Diversity and evolution of the P450 family in arthropods

Wannes Dermauw a,1, Thomas Van Leeuwen a, René Feyereisen a,b,*

a Laboratory of Agrozoology, Department of Plants and Crops, Faculty of Bioscience Engineering, Ghent University, Coupure Links 653, 9000, Ghent, Belgium
b Department of Plant and Environmental Sciences, University of Copenhagen, 40 Thorvald tegades vej, DK-1871, Frederiksberg C, Copenhagen, Denmark

1 Corresponding author. Laboratory of Agrozoology, Department of Plants and Crops, Faculty of Bioscience Engineering, Ghent University, Coupure Links 653, 9000, Ghent, Belgium.
E-mail address: rene.feyereisen@gmail.com (R. Feyereisen).

ABSTRACT

The P450 family (CYP genes) of arthropods encodes diverse enzymes involved in the metabolism of foreign compounds and in essential endocrine or ecophysiological functions. The P450 sequences (CYPome) from 40 arthropod species were manually curated, including 31 complete CYPomes, and a maximum likelihood phylogeny of nearly 3000 sequences is presented. Arthropod CYPomes are assembled from members of six CYP clans of variable size, the CYP2, CYP3, CYP4 and mitochondrial clans, as well as the CYP20 and CYP16 clans that are not found in Neoptera. CYPome sizes vary from two dozen genes in some parasitic species to over 200 in species as diverse as collembolans or ticks. CYPomes are comprised of few CYP families with many genes and many CYP families with few genes, and this distribution is the result of dynamic birth and death processes. Lineage-specific expansions or blooms are found throughout the phylogeny and often result in genomic clusters that appear to form a reservoir of catalytic diversity maintained as heritable units. Among the many P450s with physiological functions, six CYP families are involved in ecdysteroid metabolism. However, five so-called Halloween genes are not universally represented and do not constitute the unique pathway of ecdysteroid biosynthesis. The diversity of arthropod CYPomes has only partially been uncovered to date and many P450s with physiological functions regulating the synthesis and degradation of endogenous signal molecules (including ecdysteroids) and semiochemicals (including pheromones and defense chemicals) remain to be discovered. Sequence diversity of arthropod P450s is extreme, and P450 sequences lacking the universally conserved Cys ligand to the heme have evolved several times. A better understanding of P450 evolution is needed to discern the relative contributions of stochastic processes and adaptive processes in shaping the size and diversity of CYPomes.

1. Introduction

CYP genes encoding P450 enzymes are consistently found among the largest gene families in plants, animals and fungi. These enzymes are at the interface of environmental responses, metabolism and endocrine regulation by catalyzing the transformation of a myriad of exogenous and endogenous substrates by hydroxylation, epoxidation, dealkylations and a great variety of other reactions. With 160,387 sequences, the P450 domain is the 18th largest of 18,259 proteins families in Pfam (Pfam.xfam.org/family/PF00067, accessed July 12th, 2020). Although P450 sequence diversity is considerable, with no universally conserved residue over their ±500 amino acid length (Sezutsu et al., 2013), the increasing number of crystal structures reveals a conserved three dimensional architecture (Werck-Reichhart and Feyereisen, 2000; Guengerich et al., 2016; Di Nardo and Gilardi, 2020). A key element of P450 enzymes is their heme-centered active site, with a Cys axial ligand, which makes them heme-thiolate proteins, predominantly active as monooxygenases, despite their traditional designation as cytochromes. By activating molecular oxygen and their ability to insert an oxygen atom in a great variety of substrates, P450 enzymes have been called nature’s blowtorch. P450 sequences are found in virtually all organisms including some viruses, with a few exceptions such as Escherichia coli. Even before the avalanche of sequences from genome projects, a nomenclature scheme was designed with CYP gene family and subfamilies distinguished by their percentage sequence identity, 40 and 55% respectively (Nebert et al., 1991). With the increasing number of sequences, and therefore CYP families, a higher level of nomenclature, the CYP clan, was introduced (Nelson, 1998). Broad overviews of P450 diversity revealed a non homogeneous distribution of sequences in CYPomes, with many families with few genes and few families with
many genes. This pattern, not specific for the P450 family, is seen in
plant, animal and fungal CYPomes (Feyereisen, 2011; Sezutsu et al.,
2013). Until recently, it appeared that only four clans of P450s were
represented in insects (Feyereisen, 2006, 2011), far less than the ten
clans of human P450s. Yet, while arthropods constitute the largest
group of animals on earth, there has been no arthropod-wide study of the
P450s, their diversity and evolution. The increasing number of available
arthropod genomes, including several non-insect genomes, therefore
calls for an arthropod-wide study. Moreover, the error-prone genome
annotations of gene families (Tawal et al., 2014), and the limited use of
approved nomenclatures creates a risk of cascading errors and mis-
conceptions in databanks and literature alike.

Here we provide a very large collection of curated arthropod CYPomes. It includes full CYPomes for 31 species as well as 10 partial
CYPomes assembled from transcriptome shotgun assembly (TSA) data or selected from additional species including the arthropod sister group
Onychophora, bringing the total number of sequences to 2942. In
arthropods, P450 enzymes perform multiple functions, from basic physi-
ological processes such as hormone biosynthesis to synthesis of species-
specific pheromones, and from metabolism of host-plant chemicals to
detoxification of pesticides. Only limited information is currently
available on which P450 performs which function, and developing
predictive tools must start with a better understanding of P450 diversity.
We show large variations in the number of P450 genes in each genome
(the CYPome), as well as large variations in the composition of each
CYPome over the 500 MY of arthropod evolution.

2. Methods

2.1. Data mining and P450 sequence curation

CYPomes were manually curated, and sequences were assigned a
name using the approved CYP nomenclature (David R. Nelson, Univ. of
Tennessee). Some gene names may have changed since publication of the
original CYPome descriptions. When available, the latest NCBI
RefSeq release was used (see Supplementary Table S1). Major discrep-
cyencies were solved by a parsimony approach using transcriptome data,
alignments with related sequences in closely related species as well as
conserved intron position and phase. Webscipio (www.webscipio.org;
Hatje et al., 2011) was helpful for the latter. Official CYP names are
associated with the most accurate accession in NCBI, with an “m”
suffix when the NCBI sequence was modified by manual curation. CYP9P2
of honeybee is an example of an intronless gene, cut by RefSeq
(XP_006562368.1) with one part spliced with another gene, while the
second part is not called. Yet there is full TSA support (HP578228) for
the manually curated sequence. Another example, not uncommon in
NCBI, is the fusion of two adjacent P450 genes, as CYP9S1 and 9R1 in
the honeybee (XP_026300962.1), that we cut into its two genes, both
fully supported by TSA. We use mN and mC as suffixes for two sequences
that are fused in the RefSeq release. In some taxa, (e.g. Protura) no
genome is available, so TSA were mined extensively for P450 sequences,
and the data were curated for removal of obviously unprocessed introns.
In a few cases, TSA data was limited or not available and de novo tran-
scriptomes were assembled from raw sequencing data using CLC Ge-
nomics Workbench 10 and default settings for de novo transcriptome
assembly (see Supplementary Table S2). Sequences mentioned in the
text that do not appear in Fig. 1 (and related Supplementary Fig. S1,
Supplementary File S1) are listed in Supplementary File S2. Pseudogenes
and partial genes (gene fragments and “loose exons”) were not retained
for our analysis. Some pseudogenes have a single stop codon defect in
their orf, and this stop codon is also found in TSA, so this would be
considered a “recent” pseudogenization, for instance CYP6A58P4 in
Manduca sexta (Supplementary File S2). Although it is recognized that
“recent” pseudogenes or “loose exons” can be used in gene conversion
events with neighbouring genes, it is difficult to estimate a reasonable
cutoff for consideration, so no pseudogenes were used in our study.

