11 genes) from TWIST1 (figure 5, line E). LSS, CYP51, PNPO, ATOH7, METTL6 and TWIST1 are all adjacent in the lancelet (figure 5, line F). Another Hox cluster in medaka (chr 19 17.5 Mb) is flanked by WNT3, PDK2, PNPO on the one side and the genes CHUK, CASKIN1, METTL9, KDEL2 and FOXX1 on the other side (figure 5, line M). This last set of genes is found in the paralogons with the CYP2W1/CYP2D6 and CYP3A in some vertebrates (figure 6c). In human and chicken, CHD7 is six and seven genes from CYP7A1 (see the electronic supplementary material, figure S22). CHD7 is adjacent to Nematoselis XM_001639310, a Clan 4 CYP sequence. In lancelet, CHD7 is 375 kb from a CYP2 gene cluster. Additional evidence for Clan 2 linkage to Clan 7 includes VPS24 being adjacent to CYP7 in lancelet. In chicken, VPS24 is adjacent to KDM3A/JMJD1A and about seven genes from EIF2AK3. EIF2AK3 is adjacent to the Clan 7 member CYP7.b in Ciona (electronic supplementary material, figure S4). EIF2AK1 is in the CYP2W1 paralogon (figure 6a), providing an indirect linkage of Clans 2 and 7. These observations link Clan 7 to CYP2W/CYP2D, CYP3, CYP4, CYP27, CYP51, CYP74 and the Hox gene clusters.

The region surrounding human CYP7B1 on chr 8 is on a paralogon with the PDE7B, MYB and EYA4 genes on chr 6 from 133.5 to 136.5 Mb. The human paralogon has the homologues EYA1 and PDE7A found near CYP7B1 (see the electronic supplementary material, figure S22). In Trichoplax. MYB is next to the trox-2 (Gox) gene, the only Hox/ParaHox-type gene in Trichoplax and proposed as a possible ancestral ProtoHox gene [84]. The linkage provides another tie between Clan 7 and the Hox genes.

The reconstructed CLG include all the Hox gene clusters in CLG16 [23]. CYP genes near these clusters would have been duplicated with the clusters during the 2R WGD events. Human CYP27B1 is 4 Mb from the HoxC cluster and CYP27A1 and CYP27C1 are on the q arm of chr 2 with the HoxD cluster, although the CYP27A and C genes and HoxD are spread out over 92 Mb. Human CYP51A1 on chr 7 with the HoxA cluster, but they are on opposite sides of the centromere. No human CYPs remain near HoxB, though a FOXXI homologue is 1.1 Mb from the HoxB cluster on chr 19 in medaka and CYP26A1 is 4 Mb from that same HoxB cluster (figure 5, line M). CYP3 genes are adjacent to FOXX1 in chicken. These facts are consistent with CYP27B being an ohnologue with the CYP27A/C pair ancestor. The lancelet Hox gene cluster has these genes near it: KCNB1, KCNJ4, SOX10, LBE25, LFGN, SLC26A11, HIBADH1, TTTL12, WHRN, with other genes intermingled (figure 5, line Y). Most of these genes are found in figure 5, in the CYP2W/CYP2D paralogons. KCNB1 is next to CYP51A1 in human and HIBADH1 is seen near the HoxA clusters in human medaka and chicken (figure 5, lines B–E).

The CYP8 family genes appear to have originated between lancelet and jawless fishes. No CYP8 genes were found in the lancelet genome, but at least three intron-containing CYP8B-like genes are present in lamprey (Petromyzon marinus). Interestingly, there are no CYP8A-like genes in lamprey, yet bony fish genomes have clearly recognizable CYP8A genes, with introns in the same locations as the introns in the CYP8B-like genes in lamprey. CYP8B genes in fish have no introns, suggesting retrotransposition of ancestral CYP8A mRNAs into early fish genomes. The CYP8B genes of fish are more like the CYP8 genes in lamprey than they are like the CYP8A genes in fish. These data suggest that the original CYP8 gene was CYP8B-like. The retrotransposed gene with no introns retains this character, whereas the intron-containing gene diverged to form the CYP8A ancestor. Thus, CYP8A was derived from a CYP8B-like precursor. The CYP8A gene is near five genes that have clear paralogues near the CYP7A and CYP7B genes in human (see the electronic supplementary material, figure S22). These paralogues are not found near the CYP8B gene, consistent with a retrotransposition event.

Lampreys make petromyzonol and the pheromone petromyzonol 24-sulphate that guides lamprey migration back to their spawning grounds [85]. Petromyzonol 24-sulphate (7α, 12α, 24-trihydroxy-5α-cholan-3-one 24-sulphate) is hydroxylated on the 7, 12 and 24 positions. Therefore, the earliest lamprey CYP8 probably made this pheromone. CYP8A is a prostaglandin 12 synthase (PGTIS). PG12 or prostacyclin appears to be a novel biochemically made only after lampreys split from Euteleostomes (bony vertebrates). It is probable that CYP8 arose from CYP7A by WGD duplication and divergence. Dehal & Boore [29] suggested that regions containing CYP7A and CYP8A were on paralogons or chromosome pieces quadruplicated in the 2R WGD process leading to vertebrates (see the electronic supplementary material, figure S22). This would be a logical way of explaining the origin of CYP8 genes between lancelet and lampreys.