2.2. Species studied

We selected phylogenetically representative species in the major
lineages of arthropods, and whenever possible from high quality ge-
nomes backed up by transcriptomes. Because the genome coverage
of arthropods is very unequal, this may have resulted in higher quality
insect CYPomes than from some other groups. Supplementary Table S1
gives the species studied, their taxonomical position as well as the as-
sembly and release on which the CYPome annotation was based. For
little studied groups, where we present a single CYPome from a single
species, when possible we also annotated a subset of genes from a second
species, e.g. from the CYP2, CYP20 and mito clans to gain a better
confidence in the conclusions about presence or absence of genes (col-
lemobolans, copepods, scorpions). In some cases, CYPomes presented
were obtained from mining of TSA (Zygentoma, Protura, Ostracoda,
Diplopoda), or partial CYPomes were obtained from lower quality ge-
nomes (Arachnogbatha, Diptura). We also used a small collection of
“founder” genes, i.e. the first representatives of a particular CYP family,
when these genes were not in the CYPomes covered (Supplementary
Table S5). Since the nomenclature is based on percentages of identity,
the position of these “founder” genes in the phylogeny helps describe the
sequence space of a CYP family. Other genes of known physiological
function and particular interest that are not part of our complete
CYPomes were also added for reference to help in the discussion
(Supplementary Table S5). The number of estimated species in each taxon
listed in Supplementary Table S1 was obtained from (www.catalogueof
life.org) last accessed March 4, 2020, in the “estimated number of
species known to taxonomists” category.

2.3. Maximum likelihood phylogenetic analysis

Of the 2942 curated CYPs (Supplementary File S1), four CYPs (Sinella
curviseta CYP3728A1b, Mesobuthus martensi CYP3001AF1b, Centru-
roides sculptatus CYP41T5b and CYP307Mb) were identically dupli-
cated and, as advised by the IQ-TREE 1.6.10 software (Minh et al.,
2020), were removed from further analysis. Eighteen L. migratoria
sequences of an improved annotation were obtained after completion of
our analyses and are not included in the tree. The CYP sequences were
aligned using the online version of MAFFT version 7 (Katoh et al., 2019)
and the “G-large–INS-1” setting (“mafft –reorder –globalpair –large
input”). Maximum likelihood (ML) phylogenetic analysis was conducted
on this alignment using IQ-TREE 1.6.10 (Minh et al., 2020) on X-SEDE
on the CIPRES web portal (Miller et al., 2018). In a first run, the
following settings were used: LG + I + G4+F as the best-fit substitution
model (determined by ModelTest-NG (Darriba et al., 2020) and using
the Akaike Information Criterion), 3000 ultrafast (UF) bootstraps
(Hong et al., 2018) and default settings for reduction of overestimation
of bootstrap support (–bnni) and minimum correlation coefficient (–bcor)
for UFBoot convergence criterion (“ –in file.txt -bb 3000 -bnni 3000
-bcor 0.99 -st AA -m LG + I + G4+F -nt AUTO”). In a second run, a
Shimodaira-Hasegawa like approximate likelihood ratio test (SH-aLRT)
with 3000 replicates was computed on the consensus tree (“contree.tre”)
resulting from the first run (“ –in file.txt -te contree.tre -alrt 3000
-st AA -m LG + I + G4+F -nt AUTO”). We also ran an IQ-TREE ML analysis with non-parametric (standard) bootstrap support (SBS). First, thirty inde-
dependent bootstrap runs each with 10 replicates were performed using
the same alignment and substitution model as in the first run and with
“maximum number of iterations to stop (–nm)” set to 1000 (“ –in file.txt
-bo 10 -st AA -nm 1000 -m LG + I + G4+F -pre boot X -nt AUTO”, with X
= 1–30). Next, the 300 bootstrap trees were combined (“alltrees”), a consensus tree (“alltrees.contree.tre”) was created (“–con –alltrees”) and
branch lengths of the consensus tree were estimated using the
original alignment (“–te alltrees.contree.tre -in file.txt -st AA -m LG + I
+ G4+F -nt AUTO”) as outlined in the FAQ of the IQ-TREE website
org/doc/Frequently-Asked-Questions). Finally, SBS values were added
to the consensus tree containing UF bootstrap and SH-alRT support using TreeGraph 2 (Stover and Müller, 2010) (see http://treegraph.bio infweb.info/Help/wiki/Tutorial:Adding_support_values). The tree with UF/SH-alRT/SBS values was midpoint rooted using SeaView version 4.0 (Gouy et al., 2009) and uploaded to Iroki (Moore et al., 2020) for tree decoration and customization. Orthologous CYP groups were identified as those branches with UF/SH-alRT/SBS values higher than or equal to 95/80/65, and containing at least three 1:1 orthologs in three or more species (with the exception of the 408 B-358 A group, of which only two representatives (Locusta migratoria CYP408B1 and Pediculus humanus CYP358A1) were included in our phylogenetic analysis but which have 1:1 orthologs in other species, see section 3.7 (Supplementary File S2).

2.4. Maximum likelihood phylogenetic analysis of CYP16s and CYP307s

Twenty-four metazoan CYP16 were aligned with eight CYP26/CYP120 protein sequences and 118 arthropod CYP307 were aligned with four arthropod CYP306 sequences. (Supplementary File S2) using MAFFT version 7 (Katoh et al., 2019) with the G-INS-I settings (“–reorder–maxiterate 1000 –tree e–globalpair input”). Phylogenetic trees were constructed in an identical way as in the section above, using the LG + I + G4+F and LG + I + G4 as best fit substitution model for ML analysis of CYP16 and CYP307, respectively. The resulting CYP16 and CYP307 trees were outgroup rooted using CYP26/CYP120 and CYP306 sequences (Supplementary File S2), respectively.

2.5. Evolutionary relationships and distance between arthropod species

Evolutionary relationships and distance between arthropod species of which the complete/partial CYPome was investigated in this study were derived from different studies (see Supplementary Table S4 for detailed information on divergence times and references).

2.6. Subcellular localization of arthropod CYPs

Subcellular localization of the curated arthropod CYPs was predicted using DeepLoc 1.0 (Armenteros et al., 2017), LocTree3 (Goldberg et al., 2014) and SubCons (Salvatore et al., 2018), while the presence of an N-terminal mitochondrial transit peptide was predicted using TargetP 2.0 (Armenteros et al., 2019).

3. Results and discussion

3.1. Overview

We studied the CYPomes of 40 arthropod species, including 31 species for which we annotated or reannotated the CYPomes from genomic data. These include 22 hexapods with 18 insect species, 6 crustaceans, 2 myriapods and 10 chelicerates. We have tried to cover major insect pests and vectors of disease, and important model species. This sampling is biased by the availability of genomic data, but we have attempted to balance phylogenetic diversity with species abundance. Orthologous CYP groups were identified as those branches with UF/SH-alRT/SBS values higher than or equal to 95/80/65, and containing at least three 1:1 orthologs in three or more species (with the exception of the 408 B-358 A group, of which only two representatives (Locusta migratoria CYP408B1 and Pediculus humanus CYP358A1) were included in our phylogenetic analysis but which have 1:1 orthologs in other species, see section 3.7 (Supplementary File S2).

Fig. 1. Six CYP clans in arthropods. Maximum likelihood tree of 2920 P450 sequences from arthropods. (6 in clan 16, 19 in clan 20, 306 in mitochondrial clan (M), 733 in clan 2, 778 in clan 4 and 1078 in clan 3). An asterisk indicates bootstrap support values (UF/SH-alRT/SBS) that are higher than 95/80/65. The full tree in svg format with bootstrap support for each branch is provided as Supplementary Fig. 1. The sequences are listed in Supplementary File S1 with their CYP name while CYP info (accession numbers, taxonomic position, clan, and subcellular localization) can be found in Supplementary Table S5.

Whenever possible we provide matches between NCBI accessions and CYP nomenclature names, as it can be difficult to navigate through the haphazard way P450 genes are named in NCBI. The manual curation revealed about 27% error rate in the sequence of the RefSeq calls, a known problem (Fawal et al., 2014), and a very small minority of sequences with a correct CYP name. Our FASTA file (Supplementary File S1) may thus serve as reference for work on arthropod P450s. We also provide a list of CYP families by CYP clan, as a resource for future work on arthropod P450 (Supplementary Table S3).

3.2. CYPome sizes

Fig. 2 (Supplementary Table S1) gives the CYPome size of the species used in this study. For the species with a full genomic-based CYPome, the tomato russet mite Aculops lycopersici has the smallest CYPome with just 23 genes, and the collembolan S. curviseta has the largest with 261 genes. Full length pseudogenes (17) are not included in this count in S. curviseta. Within insects, Nicrophorus vespilloides had 129 P450s, and the body louse 36. Literature data indicate that the fig wasp Ceratosolen solmsi, an extreme specialist, carries 34 P450 genes (Xiao et al., 2013) with much larger numbers in mosquitoes (196 in Culex quinquefasciatus, Yang and Liu, 2011; 186 in Aedes albopictus, Chen et al., 2015). Beyond insects, the salmon louse Lepeophtheirus salmonis, a copepod ectoparasite of fish, has just 21 P450 genes (Humble et al., 2019) while Campodea augens (Diplura) has 202 (Manni et al., 2019). In this study, there was no relation between CYPome size and genome size (n = 31, r = 0.212), or between CYPome size and total gene number (n = 31, r = 0.080).

Although partial comparative surveys are published in some genome papers, the numbers given can be approximations and the criteria used to obtain them are not always stated (e.g. full-length curated sets or Pfam hits ? with or without pseudogenes ? were gene predictions run with and without RepeatMasker in order to identify and recover all tandemly duplicated P450 genes ?). This might lead to fluctuations that...
obscure the underlying biology. For instance *Drosophila sechellia* which recently (about 3 MYA) diverged from *D. melanogaster* and is now adapted to feed on the otherwise toxic *Morinda citrifolia* fruit, has been reported to have 14 P450s fewer than *D. melanogaster* by manual annotation (Good et al., 2014) or 9 more by automatic prediction (Rane et al., 2019). Yet the latter study, based on a rigorous methodological pipeline claims to give results within 2% of the data obtained by manual annotation, which is not the case for this classical example. Here we give numbers for full-length genes, and distinguish them from clear pseudogenes, or incomplete partial genes and loose exons, so that our manual curation is consistent. Our CYPome sizes are generally in agreement but marginally lower than those of Rane et al. (2019)(Supplementary Fig. S3), with the notable exception of the damselfly *Calopteryx splendens* for which Rane et al. reported four times as many P450s. Unfortunately, most genomes are incomplete so it can be difficult to determine the nature of the sequences that we did not include. Our CYPome sizes are therefore a minimum, close to the actual number for high quality genomes, and possibly slightly underestimated for others. Some genes may have a portion of the sequence missing because of a rearrangement (i.e. a type of pseudogene) and some genes are truncated by a gap in the genomic sequence (i.e. a sequencing artefact). In some cases, the latter therefore a minimum, close to the actual number for high quality genes. P450 numbers for full-length genes, and distinguish them from clear pseudogenes in one genome can be pseudogenes in another genome of the same species. This may affect total gene numbers. Perhaps an extreme case is the fall armyworm *Spodoptera frugiperda*. The genomes of two strains (corn and rice) were sequenced, with CYPomes sizes of 117 and 135 respectively (unpublished results). In contrast, deep transcriptomes of *Helicoverpa armigera* and *H. assulta* yielded 169 vs 155 P450 genes respectively (Li et al., 2013). Even if these numbers differ from those of the reference genome, they are not always taken into account. The CYPome size of the highly polyphagous mite *Tetranychus urticae* is about double that of the oligophagous *T. evansi* or monophagous *T. luteolus* (unpublished results). In contrast, deep transcriptomes of *Helicoverpa armigera* and *H. assulta* yielded 169 vs 155 P450 genes respectively (Li et al., 2013).
from ours (for *H. armigera*), there is little reason to believe that the authors of that study used vastly different methods and counting criteria for the two species. Therefore there is probably not a very large difference in CYPome size between those very close species, which are known to hybridize. Yet *H. assulta* is an oligophage on Solanaceae while *H. armigera* is notoriously very polyphagous. In both strains of *S. frugiperda*, as well as in the *H. armigera/assulta* pair, a closer examination of the differences in CYPome sizes would be of great interest.

### 3.3. Clans represented

Until recently insect CYPomes were thought to be comprised of sequences from just four clans the CYP2, CYP3, CYP4 and mitochondrial clans (Feyereisen 2006, 2012), but our survey now widened to arthropods establishes the presence of P450s from two additional clans, the CYP16 and CYP20 clans (Fig. 1). CYP clan diversity in arthropods is therefore much lower than in plants (11 clans) fungi (at least 17 clans) or vertebrates (11 clans) (Nelson, 2018). Arthropods are lacking the conserved CYP51 clan (sterol 14-demethylase) and are known to be staro heterotrophs, but CYP51 is present in the sister clade Onychophora (Nelson, 2018). A CYP51 sequence found in the collemobolan *Fotonia candida* (GASX02003270.1) appears to be an algal contaminant (Helicospordinum KDD76879.1).

All six arthropod clans are well supported: CYP2 (100/98/68), CYP3 (100/100/74), CYP4 (98/99.7/75), CYP16 (100/100/100), CYP20 (100/100/100) and the mitochondrial clan (100/100/91) (Fig. 1). Only six genes from three CYP families in *Metaseiulus (Galenordus) occidentalis* (CYP3104-6) fall outside this clan framework, and form a single clade (100/100/100) between the CYP2 and CYP16 clans, so they are outliers that we included in the CYP2 clan.

The relative proportion of clans is quite variable when arthropods are considered (Fig. 2, Supplementary Table S6) with apparently a shift from CYP2 clan predominance in some chelicerates, crustaceans and collemobans towards the CYP3 and CYP4 clans in insects. These major shifts in the proportion of P450s in various clans illustrate the dynamics of gene births and deaths over evolutionary time affecting the whole CYPome. The number of families per clan may be taken as one measure of diversity and the four major clans (CYP2, 3, 4 and mito) currently comprise 95, 156, 101 and 44 families respectively (Supplementary Table S3). However, some families such as the CYP4 family are very rich in subfamilies, so that diversity reaches as expected much deeper than the family cutoff (40% sequence identity). While in vertebrates the number of CYP families has “saturated” with few new families expected to be named from newly sequenced genomes (Nelson, 2011), this is clearly not the case in arthropods where the annotation of new CYPomes brings many new CYP families particularly in lesser studied branches of the phylogeny as shown here with 42 new families in *S. curviseta* alone.

Our phylogenetic analysis is generally very supportive of the CYP nomenclature, in that most families and subfamilies cluster together. However, this is not always the case, with a few families or subfamilies that are dispersed or simply “contaminated” with non homonymous sequences. This should not be too worrisome, as the CYP gene nomenclature is a tool to facilitate scientific exchange and not a substitute for phylogenetic analysis. We can identify at least two causes for the discrepancies. On one hand, CYP family and subfamily designations are based on percentage identities (40 and 55%). A rapidly evolving duplicate may therefore “escape” its original CYP name. Almost trivial contaminants with non homonymous sequences. This should not be too worrisome, as the CYP gene nomenclature is a tool to facilitate scientific exchange and not a substitute for phylogenetic analysis. We can identify at least two causes for the discrepancies. On one hand, CYP family and subfamily designations are based on percentage identities (40 and 55%). A rapidly evolving duplicate may therefore “escape” its original CYP name. Almost trivial contaminants are known for the CYP4 family, where the first insect representative, CYP4C1 was placed in the CYP4 family, initially of vertebrates, and therefore artificially bloated the size of the CYP4 family.