CYP39 is found in the sponge A. queenslandica, indicating that the CYP39 family is very old. Monosiga brevicollis (a single-celled choanoflagellate) also has a CYP39-like gene. If it is correct that CYP39 belongs in Clan 7, then Clan 7 also is very old and probably was present in the CYP genesis locus very early in the history of animals. CYP39A1 is 3 Mb (18 genes) from CYP2AC1P in human. CYP39A1 is only seven genes from the CYP2AC gene cluster in lizard (ranked by MLI1 and RHAG). MLI1 is only eight genes from CYP46A1 in medaka.
CYP7 is found only in chordates and it may have diverged from a CYP39 precursor. There are, however, CYP7-like genes in fungi (see the electronic supplementary material, figure S23), which raises a question of the origin of the fungal CYP7-like genes. Is this a case of convergent evolution or lateral transfer, or is Clan 7 older than Opisthokonta (fungi + animals)? The evidence for CYP7 being linked to the CYP genesis locus argues for Clan 7 originating with animals and not before. We suggest, therefore, that the fungal Clan 7 genes are probably derived from a lateral transfer from animals to fungi, and specifically to some filamentous fungi.

If CYP7 diverged from CYP39 by segmental duplication only in the chordate lineage, then the gene neighbours of CYP7 should be more like the gene neighbours of CYP39. As mentioned earlier, CYP7 genes are adjacent to HERB and Pip4K2C, which associate themselves with CYP11 and CYP27 sequences (both mito Clan members). There is no evidence linking the CYP7 and CYP39 neighbourhoods. This would argue against CYP7 deriving from CYP39 during chordate evolution. An older origin for CYP7 would require many gene losses to produce the current picture.

The CYP39A region in lancelet includes the genes OSCP1, clorf102, CYP39A, NUDC. In human, the OSCP1 gene maps to a region on chr 1 that is a paralogon of the CYP7A1, CYP7B1, CYP11B1, CYP11B2 region of human chr 8. The paralogon relationship shows that CYP7, CYP11 and CYP39 genes were present in a single region in the common ancestor before the 2R WGD event. However, the CYP39 gene is in a different location in human. It should be borne in mind that lancelet has three CYP7 genes. The ancestor certainly had both CYP7 and CYP39 genes, indicating that CYP7 is not the product of the 2R WGD.

The OSCP1 homologue on medaka chr 16 is about 700 kb from a CYP4T12/CYP4T13P gene/pseudogene pair. The only OSCP1 homologue detected in Nematostella is three genes from a CYP3-like gene pair, which would be in Clan 3. This links CYP39 to CYP Clan 3 and CYP Clan 4 in the animal ancestor. Both CYP39 and CYP7 seem to be very old. Both map back to a similar CYP-rich neighbourhood. CYP7 does not appear to be significantly newer than CYP39. This suggests that there was a loss of CYP7 in sponges, cnidarians, protostomes and ambulacarians (hemichordates + echinoderms). This may change as more genome data become available. The possibility exists that CYP7 and CYP39 are in separate clans, and if so, this would make 12 animal CYP clans.

(1) The CYP Clan 26 and the use of retinoids in development

The CYP26 enzymes hydroxylate retinoic acid and thereby inactive retinoid signalling, creating sharp retinoid boundaries crucial in the developing hindbrain [86], spinal cord [87] and retina [88]. The CYP Clan 26, which includes CYP16s, is quite old in animals. CYP16s are found in Trichoplax and cnidarian genomes and even among sponge ESTs (EC374157, Osarella carmela). The CYP16 gene has been lost in mammals and is absent from the zebrafish genome, but found in other fish and many invertebrates (D. Nelson 2010, unpublished data).

Figure 9 shows synteny around the CYP26 and CYP16 genes. The Trichoplax CYP16 gene is six genes from HEXB and is the only Clan 26 member in Trichoplax. As noted earlier, HEXB is adjacent to CYP7 genes in the lancelet and only three genes away from CYP11A1 in medaka (figure 9). Gene order around CYP16 in Trichoplax includes the SLC26A11 orthologue, in the order SLC26A11 CYP16 DCLRE1A ADD2 HEXB. In human, ADD2 is 1.5 Mb from CYP26B1. In lancelet, CYP16B is two genes from DCLRE1A. In lancelet, CYP16B also is 585 kb (10 genes) away from CYP20. The lancelet TIMP1/2 orthologue lies between CYP20 and DCLRE1A, and it is found in the paralogs in figure 7 between CYTH1 and CARD14 on chr 17. TIMP3 is on chr 22 in figure 7. The best hit to the SLC26A11 gene in lancelet is four genes from a CYP2U-like gene in Clan 2. This lancelet region is similar to the single Hox cluster in lancelet and it contains GNA12, ADSL, LFNG, UBE2S and SOX10 (figure 5, lines Y and Z), which is similar to the CYP2W1 and CYP2P6 regions in vertebrates (figure 6c). Furthermore, CYP26A1 is only about 13 genes from the CYP2C cluster in human (approx. 1.6 Mb). Examination of zebrafish chr 17 and zebra finch chr 5 reveals CYP26C1 and CYP46A1 are at opposite ends of a conserved block of genes. The linkages in figure 9 cover sequences in CYP Clans 2, 4, 7, 20, 26, 46 and mito. These associations place CYP26 with other CYP genes discussed earlier, in the CYP genesis locus.