#### 3.3.1. CYP clan 20

P450 sequences of the CYP20 clan are widely but not uniformly distributed in animals. The function of CYP20 remains enigmatic and even the function of the human CYP20A1 gene is still unknown. The presence of a CYP20 homolog in the damselflyy, *C. splendens* was first reported by Ioannidis et al. (2017). We mined genomes and transcriptomes for CYP20 sequences, and found them in various, but not all taxa of Chelicerates, Myriapods and Pancrustacea (Supplementary File S2).

Among Chelicerates, CYP20 sequences are found in ticks, but not in mites, except orbibatid mites. CYP20 are found in some Aranea, Acanthocurria genilicata (as intronless gene) and Latrodectus geometricus, but not in Parasteatoda tepidariorum. They are also found in all scorpions and horseshoe crabs (Xiphosura). CYP20 sequences are found in myriapods, both Chilopoda (Strigimurus maritimus and Scalopendra dehaani) and Diplopoda (Chamberlinius hauienensis, Glomeris pastulata, Polydesmus angustus, Helicorthomorpha holstii and Trigoniulus corallius). In “Crustacea”, CYP20 sequences have a very patchy distribution, perhaps owing to the limited genomic coverage (when compared to insects). They are found in copepods, *T. japonicus* and Paracyclopina nana, in Gammaerus fossoram (amphipod) and in barnacles (Cirripedia) but not in *Hyalella azteca* or *Daphnia pulex*. They are also found in Branchiura and Ostracoda. We did not find CYP20 in Collembola, but they are found in Diplura and Protura. In insects, CYP20 are restricted to Archaeognatha, Zygentoma, Odonata and Ephemeroptera (Fig. S3). In most cases, CYP20 is a single gene, but there are six genes in the centipede *S. maritima* and two in the millipede *P. angustus*. The CYP20 clan P450s appear to be polyphyletic in arthropods because the gene tree does not follow the species tree. The patchy distribution of CYP20 homologs indicates multiple losses in arthropods.

The CYP20 sequences differ from usual P450 sequences (Fig. 3). The typical FxxGxxC Cys pocket motif has a codon insertion between Phe and Gly, and in orbibatid mites there is a further codon insertion between Gly and Cys. Pro always follows the Cys codon and while this is not unique to CYP20 sequences, it is relatively uncommon. In Cirripedia, the CYP20s have lost the conserved Cys altogether.

#### 3.3.2. CYP clan 16

The CYP16 family was the last new family described in vertebrates (Nelson, 2011) and related sequences have only recently gained the status of clan, distinct from the related CYP26 clan (D.R. Nelson and J.V. Goldstone, personal communication). No function has yet been described for CYP16 clan P450s. The related but distinct CYP26 clan includes retinoic acid hydroxylases, and the related cyanobacterial CYP120A1 is also a retinoid acid hydroxylase (Alder et al., 2009). We found sequences belonging to the CYP16 clan in all major lineages of arthropods, but their distribution was very patchy (Fig. 4). Xiphosura is the only lineage of Chelicerates where we found CYP16 sequences, in *L. polyphemus* and Carcinocoris rotundicaud. In myriapods, there is a CYP16 clan sequence in *S. maritima* (but no evidence in the TSA of *S. dehaani*, GBIM01, and no evidence in the millipedes *C. hauienensis* (TSA) *H. holstii* or *T. corallius* (genomes, JAAFCF01 and JAAFCF01). We also found CYP16 sequences in some copepods and decapods, as well as barnacles and oysters. The presence of CYP16 was restricted in insects to Archaeognatha (*Machilis hrabei*) and some Zygentoma (partial sequence found in *Atelura formicaria*) (Supplementary File S2). In all cases, there was a single CYP16 sequence in each species, indicating that despite the multiple losses in arthropods, the CYP16 homologs are highly conserved and do not appear to bloom as seen in all other clans. The arthropod CYP16 clan sequences are rather peculiar, in that they lack the C-helix WxxR motif. They have a typical HWHGXxxSXP
motif in the helices F-G loop region and have an unusual and shifted I helix motif ALxPKxxASxLTS. However, they have a highly conserved ETLR and SRS5 motif ExxRxxPPPxxGGRxR, similar to CYP26 clan structures. Most notably, the arthropod CYP16 clan proteins lack the typical N-terminal membrane-anchoring sequence.

Fig. 4. Maximum likelihood phylogeny of the CYP16 clan. The CYP16 sequences (Supplementary File S2) are indicated with a bordeaux color, while the CYP26/120 outgroup is shown in black. Fully supported clades are marked by an asterisk. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

3.3.3. Mito clan

The mito clan is generally small, down to 4 sequences in the collembolan S. curviseta, but can be larger than the CYP2 or CYP4 clan as in the cat flea Ctenocephalides felis where mito clan P450s comprise 30% of the CYPome. Sequences belonging to this clan are considered to be located on the inner membrane of mitochondria, as are the vertebrate members of this clan (CYP11, CYP24, CYP27). This clan, rather than being called CYP11 clan, has therefore the mitochondrial designation. Strong evidence that house fly CYP12A1 is indeed a mitochondrial enzyme comes from its localization by immunogold histochemistry and absolute dependence on mitochondrial electron donors adrenodoxin reductase and adrenodoxin (Guzov et al., 1998). Such stringent criteria have not been applied to other members of this clan in arthropods, and assumption of mitochondrial localization is generally accepted, although Drosophila CYP314A1, 302A1 and 315A1 tagged and expressed in S2 cells colocalize with a mitochondrial marker by confocal microscopy (Petryk et al., 2003). CYP12A1 is predicted to be targeted to mitochondria by DeepLoc, LocTree3 and SubCons. However, this is not the case for all mito clan P450s. Of 306 mito clan P450s in our study, while 97.4% were predicted to be mitochondrial by LocTree3, only 81.4% and 72.9% were predicted to be mitochondrial by DeepLoc and SubCons respectively (Supplementary Table S5). The case of CYP314A1 is of interest because it encodes an ec dysone 20-hydroxylase which is known to be found in both mitochondria and microsomes, depending on the species, the tissue and the developmental time (Lafont et al., 2012). The Bombus mori 20-hydroxylase in eggs is microsomal and inhibited by (microsomal) NADPH cytochrome P450 reductase antibodies (Horike and Sonobe, 1999), and so is the larval cockroach enzyme (Halliday et al., 1986). It is not known whether these enzymes are encoded by CYP314A1, which is predicted to be located in the ER, not mitochondria by SubCons and DeepLoc in B. mori and L. migratoria. While LocTree3 predicted all 38 CYP314A1 to be mitochondrial, this was only the case for 12 by DeepLoc and just 9 by SubCons. The “mitochondrial” in the name of this clan should therefore not be taken as definitive evidence of subcellular localization. Some P450s of this clan may be microsomal or have dual localizations.
3.3.4. CYP clans 2, 3 and 4: the major arthropod clans

Fig. 2 shows that these three clans are predominant in arthropod species, with the CYP3 and CYP4 clans generally predominant in insects. In the spider mite, the CYP2 clan includes the most CYP sequences, but it is much reduced in the tomato russet mite A. lycopersici with just one sequence, while Dermatophagoides pteronyssinus and Varroa destructor have just three and four CYP2 clan sequences, respectively. The distribution of genes within the CYP3 clan can change dramatically in less than 100 MY (e.g. in bees, Megachile rotundata and Apis mellifera), and result in presence or absence of subfamilies associated with particular detoxification capacities (Hayward et al., 2019). In Lepidoptera, Calla et al. (2017) showed a scatter (sic) plot with data from seven species to conclude that “it is evident from the data that CYPome size in Lepidoptera is directly correlated with the size of the CYP3 clan”. This correlation, based on their data, is not significant ($r = 0.6101$, $n = 7$, $P = 0.146$). A careful analysis and presentation of the data for 19 species shows that the CYPome size of Lepidoptera is dependent on both the CYP3 and CYP4 clan sizes (Fig. 5).