(m) CYP Clan 46

Synteny in the neighbourhood around CYP46 is summarized in figure 10. CYP46 is four genes from SETD3 in human, chicken, anole lizard and frog. SETD3 is only nine genes from CYP26C1 in medaka, two to four genes from AK7 and VRK1 in chicken and six to eight genes distant from AK7 VRK1 in human. As mentioned earlier, a CYP7 gene is adjacent to AK7 in lancelet. VRK1 is only two genes from SETD3 in zebrafish.

In medaka, the CYP46A1 gene is eight genes away from MLIU. MLIU is adjacent to the very large CYP2AC subfamily gene cluster in X. tropicalis. MLIU is also next to CYP2AC1P in human and in some other vertebrates. In Nematostella, there are two CYP46-related genes. They are separated by an ORC5L homologue. The best hit for ORC5L in M. leidyi (a ctenophore) is 800 bp from a CYP Clan 46 gene. ORC5L is 4.4 Mb from the CYP3A cluster in human. MRPL1 is adjacent to one of the CYP46 genes in Nematostella. MRPL1 is four genes from ZDHHC12 in Tetradon nigroviridis (freshwater puffer), and figure 3 shows that ZDHHC12 has many linkages to the 2, 74 and mito Clans. In medaka, CYP11A1 is only 16 kb away from HEXB and ARH1. HEXB is adjacent to two CYP7 sequences in lancelet, and ARH1 is next to a CYP Clan 2 member in the coral Acropora. A CYP46 pseudogene, CYP46A-scl1[12:13:14], is only six genes from the CYP4ABXZ locus on human chr 1. In sea urchin, the sequences GLEAN3_02660 and 15568 are two adjacent CYP Clan 46 sequences. They are also next to a CYP Clan 2 gene, GLEAN3_02658, on scaffold 59540 (see the electronic supplementary material, figure S25). These associations link CYP Clan 46 with the 2, 4, 7, 26, 46 and mito Clans.

(n) ARF gene linkages to CYPs

ARF-like GTPases apparently shared a close association with CYPs in the animal ancestor. ARF and ARL genes are ancient eukaryotic 21 kDa GTPases [89,90]. ARL genes do not appear to have a syntenous link to CYPs. By contrast, there is a frequent association of CYP genes and their neighbours with ARF genes. Animals have three classes of ARF genes (I, II
and III). Phylogenies of ARF proteins predict that the first animals had two ARFs, one class I + II and one class III (also known as ARF6) that gave rise to all animal ARFs [91]. The ARF class (I + II) genes have 12 associations with CYP genes or their close neighbours. These ARF–CYP linkages are described below.

A CYP Clan 46 gene (gw.166.6.1 [Nemv1:11772]) in *Nematostella* scaffold 334 is found only 3.5 kb from an ARF1 homologue. Another ARF1-like sequence in the lancelet is found next to *HHIP1*. Human *HHIP1* is adjacent to *CYP46A1*. This permutation suggests that all three genes were neighbours in a common ancestor. Another ARF1-like gene is only four genes from *CYP51* in *Trichoplax*. An ARF4-related gene in lancelet is adjacent to a Clan 2 member (see the electronic supplementary material, figure S26). In sea urchin, an ARF2-related gene is only two genes away from *HHIP1* homologue. An indirect linkage in Rhesus macaque has ARF6 next to a *PDLIM1* pseudogene (electronic supplementary material, figure S31). As described earlier, *PDLIM1* is near the CYP2C genes, whereas *PDLIM3* is near CYP4V2 orthologues (figure 5, lines H–L). A similar *PDLIM* pseudogene is found in the same place in human. Because this is a pseudogene, and it does not appear to be on a paralogon of *PDLIM1* or *PDLIM2* regions, the linkage to ARF6 could be accidental. Another example of a spurious link between ARF6 and CYP genes is found near the CYP3 cluster in human. Thus, CYP3 is adjacent to a zinc finger protein ZNF498 in many eutherian mammals but not in marsupials, or in other vertebrates. Humans have an ARF6 pseudogene between CYP3A5 and ZNF498. Although this is a very tight association between ARF6 and CYP genes, the pseudogene is found only in great apes: human, chimp and orangutan, not in mouse, rat, dog, cow or horse. This had to be a recent insertion, so it is not relevant to early CYP neighbourhoods. Furthermore, the CYP3 cluster in human has moved from its original location close to the CYP2W1 gene as described earlier. Thus, in some rare cases, linkages of ARF6 genes with CYP genes appear to be chance occurrences, but this does not diminish the strength of most ARF–CYP linkages as evolutionarily conserved.