The CYP4 clan genes are highly abundant in arthropod genomes, but paradoxically, one of the least studied in insects (Feyereisen, 2006). This remains the case. The size of the CYP4 clan is reduced in several lineages. While the honey bee and other Apoidea have just four CYP4 clan genes (Claudianos et al., 2006; Beadle et al., 2019), there are only two CYP4 clan genes in Tyrannus japonicus and none reported in the salmon louse L. salmonis (Humble et al., 2019). The TSA of another coperop P. nana apparently lacks CYP4 clan genes (Han et al., 2015) and so does the TSA of Heterocypris incongruens (Ostracoda, ICLE01). Protura TSA are also depauperate in CYP4 sequences, with just one in Acerentomon sp., and at least three in Sinentomon sp.

We took a closer look at the four CYP4 clan genes of the honey bee to discern any pattern that might explain why just these genes are representing this clan. CYP4G11 is a conserved CYP4G essential in Neoptera for cuticular hydrocarbon biosynthesis and its evolution in Apidae has been described (Feyereisen, 2020). CYP4A1 is generally a single copy gene in a perfectly supported monophyletic clade from Polyneoptera to Holometabola, with the notable exception of Lepidoptera and Condylognatha. Its function is unknown. CYP4AB3 is a single gene in bees, but this subfamily can be greatly expanded (blooms) in ants and wasps, with 20 genes in Nasonia vitripennis. CYP4AV1 is part of a well supported clade (100/100/95) of unique genes from various CYP4 subfamilies. These include the wasp CYP4AV5, the aphid CYP4CK1, the louse CYP4BT1, the damselfly CYP4PA1 and is anchored by Diploptera punctata CYP4C7. The latter P450 plays an important role in the cockroach corpora allata, clearing all JH precursors at the end of a gonotrophic cycle of activity (Sutherland et al., 1998, 2000).

3.4. Diversity - The many and the few

Taking the number of CYP subfamilies as an arbitrary measure of diversity (members of different subfamilies having generally less than 55% identity), showed that there is a significant correlation between CYPome size and diversity ($r = 0.83$, $n = 31$). However, smaller CYPomes are more diverse than larger ones, as shown by plotting the CYPome diversity relative to CYPome size against CYPome size ($r = -0.5944$, $n = 31$) (Fig. 6). The larger CYPomes have remarkable blooms of P450s, which are mostly consisting of closely related sequences, thereby decreasing their contribution to overall diversity. Some of the CYPomes that are both smaller and with smaller relative diversity, i.e. outliers in Fig. 6, appear to be from aquatic organisms, D. pulex, H. asteca, L. polyphemus and to some degree Penaeus vannamei and S. maritima. Data from more CYPomes would be needed to follow this apparent trend.

In the species studied here, the distribution of sequences into families is unequal, with a pattern of many CYP families with few genes and few families with many genes (P450 blooms). This pattern is most obvious for large CYPomes which represent the tail end of the distribution in Fig. 6. However, even the very small genomes follow this trend. A. lycopersici with only 23 genes, has one well supported clade (96/97.2/76) of ten genes, and therefore a bloom of its own, with five genes in the CYP3120 family and the five other more divergent named in different families. In V. destructor with 26 genes there is a small bloom of four CYP3011 genes that is also shared with M. occidentalis. The exception to this power-law pattern of many CYP families with few genes and few families with many genes is the L. polyphemus CYPome of 42 genes (see below, onhologs), in which the only large clade is formed by four CYP3001T genes.

This general power-law pattern has been noted long ago for the sizes of gene families in genomes (Huynen and Van Nimwegen, 1999), and has been described throughout eukaryotic CYPomes as well (Reed and Hughes, 2004; Feyereisen, 2011; Sezutsu et al., 2013). Mathematical birth/death models can closely approximate such patterns (Qian et al., 2001; Karev et al., 2004; Reed and Hughes, 2004). The pattern can be observed in large and small CYPomes in hexapods, crustacean, myriapods and chelicerates. Calla et al. (2017) saw a “problem with this framework” invoking inconsistencies within the CYP nomenclature. Yet in the S. curvata CYPome, the power-law pattern (few with many and many with few) is observed by counting families following the CYP nomenclature, but also by strictly following the supported clades in our
phylogenetic analysis (Fig. 7). Since the latter is dependent on the sequence phylogeny and not on the names, the specious argument of Calla et al. is refuted. Similarly, Good et al. (2014) objected that birth-death models dealt with gene numbers and ignored gene sequence subjected to molecular evolution. This was already refuted by Axelsen et al. (2007) who reported on the power law behavior of paralog sequence identities in proteomes rather than just number of genes per family. The pattern of many CYP families with few genes and few families with many genes is seen in free living arthropods and in obligate parasites, in monophagous and polyphagous species, in carnivores, herbivores and omnivores, in terrestrial, aquatic and subterranean species. It is seen in Holometabola where complete metamorphosis can lead to a complete change in dietary pattern and diversity (e.g. from detritivore to blood feeder), but also in arthropods which occupy the same ecological niche throughout their life cycle. With the rare exception of L. polyphemus then, the many and the few constitute the defining pattern of sequence distribution in CYPomes. At the ends of this distribution are CYP blooms on one side and the small families of conserved orthologs on the other. As noted by Sezutsu et al. (2013) “patterns of gene duplications and death in the CYPome are not dependent on the ecology and life-history traits of the organism. Plant–animal chemical warfare, widely thought to be a root cause of the proliferation of CYP genes, does not shape the distribution of P450s into families”. Strangely, some authors added a corollary implying that the distribution pattern would therefore be “completely dependent on stochastic changes” (Calla et al., 2017) and that it would therefore be “necessary to invoke adaptation to explain P450 gene number change” (Good et al., 2014).

These are non sequitur because while birth/death models do not explicitly require selection, they do not explicitly exclude it, which is quite different. Sezutsu et al. (2013) clearly stated that P450 evolution is the result of evolutionary forces of which three (mutation/duplication, recombination and drift) are stochastic and non-adaptive, while the fourth, natural selection, plays a role that is highly variable in time and space. Good et al. (2014) showed exactly this, with some but not all Drosophila genes having seen selection favoring sequence changes in some lineages during at least part of their evolutionary history. And yet nothing in our current knowledge of P450 function would explain why Drosophila CYP306A1 (a conserved edcdysteroidogenic gene) or CYP318A1 (a “stable” gene in the “unstable” CYP313 clade) are under positive selection.

Similarly, Seppey et al. (2019) addressed the question of dietary shift in Coleoptera by comparing Adephaga (mostly carnivorous) and Polyphaga (mostly phytophagous). These authors took a non-biased approach, comparing sizes of gene families from these two lineages, and testing for adaptive expansions in either one. Of 22 orthologous groups (OG) of P450 genes in their study, selection for larger size in Polyphag was shown for only one. This selected P450 OG (EOG805V7G), includes 22 genes in Dendroctonus ponderosae and 45 genes in Tribolium castaneum, belonging to CYP6 subfamilies. Thus two thirds of the P450s were not selected by dietary shift (we thank Robert Waterhouse for access to sequences belonging to OG EOG805V7G). In the comparison of the rice and corn strains of S. frugiperda, (Gouin et al., 2017), only two P450 genes, CYP6AE74 and CYP340L16 showed marked positive selection, and detoxification genes were not over-represented in genes showing positive selection. Johnson et al. (2018) showed positive selection in some CYP6AS genes of social bees, which they associated with the consumption of flavonoids. Perhaps more important in the dynamics of CYPome composition is the gene turnover rate, rather than changes in gene count, as shown in Scaptomyza flava (Gloss et al., 2019). This herbivorous drosophilid has the smallest CYPome when compared to seven other Drosophila, but among the highest P450 gene birth and gene death rates, and a small bloom of CYP6G genes.

3.5. Blooms and clusters

The expansion of specific branches of the gene tree by multiple duplications or blooms of P450 genes is not restricted to a particular CYP clan. CYP3 clan blooms in insects include the CYP6AE in Lepidoptera or the CYP6CY in aphids and in the CYP4 clan, CYP4AB in N. vitripennis or CYP325 in Anopheles gambiae. While modest in size in insects, the CYP2 clan shows massive blooms in other groups, as the 75 CYP3705 in S. curviseta or the 37 CYP392 in Tetranychus urticae. Blooms in the mitochondrial clan are observed, as the 7 CYP333B in M. sexta. and the 26 CYP12 in the cat flea C. felis (our unpublished results). Blooming also occurs in CYPomes of small size as the CYP3120 in A. lycopersici or the CYP6AS in bees. Genes that are mostly single orthologs throughout the phylogeny (see below) occasionally bloom as well, for instance the six CYP4G1 in Musca domestica, the eight CYP4G15 genes in Halyomorpha halys, the ten CYP4G15 genes in Pyrococelia pectoralis (Feyereisen, 2020) or the eleven CYP303 in Photinus pyralis (Fallon et al., 2019). Even the elusive CYP20 clan shows a bloom of five genes in S. maritima. CYP blooms have been triggered at any time during evolution, and their age can be roughly determined by their distribution in extant species. In many if not most cases, the duplications leading to blooms result in the formation of clusters of CYP genes in the genomes. Such clusters of CYP genes have been recognized before the age of genomics but they evolved

![Fig. 7. Distribution of the Sinella curviseta P450s into groups showing the frequency of groups of various sizes against the number of genes per group. Groups are defined as CYP families according to the CYP nomenclature, or as CYP clades as defined by supported branches of our arthropod phylogeny. The regression lines in both cases show a power-law distribution, few groups with many genes and many groups with few genes.](image-url)

![Fig. 8. CYP6 clusters in Diptera. Top: the Musca domestica 180 kb scaffold of chromosome 5. Ten genes (in blue) were previously cloned (Feyereisen et al., 1989; Cohen and Feyereisen 1995; Zhu et al., 2008; Gao et al., 2012). Gene fragments (open arrow) or pseudogene (grey) are marked. Multiple genomic gaps suggest that some gene fragments may represent further genes in this cluster. Bottom: the orthologous cluster on chromosome 2 R in Drosophila. CYP317A and kank (S’ from CYP317, not shown) define the two clusters as orthologous, but there are no 1:1 orthologs between them. Note the different scale. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)](image-url)
from curiosity to norm with genome projects. Fig. 8 shows the CYP6A cluster of the house fly, *M. domestica*, showing the 22 genes, pseudogenes and gene fragments on chromosome 5, far more than the 6 genes laboriously cloned from lambda genomic libraries (Cohen and Feyerisen, 1995). This cluster predates the divergence between the house fly and *Drosophila* (at least 65 MY, Wiegmann et al., 2011), as the latter has a homologous cluster on chromosome 2R (Fig. 8). The comparison of closely related species can often date duplications to either before or after speciation events (see the CYP9A cluster example, d’Alençon et al., 2010). Gene conversion between closely clustered genes may cause some blooms to look “younger” than they are, but cannot deny their importance in shaping the diversity of CYPomes. Blooms are probably top-heavy (Feyerisen, 2011), i.e. with a more rapid expansion when “young”, followed by a slower rate of decay. This is in part because recently duplicated genes have twice the probability of duplicating again. This pattern is predicted by gene birth/death models, and is apparent on the tree, with short branch lengths for massive blooms and longer branches for blooms of lesser amplitude.

Although Lepidoptera have a very high rate of local gene rearrangement (d’Alençon et al., 2010), “old” CYP clusters are recognized in distant lineages, such as the CYP6AE cluster in Bombycoidea and Noctuoidea, and their genes are not fully dispersed in the genome. Much is known about the functions of the CYP6AE genes in the cotton bollworm (Shi et al., 2018; Wang et al., 2018). Fig. 9 shows the activity of ten CYP6AE enzymes of *H. armigera* on ten different xenobiotic substrates. Nine of the genes are clustered on chromosome 16 and one is on chromosome 21. There is no pattern or relationship between the position on the cluster, the phylogenetic relatedness or the catalytic competence of these P450s. This would suggest that there is a selective advantage for maintaining the cluster as a heritable unit, providing a reservoir of catalytic diversity, thus facilitating transitions from one ecological niche to another. Our phylogenetic analysis (Fig. 1) shows that the CYP6AE subfamily has 11 members in *H. armigera* (polyphagous) and 7 in each *B. mori* (monophagous) and *M. sexta* (oligophagous), so that there is little support for the proposition that a specialized diet is associated with a larger repertoire of CYP6AE genes (Calla et al., 2017; Amyelois transitella, polyphagous, 7 genes; *Depressaria pastinacella*, specialist, 9 genes). The data of Fig. 9 also shed some light on the fate of duplicated genes, and suggest that the classical distinction of sub- and neo-functionalization is at best an oversimplification representing extreme outcomes that are probably rarely encountered in the P450 family. Rather, the catalytic competence of a P450 may be viewed as a cloud of substrate structure space, here oversimplified by testing “only” ten substrates. Upon duplication, this cloud can change shape and volume but remains overlapping. It is rarely split in two (sub-functionalization) or giving rise to a new cloud (neofunctionalization).

Most clusters are related genes, syntetic as a result of their relatively recent duplication(s), but there are relatively rare cases of clusters of highly divergent P450s, with for example the CYP3 clan CYP347AA1 of *N. virescens* less than a kb distant from a cluster of three CYP4 clan genes. Most of the CYP3 clan P450s in this species (62/68) are distributed in 8 clusters. Whether CYP347AA1 is related to its CYP4 clan neighbours by function (e.g. forming a metabolic cluster) or whether its position is random is currently unclear (see below the md - CYP307A metabolic cluster).

### 3.6. Conserved orthologous groups

In a gene family with so many births and deaths as the P450s, it can be difficult to assign true orthology (homologous genes separated by speciation events). Paralogs (homologous genes separated by duplication) can blur the picture by differential loss. The somewhat looser term “orthologous group” can therefore be quite useful (Gabaldon and Koonin 2013). An example in arthropod CYPomes is the case of the neopteran insect CYP4G genes which despite multiple births and deaths have maintained two clearly distinguishable orthologous groups (Feyerisen, 2020). Mounting evidence shows that they have also maintained essentially the same function in cuticular hydrocarbon biosynthesis over 400 MY.

The study of genomes from closely related species provides an excellent opportunity to identify orthologs. This was done by Good et al. (2014) who have followed the fate of the 77 P450 genes inferred from the ancestral *Drosophila* genome on the species tree of 12 extant species. Of these 77 genes, 30 were considered evolutionarily “stable”, (“meaning that they have only one gene from each of the 12 species”). Given the fine grain of the Good et al. study, we can extend their definition to infer that “stable” genes are those incurring no documented gene gain or losses in the branches of the tree under study. These 30 genes can therefore be considered 1:1 orthologs in the strictest sense, although Good et al. simply called them stable genes, following Thomas (2007). Although the loss or duplication of a conserved gene in one lineage removes this gene from the stable category, loss of a gene does not negate the probable orthology of the remaining genes of that group. Keeping this caveat in mind, we extended the dataset of Good et al. to the phylogeny of arthropods to determine the fate of stable genes over time. The data (Fig. 10) show that “stability” follows a near exponential decay, with Collembola retaining just three stable genes (relative to the *Drosophila* set), and Myriapods one or none, depending on the branch considered. The curve may in reality be even steeper as it is based on the very limited sampling of arthropod CYPomes. The estimated “half-life” of a stable gene from this curve, about 50 MY, is therefore a generous upper limit. A similar curve can be obtained by starting at an opposite branch of the tree, or indeed by studying plant or vertebrate CYPomes (results not shown). Having established the relative nature of the stable
gene concept (when is it measured ?), it follows that the dichotomy between stable and unstable genes is itself relative. The notion of conserved orthologous group would therefore seem more appropriate than stability because whether a gene is duplicated or lost in a particular lineage does not falsify the importance of orthology, which brings with it a degree of sequence identity and possibly of function conservation as well, both meriting closer attention.