**Figure 10.** Synteny mapping of the CYP46 clan to CYP7 and CYP26 clans via AK7, to ZDHHC12 via MRPL1 (see figure 2 for ZDHHC12 linkages), to CYP2 and mito clans via STOML1. CYP46 is connected by direct linkages to CYP Clans 2 and 4.

The origin of CYP19 (aromatase) remains a mystery, as CYP19 sequences are distinct from other CYP clans and offer no clues to the parent origin. Aromatase synthesizes...
oestrogen from testosterone via a three-step demethylation [92]. This process is at the end of a long steroid pathway initiated with the CYP51-mediated synthesis of cholesterol that also includes CYP11A1 for side-chain cleavage and CYP17 for 17/20 lyase and hydroxylase activities [93].

A gene duplication to create CYP19 would have had to precede lancelet, which has CYP19. This would be earlier than the 2R WGDs, although, some sequence relatedness would be expected between CYP19 and its parent clan if the duplication were early in chordate evolution. It seems unlikely that a duplication within deuterostomes or even bilaterians would create a new CYP clan, because the 2R WGDs (more than 525 Ma) failed to create even a new family (except apparently CYP8). Therefore, it is plausible that CYP19 was an original member of the CYP genesis locus and it was lost in many lineages just like Clan 7. A mechanism to explain multiple CYP clan losses may require that the lost genes were adjacent in a tight cluster. For example, insects, crustaceans and C. elegans, in the Ecdysozoa, all lack the same seven CYP clans. Possibly, these genes were adjacent and lost together in a block, a difficult hypothesis to test.

The lancelet has three different CYP19 sequences in two different subfamilies (CYP19B1, CYP19C1 and CYP19C2), the subfamilies are 41 per cent identical to each other and both are approximately 40 per cent identical to human CYP19A1. Mizuta & Kubokawa [94] cloned a CYP19C3 orthologue from ovaries of the related species Branchiostoma belcheri, but the gene was not heterologously expressed and assayed. Radioimmunoassay of ovarian steroids did show oestriadiol was being synthesized, but which of the CYP19 genes was responsible is not known. Comparing the CYP19 location in vertebrates and the lancelet shows that they are different, suggesting that CYP19 has moved to a new location in vertebrates, which would not be relevant to connecting CYP19 to the CYP genesis locus. This possibility was a concern of Castro et al. [95] who examined the synteny around the CYP19 genes in vertebrates, but they did not have the lancelet synteny data. The lancelet has kept the original location of CYP19, and the genes surrounding it are informative.

The lancelet region in the electronic supplementary material, figure S32 contains CYP19B1v1 with an adjacent pseudogene, in a region that also has two TGFβ-like genes. The orthologue of one gene (TGFβ1) is found four genes from the CYP2ABFGST cluster in human. The lancelet region also has TLL5 that is adjacent to TGFβ3 in human. TGFβ3 is not near a CYP, but it is on a paralogon to the CYP2ABFGST cluster. TGFβ2 in human is also on a third paralogon to the CYP2ABFGST cluster. TGFβ2 in lancelet is next to CYP4T1.

The lancelet CYP19B1v1 gene is adjacent to a NOP14 gene, and to EXOC6 on the other side. In medaka, there is only one gene similar to NOP14, located four genes from RAD23B. In Ciona, EXOC6 is about 300 kb from a RAD23-like gene on chr 8. A RAD23-like gene occurs in the CYP genesis locus often near PRDX genes (see §30). In human, the NOP14 region on chr 4p is a paralogon of the CYP26B-containing segment on human chr 2 and the CYP26A and CYP26C segment on human chr 10. EXOC6 is also adjacent to the CYP26C–CYP26A1 gene pair in human, chicken, lizard and frog. CYP26A1 is next to EXOC6 in medaka (figure 5, line M), but CYP26C1 has moved to a new location. The presence of genes in line M that are also found in the CYP2W1/CYP3A paralogon ties CYP26 and CYP19 to the CYP2 and CYP3 Clans. After the 2R WGD events, CYP26B was duplicated to form a new CYP26A/C precursor (table 1). The gene EXOC6 was also duplicated next to it and is now EXOC6B. In zebrafish, EXOC6B is adjacent to CYP26B (chr 7 approx. 16.23 Mb; though only the N-terminal of the gene is present in the UCSC browser danRer5 assembly). EXOC6B is adjacent to CYP26B1 in frog, chicken and human. These synteny relationships are indicators that CYP19 and CYP26 were close neighbours in the past.

Ciona intestinalis has EXOC6 adjacent to COG3 on chr 8. COG3 in lizard is next to the SORBS1 PDLIM1 HELLS trio (figure 5, line L). These genes are adjacent to CYP2C in human and chicken (figure 5, lines J, K), but the lizard contig ends here. COG3 matches a gene in lancelet that is six genes from a CYP cluster containing CYP3 and CYP2 sequences (see the electronic supplementary material, figure S6). Further, synteny evidence links CYP19 in lancelet to CYP2UI in human (see the electronic supplementary material, figure S37). The CYP2UI gene on human chr 1 is flanked by PAPSS1 on one side and HADHSC and LEF1 on the other side. These same three genes flank the CYP19C genes in lancelet (see the electronic supplementary material, figure S37). These observations link CYP19 to the CYP Clans 2 and 3.

Additional support is given by the gene LCT, which is adjacent to NOP14 in the lancelet. In medaka, this gene is situated between FAIM on one side and HDAC9 TWIST1 on the other side (see the electronic supplementary material, figure S33, chr 21). FAIM is found in figure 5, line Y next to the lancelet Hox cluster. HDAC9 and TWIST1 are seen in figure 5, lines B and D–F near CYP51, the HoxA cluster in chicken and medaka and the CYP11B gene in medaka. These genes are part of the CYP2WI paralogon on human chr 7.

The gene EML1 is another neighbour of the CYP19B1v1 gene in lancelet (see the electronic supplementary material, figure S32). EML1 is adjacent to CYP46 in human. There is one other location for the CYP19B1v2 gene in lancelet. However, the CYP19B1v1 and v2 genes are 99 per cent identical and probably represent alleles in the genome assembly. Both have tail-to-tail NOP14 neighbours. Other neighbouring genes differ, but there are gaps in the assembly so some genes may be missing. CYP19B1v1 has NOP14 on one side and ABCC4/CFTR on the other side. ABCC4/CFTR is one gene from WNT2 in human. The related gene ABC3 is also found in the CYP2WI paralogon of chicken. These observations have placed the CYP19 ancestor into the same region as precursors of the 2, 3, 4, 26, 46, 51 and mito Clans.

(p) CYP20

The genes flanking CYP20 in lancelet are RPL27, AAAS and SLC11A2 (see the electronic supplementary material, figure S34). AAAS is 800 kb from the HoxC cluster in X. tropicalis and 700 kb from the chr 12 HoxC cluster in human. CYP27B1 is on the opposite side of this HoxC cluster in human (figure 5, line T). The SLC11A2 gene is 900 kb from CYP27B1 on medaka chr 5, but it is only five genes (250 kb) from CYP27B1 in X. tropicalis. Opossum has SLC11A1 on chr 7 380 kb from CYP27A1. In human, SLC11A1 is on chr 2q in the same chromosomal region as CYP27A (385 kb away) and CYP20 (15.1 Mb away). It appears
that CYP27, CYP20 and SLC11A were duplicated during the 2R WGD, and CYP27B became linked to SLC11A2, as seen in medaka and *X. tropicalis*, while CYP27A is linked to SLC11A1. Apparently, CYP20 moved away from the original site and lost the duplicate copy near CYP27B. RPL27 is two genes from RPL3 in zebrafish and it is four genes from a RAC parologue. Both RPL3 and RAC2 are found in the paralogon containing CYP2D6 (figure 7, chr 22). The human chr 7 paralogon includes RAC1, CYP2W1 and the original vertebrate CYP3 locus (figure 7). PDE1A is 300 kb from CYP20 in chicken. PDE1C is 870 kb from CYP8A1 in stickleback, and PDE1B is 3.3 Mb from CYP2B1 in human. The PDE1A,B,C genes in human are on paralogons containing the *Hox* D, C and A clusters, respectively [23].

More ancient neighbours are found in *Nematosella*. There are two adjacent CYP20s in *Nematosella*. The best match to the right-hand-side neighbour (fgenesh1_pm. scaffold_58000009 [Nemwel:229175]) is MCM9 in lancelet. Surprisingly, this is adjacent to a CYP2 gene cluster with nine members (see the electronic supplementary material, figure S35). The gene on the other side of CYP20 in *Nematosella* matches with RAB18. RAB18 is 370 kb from a four gene CYP20 cluster in lancelet (see the electronic supplementary material, figure S34). RAB18 is two genes from ABI1 in medaka, and ABI2 is adjacent to CYP20 in human, and in opossum, lizard, frog and medaka. The medaka RAB18 is approximately 700 kb from PDRX6 and FGGY. FGGY is two genes from CYP2J2 in human. ABI was duplicated in the 2R WGD events and CYP20 followed ABI2. This shows that RAB18 to ABI2/1 tracks the CYP20 trajectory from *Nematosella* to human. RAB18 is close to ARM4 in medaka, chicken, opossum and human. In *Ciona*, ARM4 is only 90 kb from a pair of CYP4 genes on chr 9. CYP20 is 1.6 Mb from Dicer1 in medaka. Dicer1 is three genes from CYP26C1 in zebrafish (figure 9). CYP20 is 685 kb from CHD7 in fugu. CHD7 has links to CYP7 in vertebrates (electronic supplementary material, figure S22) and to Clan 2 genes in lancelet. Through these interactions, the CYP20 gene is also associated with the 2, 3, 4, 7, 26 and mito Clan neighbourhoods.