Our CYPome phylogeny reveals a number of such conserved orthologous groups in all clans. **Fig. 11** shows their distribution for groups that had at least three 1:1 orthologs and were found in three or more species of the 40 that we surveyed. There were about equal numbers of orthologous groups in each of the four major clans (9 in clan 2, 12 in clan 3, 12 in clan 4 and 12 in the mito clan). No P450 has a strict 1:1 ortholog in all species surveyed. Of the 47 orthologous groups, only 24 are strict 1:1 orthologs, and 23 are 1 to more. Some are found in virtually all arthropods (CYP18, CYP302, CYP307, CYP314, CYP315), or in all Mandibulata (CYP306). Some are found in all Neoptera as mentioned above (CYP4G1, CYP4G15) or are restricted to an order, where they are highly conserved. **Fig. 11B** shows fourteen such groups in Lepidoptera. Most order-specific orthologous groups in Lepidoptera are from CYP clan 3 in our analysis. This is in contrast to a previous study which claimed little conservation in CYP clan 3 with orthologies rarely recognizable (Calla et al., 2017). It is clear that surveying more species within an order would reveal more orthologous groups of one or few genes. For most of the conserved orthologous groups, the function of their members is unknown, however, their function is known or suspected for a few and these will be described first.

**Fig. 10. Decay of CYP gene “stability” over time.** Starting with the 12 Drosophila species study of Good et al. (2014), the number of stable genes (defined as genes that do not undergo gene gain or loss in the branches studied) was followed along the arthropod phylogeny to myriapods. The timescale is the same as in **Fig. 2**.
3.6.1. Ecdysteroid metabolism

The role of P450 enzymes in ecdysteroid metabolism has been known for many years (review in Lafont et al., 2012) and several CYP genes have clearly been associated with ecdysteroid biosynthesis. These are the CYP2 clan CYP306A1 and CYP307 as well as the mitochondrial clan CYP302A1, CYP314A1 and CYP315A1 (Fig. 12). These genes are often called by their confusing Drosophila mutant names, Cyp306a1 (phantom), Cyp307a1 (spook), Cyp302a1 (disembodied), Cyp314a1 (shade), and Cyp315a1 (shadow). They were collectively dubbed Halloween genes, a term that does not appear in the original (1984) publications describing the mutants, and that obscures the relationship with CYP18 discussed below. We will refer to them as ecdysteroidogenic P450s for convenience (see also review in Niwa and Niwa, 2014). This denomination is clearer and allows expansion to the future P450s that remain to be characterized. CYP18A1 is another P450 involved in ecdysteroid metabolism, by hydroxylation and further oxidation at C26, which constitutes hormonal inactivation in Drosophila (Guitard et al., 2011).

The origin of the P450 genes involved in ecdysteroid biosynthesis is of great evolutionary importance, as one might expect “Ecdysozoa” to be capable of making a molting hormone, yet the presence of ecdysteroids, or indeed of the ecdysteroidogenic P450s, is clearly not a synapomorphy of “Ecdysozoa”.

3.6.1.1. CYP307. CYP307 are enigmatic P450s of the CYP2 clan. Namiki et al. (2005) first showed that the B. mori gene and its D. melanogaster homolog (spook) were involved in ecdysone biosynthesis. D. melanogaster has two CYP307 genes, CYP307A1 (spook) and CYP307A2. The latter was found as putative pseudogene in the initial genome release (Tijet et al., 2001), but later obtained from a difficult to sequence heterochromatic region (3R-47.1). The function of the CYP307 enzymes is still unknown 15 years after being characterized as a key area in studies of ecdysteroid biosynthesis (Namiki et al., 2005), but it is thought that CYP307 are involved in the “black box” steps downstream of 7-dehydrocholesterol. CYP307 genes are known to be “unstable” with multiple instances of birth and death (Ształ et al., 2007; Rewitz and Gilbert, 2008; Sezutsu et al., 2013). We expand upon these earlier results, and confirm the presence of two orthologous groups in Pancrustacea, the CYP307A and CYP307B genes. These cover Hexapoda, Branchiopoda (D. pulex) and Copepoda within “Multicrustacea” (sensu Schwentner et al., 2017). Depending on the lineage, these two groups are either present together or singly, with no phylogenetic pattern (Supplementary Fig. S4).

Rewitz and Gilbert (2008) noticed the tail to tail location of CYP307A2 and the neverland (nvd) gene in D. pulex, A. gambiae and Drosophila willistoni. The nvd gene encodes a cholesterol 7-dehydrogenase, i.e. an enzyme upstream from CYP307 in ecdysteroid biosynthesis (Yoshiyama et al., 2011). We found that in D. melanogaster the two genes are only 0.1 cM apart. The synteny is also maintained in Aedes aegypti, at least four lepidopteran species, as well as N. vitripennis, Cephus cinctus, Neodiprion lecontii, Athalia rosae (tail to tail), Myus persicus, Frankiniella occidentalis, P. humanus and Zootermopsis nevadensis. In Aphis gossypii there are four genes, one CYP307A (XP_027846257) located tail to tail with a nvd pseudogene (XP_027846250) and nvd (XP_027846249) and two more in tandem array (XP_027848023 and XP_027848010), as well as one CYP307B gene (XP_027854219). CYP307A2 and neverland therefore form a functional cluster.

More distant from the CYP307A and 307B sequences are clades of CYP307 in Malacostracca, Cirripedia as well as Diplura, whose relationships to the CYP307A/B clades are unclear (Supplementary Fig. S4). There are two more clades, one (100/100/99) includes Strigamia CYP307 and most Chelicerata, and another (100/100/100) is specific for Acariformes. In none of these genomes did we detect synteny of CYP307 with nvd. While the TSA of the millipede C. hualienensis revealed no CYP307, its presence in millipedes can be shown in the genomes of H. holstii and T. corrallinus. In the bark scorpion, CYP307M was recently duplicated, with the two genes differing by just four nucleotides, but with different neighboring genes, while the horseshoe crab has three CYP307 genes (Supplementary Fig. S4).

All CYP307 sequences share unusual structural features (Supplementary File S2). The C-helix motif conserved Gxx(WxFAQ) of the CYP2 clan has instead a AsxWxQQxRR motif. The 1 helix of CYP307 lacks the conserved Thr, and has a LEDxxGHHxxvN consensus where the CYP2 clan has LxDLFxAGx(E/D)TS. The conserved ExxR and PERF are present and the Cys pocket motif has the consensus FxFxaGxRxCxG. We believe that these structural features will prove determinant in explaining the complex reaction(s) catalyzed by CYP307 enzymes. Until now, no clear evidence has been presented for any intermediate between 7-dehydrocholesterol and a putative “Δ4-diketol” precursor which would then be reduced to the most likely product of the “black box”, the “5β-diketol” (Fig. 12) and, depending on the species, further to the “ketodiol” (2,22,25-trideoxyecdysone) (Lafont et al., 2012). Furthermore, there is no evidence that the introduction of the 14 alpha-hydroxyl group is independent of the formation of the conjugated 7-ene-6-one moiety. That the complex and

Fig. 12. Ecdysteroid biosynthetic pathway in arthropods. A: The presumed substrate and product of the “black box” (BB) in which CYP307 enzymes are thought to be active are 7-dehydrocholesterol and the 5β-diketol, which can then be reduced to the “ketodiol” (2,22,25-trIDEOxyecdysone). Several hydroxylations lead to the common molting hormone 20-DEoxyecdysone. Carbon numbering is shown. Ponasterone A is an alternative molting hormone in various species including T. urticae and crustacea. B: The known pathway from the “black box” of four hydroxylations leading to 20-DEoxyecdysone are shown in the middle (classical scheme), with the CYP enzymes (magenta) known to catalyze those steps in Drosophila or Bombyx. Ecdysteroids in blue are known to occur in arthropods but do not fit in the classical scheme and the enzymes have not been identified. This figure covers C27 steroids with a 3β-hydroxyl. Some intermediates may be 3-oxo or 3-epi-ecdysteroids, and some species may use C24-methyl or ethyl precursors and homologs. See also Supplementary Table S8. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)
poorly understood reactions of this famed Black Box should be catalyzed by the product of an “unstable” gene as CYP307 is somewhat paradoxical. The high degree of CYP307 “instability” contrasts with the stability and high conservation of the other P450 genes of ecdysteroid biosynthesis. It is possible that different CYP307 have different substrates but a similar product, so that comparative biochemistry may resolve the paradox, and/or that CYP307 duplications allow different timing and sites of expression as in D. melanogaster and Nilaparvata lugens (Ono et al., 2006; Zhou et al., 2020).