![Image](http://rstb.royalsocietypublishing.org/)

4. Discussion

The current abundance of animal genomes and the concept of CYP clans, inferred from deep branching in molecular phylogenies [4,96], have made it possible to trace linkages among the CYP families and clans, through analysis of syntenic relationships. The nature of relationships among the CYP clans has been difficult to determine from phylogenetic trees, except for Clans 3 and 4, which consistently cluster together, suggesting that they shared a common ancestor. CYP Clan 3 and Clan 4 members are present in cnidarians and even in sponges (figure 4), thus the divergence of these two clans occurred very early in animal evolution. In fact, figure 4 shows all the clans except Clan 19 have sequence evidence indicating that they were present at least by the origin of cnidarians. In most cases, however, the percent identity of sequences in different clans is in the low 20 per cent range, and thus comparison of the CYP sequences alone does not permit further insights into evolutionary relationships.

The examination of synteny allows new relationships to be detected based on shared neighbours. Applying this approach to members of all 11 CYP clans uncovered heretofore unknown linkages among CYP clans. Some clans have more detectable linkages than others, but in the end all 11 clans can be tied to each other through shared linkages. As an illustration, CYP20 is not linked to CYP7, but it is linked to CYP27 in the mito Clan. CYP11 in the mito Clan is linked to CYP7, so CYP7 is linked indirectly to CYP20. Our analysis indicates that all the CYP clans shared one common neighbourhood, which we call the cytochrome P450 genesis locus. This has never been hinted at before and it is a novel result of the synteny analysis.

Some of the genes in this neighbourhood that we call the CYP genesis locus are quite well known, including the *Hox* gene cluster and several *Wnt* genes. The proximity of the *Wnt* genes to the *Hox* genes has been noted before [23]. Other genes also occur close to the genesis locus. Thus, an *ANTP* megaclass consisting of *Hox*, *ParaHox* and *NK* genes has been proposed [97]. The sponge genome still maintains an NK gene cluster with six genes [98], but there are no *Hox* or *ParaHox* genes in this genome. Although not described in §3, further linkage of the sponge NK genes to CYPs is provided by the KIF11 gene that is adjacent to the NK5/6/7b gene in sponges (ACUQ11000781) and is also adjacent to CYP26C1 in zebrafish (figure 5, line P) and very close to CYP46 in lancelet (figure 9). Because there are CYP genes in the *A. queenslandica* sponge genome but no *Hox* genes, the CYP gene cluster appears to predate the *Hox* gene cluster that seemingly arose from an NK gene. Other sponge genomes may hold additional clues to the genesis locus gene environment.

A CYP3 Clan member (XM_002107799.1, 39% identical to CYP3A4 human) is 10 genes from an NK2 gene cluster in *Trichoplax* (from 4.32 to 4.36 Mb on scaffold 1), suggesting the ancestral CYP gene cluster was linked to the NK gene cluster.

Together, the Wnt, *Hox*, NK and CYP genes represent a developmental locus in early animals (figure 11). An NK-like homeobox domain is found approximately 14 kb from a CYP46 Clan member in the ctenophore *M. leidyi*. EXOC6 is 13 genes from another *Trichoplax* NK cluster containing NK3, NK6 and *Hex*. EXOC6 has been linked to CYP19 and CYP26. The forces acting to keep the NK genes, the *Hox* genes, and *Wnt* genes together may also have kept this block of CYP genes from dispersing after their formation through *cis* duplication. Studies on fish *ParaHox* gene clusters support the existence of interdigitated control regions in the clusters that select against their loss [99]. Irimia *et al.* proposed the existence of Genomic Regulatory Blocks, with transcriptional enhancers for developmental genes being embedded in neighbouring genes, thus keeping them together [51]. Furthermore, the gene functions may interact. For example, retinoic acid regulates *Hox* gene expression, and CYP26 is a retinoic acid hydroxolation enzyme that controls retinoic acid concentration. Other CYP substrates/products may influence *Hox*, *ParaHox*, NK or *Wnt* gene expression.

Consequences of gene clustering may also involve gene losses. The loss of seven CYP clans in insects, crustaceans and nematodes (all ecdysozoans) may have occurred by a block deletion of the clustered CYP genes. Such a major loss of CYP genes all at once could be expected to have profound effects on ecdysozoan development. It has been known since the time of Geoffroy de St Hilaire that the arthropod dorsoventral axis is inverted compared with vertebrates, indicating a major developmental shift in this group [100–102]. Could it be related to the simultaneous loss of so many CYP genes?
The fact that all CYP clans can be traced back to this neighbourhood has other implications as well. The CYP7, 19 and 74 Clans appear to originate in the genesis locus and not from outside it. Otherwise, it would not be possible to find linkages to the locus. This is illustrated by vertebrate CYP19, which apparently moved to a new location from that seen in the lancelet. The vertebrate location of CYP19 is unrelated to the genesis locus and cannot be traced back to it. If the CYP19 Clan existed in the genesis locus as predicted, its function would probably not be the same as modern CYP19 aromatase enzymes. The novel aromatase biochemistry of CYP19 emerged in chordates, as is seen in the lancelet. Evidence for CYP11 and CYP19 sequences in lancelet was presented earlier, and steroid nuclear receptors are not found before the chordates [93].

An animal origin for CYP74 is surprising, since the first CYP74 genes all were found in land plants, however, CYP74s are not found in algae or fungi. The dating of early metazoan divergences is a controversial subject; for this discussion, we use dates taken from Berney & Pawlowski [113] [103] and from Blair [41] and Hedges et al. [42]. The earliest appearance of the CYP74 Clan is in Trichoplax, thought to have arisen between 733 and 592 Ma. This predates the appearance of land plant fossil spores at 475 Ma [104,105]. How did a CYP74 gene make the transition from marine animals to land plants? Presumably a vector would be required, such as a bacterium that could obtain the CYP74 gene from marine animals. Indeed, two CYP74-like genes have been identified in bacteria (see [14] electronic supplementary material, figure S18). These are Methyllobacterium nodulans and Methyllobacterium sp. 4–46, plant symbiotic bacteria involved in nitrogen fixation in legumes, ideal candidates for gene transfer to plants. Lateral gene transfer to early land plants is known to have occurred for T globins [106] and for glutamine synthase IIβ [107]. Two additional examples that are involved in CYP pathways are PAL and 4CL. The first gene in the general phenylpropanoid pathway of plants, phenylalanine ammonia lyase (PAL) was acquired by a horizontal gene transfer from soil bacteria early during land colonization [108]. The gene 4-coumarate coenzyme A ligase (4CL), found next to CYP Clan 3 genes in lancelet and Trichoplax, is the third gene of the plant phenylpropanoid pathway [109], acting after CYP73A1, a CYP that encodes cinnamate 4-hydroxylase [110]. The 4CL genes were thought to be plant-specific rather like CYP74, but as mentioned above they exist in the lancelet, Trichoplax and Nematostella, as does CYP74 [14]. Analysis of the purple sea urchin detected two more of these 4CL genes [111]. A bacterial version of 4CL has been found in the soil bacterium Streptomyces coelicolor [112]. This raises the possibility that 4CL may have a similar history to PAL and CYP74.

CYP7 and CYP19 are chordate-specific. CYP7 and CYP39 are currently placed in the same clan, but CYP39 sequences are detected in the choanoflagellate M. brevicollis, sponges and Trichoplax, while CYP7 is not seen until lancelet. Possibly, CYP7 and CYP39 should be placed in different clans; the relationship between them is fairly weak. This would allow CYP7 to be newly evolved. However, CYP Clan 7 members can be linked to the genesis locus, suggesting they are in fact very old as well. A more recently derived chordate-specific CYP Clan 7 gene such as CYP8 should show a strong resemblance to another Clan 7 CYP, and it does. This argues against a de novo appearance of CYP 7 only in chordates. Instead, we suggest there have been gene losses from Clan 7 in pre-chordate animals. A similar argument applies to CYP19.

The analysis here suggests that early appearance of CYP clans could have resulted from tandem duplication of a progenitor CYP. The time window of animal origins appears to be fairly limited, perhaps 632–863 Myr [103–114]. Thus, establishing the order in which CYP clans arose via tandem duplication of an initial CYP gene may be difficult. A Clan 4 member appears in M. brevicollis, though this may be the precursor of the Clans 3 and 4. A CYP39-related gene (7 Clan) is seen in M. brevicollis, so the Clans 4 and 7 may have been very early members in the genesis locus. Monosiga brevicollis also has several CYPs most like plant CYP710/CYP61, CYP704, CYP711 and CYP745, gene families that we presume were lost on the way to animals. The ctenophore M. leidyi has the first known 20 and 46 Clan members. The placement of sponges and placozoans in evolution is not certain, but we assume in figure 4 that sponges were earlier (see [35]). Sponges have the first observed 26 and mito Clan members. Trichoplax has the first Clan 2 and Clan 74 members. All clans except 19 predate the origin of Cnidaria. The CYP Clan 19 has no sequence evidence until lancelet, but for reasons mentioned earlier, it must have a much older origin because of linkage with many other CYP clans in addition to its very divergent sequence. Gene loss has certainly occurred in some lineages and this may confuse the interpretation of first appearances. Additional genomes, especially from sponges, may help to more precisely clarify the origin.
of some CYP clans/families such as CYP2, CYP7, CYP19, CYP46 and CYP74.

Earlier attempts to discern the evolution of animal CYPs strongly suggested that all animal CYPs derived from a single CYP51 gene, based largely on the presence of a CYP51 in bacteria, plants, fungi and animals [11,18,115,116]. It has been argued that the bacterial CYP51 may have arisen from a lateral transfer from plants, but this does not alter our discussion here on animal P450s [19]. One could envision that, as a source of sterols important in cell membrane function, CYP51 could be an essential early enzyme. On the other hand, the catalytic function of extant CYP51s is sufficiently specialized that one could also argue that some other less specialized function may have characterized the progenitor CYP. Identifying a precursor to the progenitor CYP would help to determine its role and identity. Regardless of the identity of the CYP occupying the initial locus, the recognition of the CYP genomic locus as the starting point for animal CYP evolution opens the way for more detailed analysis of each clan to understand the origin of specific families. As more genome assemblies become available, reconstruction of ancestral genomes becomes possible as already achieved with the CLGs. The history of each CYP is really a vector through time that tracks gene duplications and losses, lateral transfers, WGDs and chromosomal rearrangements. Armed with more and more detailed genome histories, we will be better able to trace individual gene histories. The limiting factor becomes what time has erased, leaving no trace for us to read.

5. Definitions

Microsynteny: The preservation of gene order on a nearly 1:1 basis without large insertions, gaps or rearrangements.

Macrosynteny: The retention of orthologues in large chromosomal regions at higher than statistically expected levels. These genes are not necessarily close together and they may be spread over large regions, as in electronic supplementary material, figure S36 where 10 orthologue/co-orthologue pairs are spread over a 21 Mb region.

Co-orthologues: Tandemly duplicated genes that arose from a single orthologous ancestral gene. The current loci in two different species may not contain the same number of members, but each is a co-orthologue of the other species orthologous gene(s). Example: the CYP4T gene in fish and CYP4A, B, X and Z genes in human.

Paralogons: Multiple regions created by whole genome duplication (WGD). The two WGD events in chordates resulted in four paralogons for each original segment in the pre-duplicated protochordate genome. Originally, these were equivalent to whole duplicated chromosomes, but over time they become smaller segments owing to chromosomal rearrangements.

Ohnologues: Genes duplicated in a WGD event that survive to the present day. Most ohnologue-duplicated genes get lost as the tetraploid genome reduces the number of genes back down close to the original diploid number. Some are retained and acquire new functions.

CYP clan: A CYP clan is a clade of genes. Clans are relatively deep branchings on phylogenetic trees. There are eleven total clans of animal CYPs but only four clans in most arthropods and nematodes and five in ticks. Plants have a different set of CYP clans, except for CYP51 and CYP74 that are in common.

Unikonts: A proposed eukaryotic taxonomic group defined by the presence of only one flagellum; includes opisthokonts and amoebozoas.

Bikonts: A eukaryotic taxonomic group defined by cells with two emergent flagella; includes Archaeplastida (plants and relatives), Excavata, Rhizaria and Chromalveolata.

Excavates (Excavata): A deep branch of eukaryotes containing many parasitic, often ‘amitochondrial’ species; includes Giardia, Trypanosoma, Euglena, Trichomonas.

This work was supported in part by NSF grant no. IOS-1242275 to D.R.N. This study was supported in part by a NOAA Sea grant no. NA10OAR4170086 (to J.J.S. and J.V.G.), NIH grant nos. SROI-ES015912 (J.J.S.) and F32-ES012794 (J.V.G.), a Superfund Basic Research Center grant no. 2P42ES07381 (J.J.S.), and by the Ocean Life and Deep Ocean Exploration Institutes of the Woods Hole Oceanographic Institution. The authors have no competing interests to declare.

Endnotes
1[GenBank:BI386673 dated 20 May 2003]. This sequence was posted to the Cytochrome P450 Homepage in 2005, though it was not recognized as a CYP74 Clan member at that time.
2Unikonts, bikonts and excavates are three monophyletic divisions of eukaryotes determined by multi-gene tree building methods, using large concatenated sequence alignments [15]. Morphologically unikonts (animals, fungi, Amoebozoa) and bikonts (plants, alveolates, stramenopiles, Rhizaria) have one or two flagella at a specific point in their life cycle. Excavates have a characteristic ventral feeding groove on their cell surface (examples are Jakobids, Euglena, Giardia, trypanosomes). Cavalier-Smith argues that excavates are bikonts [16].
3Xenambulacraria = hemichordates + echinoderms + xenoturbellarians.
4The 236 P450 gene count for Branchiostoma (lancelet) seems high compared to other chordates. The paper of Putnam et al. [23] argues for a fairly strict paralogon relationship to vertebrates, so the vertebrate P450s should have an ohnolog precursor in lancelet. A large expansion is not expected. Part of the increase may be due to alleles being counted as different genes. The P450s of lancelet may not contain the same number of genes but only four clans in most arthropods and nematodes and five in ticks. Plants have a different set of CYP clans, except for CYP51 and CYP74 that are in common.
5RPL28 occurs in two different genomic locations in lancelet shown in figure 3.
Fungi have about 15–20 clans that are not completely defined yet. CYP51 and CYP7 clan members are seen in fungi. The distribution of 7 Clan members is only found in some filamentous fungi strongly suggesting a lateral transfer from animals to fungi. Therefore, it is shown as an asterisk in figure 4. The other fungal clans are distinct from the animal clans.

References


and phylogeny of the earliest vertebrates. 


100. Lacalli TC. 1995 Dorsoventral axis inversion. Nature 373, 110–111. (doi:10.1038/373110c0)