3.6.1.2. CYP18 and CYP306. These two CYP2 clan P450s are closely related in sequence and result from a duplication event, forming a strongly supported monophyletic clade (99/99/88). They are also mostly present in close synteny, with the two genes head to head in the honey bee and Drosophila (Claudianos et al., 2006) as well as in D. pulex (Rewitz and Gilbert 2008). This arrangement is also found in Coleoptera, Hemiptera, Isoptera, C. splendens, and the collembolan Holacanthella duospinosa but the CYP306 and CYP18 genes are head to tail in another collembolan Orchesella cincta and in the amphipods H. azteca and Parhyale hawaiensis. They are also head to tail in the millipede T. corallinus, but tail to tail in another millipede, H. holsti. In Lepidoptera, CYP18 is duplicated, and the two genes have a different tissue expression pattern (Li et al., 2014). The synteny of CYP306 (25-hydroxylase, Niwa et al., 2004; pmh in Drosophila, Warren et al., 2004) and CYP18 (26-hydroxylase/oxygenase, Guitard et al., 2011) maintained over 600 MY is remarkable, given that the function of the two genes is thought to be opposite (biosynthesis vs. inactivation - at least in Drosophila). CYP18A1 has been lost in A. gambiae (Feyerreisen, 2006), but this loss is restricted to the A. gambiae complex, as it is found in the closest species, An. chrystyi and beyond (Neafsey et al., 2015). Similarly, we could not find CYP18 in the genome of Blattella germanica, nor in the TSA of any other related species in the “Blattellinae” (sensu Evangelista et al., 2019), although it is readily found in Blaberidae and other Blattodea (Supplementary File S2).

The M. sexta midgut P450 2C6-hydroxylase activity is found in both microsomal and mitochondrial compartments (Williams et al., 2000). The lepidopteran midgut CYP26-hydroxylase would presumably be encod-
ed by CYP18B1 as in B. mori (Li et al., 2014). Indeed M. sexta CYP18B1 has a predicted mitochondrial location by DeepLoc1.0. While this is unusual for a CYP2 clan P450, it would explain the biochemical data (Williams et al., 2000) yet would suggest, in addition, the existence of a microsomal P450 with 2C6 hydroxylase activity, or a cryptic micro-

3.6.1.3. CYP302. CYP302A1 is a C22 hydroxylase (db) in Drosophila (Warren et al., 2002). We did not find a CYP302 gene in the centipede S. maritima. However, CYP302 transcripts were found in other Chilo-
poda, Scutigera coleoptrata (Notostigmorhapha, order Scutergiormapha) and Eupolybothrus caverniculus (full length, Pleurostigmorpha, order Lithobiomorpha) suggesting that if CYP302 was indeed lost in S. maritima (Epimorpha, order Geophiliomorpha) this loss is not common to all Chilopoda (Supplementary File S2). The latter two lineages diverged about 375 MYA (Fernandez et al., 2018). We found a CYP302 transcript of S. maritima (in SRR5602569) but this could not be assigned to the genome and is in fact a C. sculpturae contaminant. In millipedes, the other major branch of Myriapoda, while H. holsti and T. californicus each have one CYP302 gene, we found two CYP302 genes in C. hualiensis where a further duplication has led to a third gene called CYP3197A1 which is divergent but closely related to the other two CYP302 (100/96/96 support). Qu et al. (2015) reported the absence of CYP302 in ticks, but we found CYP302 in I. scapularis and a number of other tick species. Although RefSeq calls two CYP302A1 genes in Platulla xylolatella, these are probably allelic variants (by one amino acid) of the same gene, as one is on a large scaffold, the other on a very small one.

3.6.1.4. CYP315. CYP315A1 is a C2 hydroxylase (sad) in Drosophila (Warren et al., 2002). The CYP315 gene was generally present in all species we studied, but was not easily found in the cecropes T. japonicus and T. californicus. However, it was found in other cecropes, P. nana and L. salmonis, and our phylogeny revealed that the CYP315 clade does contains two T. japonicus genes, CYP3022A1 and CYP3023A1, with strong support (98/97.4/73). This suggests that these two P450s are in fact the products of a CYP315 duplication and rapid divergence. Whether or not 2-deoxyecdysteroids are found in the Tigrinopus genus is not known, and the function of these two P450s merit examination. Three CYP315 are apparent in the collembolan F. candida, whereas there is just one gene in S. curviseta. In L. polyphemus, we found two pairs of CYP315 genes. The TSA of the millipede C. hualiensis revealed no CYP315, but it was found in the genomes of H. holsti and T. corallinus. In centipedes, S. maritima has two CYP315, which is intriguing given the apparent lack of CYP302.

Gilbert and Rewitz (2009) suggested that the intronless T. castaneum CYP315A1 was a retrogene. However, it is located in an intron of a “tyrosine-protein kinase transmembrane receptor Ror-like” gene just as its ortholog in D. ponderosa which has seven exons and its orthologs in Leptinotarsa decemlineata and Aethina tumida which have two. We conclude that this is a case of differential intron loss in Coleoptera, rather than retrotransposition. Schumann et al. (2018) reported the presence of three CYP315 genes, but no other ecdysteroidogenic gene in velvet worms. Our analysis suggests that only two Eup teriopoides CYP315 belong to the weakly supported monophyletic CYP315 clade (89/92.3/59), and whether they have ecdysteroid 2-hydroxylase activity is unknown.

3.6.1.5. CYP314. Although CYP314 is found in all arthropods that we studied and forms a monophyletic clade (100/100/98), there were some exceptions. CYP314A1 is a C20 hydroxylase (shd) in Drosophila (Petryk et al., 2003). CYP314 is found in crabs (Carcinus maenas, Eurypanopeus depressus) and 20-hydroxycydon and ponasterone A are commonly reported from decapods. However, Sin et al. (2015) reported a lack of CYP314 in the shrimp Neocaridina denticulata (Decapoda, Caridea). They rightly warned of possible poor recovery of this genomic locus, as few CYP genes they reported were of full length. We did not find a CYP314 in the genomes of P. vannamei (Pacific white shrimp), Penaeus monodon or Palaemon carinicauda. In the snow crab Sagmarthus (Janus) (Pau-

utra) verreauxi TSA, no CYP314 was recovered among 42 P450 se-

sequences (Ventura et al., 2017). Although these authors claimed to have identified a P450 of the CYP4 clan with 20-hydroxylase function (which can in fact be shown to be a CYP3213 of the CYP2 clan), their functional
expression lacks proper controls and does not support their identification. The apparent lack of CYP314 from shrimp is both remarkable and ironic, as 20-hydroxylated molting hormones are found in Panuliridae, dating back to the original isolation of crustaceodysone (20-hydroxyecdysone) from *Jasus lalandei* (Hampshire and Horn 1966). Also, 20-hydroxylation of ecdysone was shown in *Panulirus argus* (James and Shiverick, 1984). The identity of the 20-hydroxylase of shrimp thus remains obscure.

The CYP314 gene is also missing from the genome of eight ants of the Myrmicinae subfamily, notably Atta (3 species), Trachymyrmex (3 species), *Cyphomyrmex costatus* and *Acromyrmex echinatior* although it is found in all ants and 23 other species of Myrmicinae. The ants lacking CYP314 are all fungus farming ants of the Attini tribe which evolved around 50 MYA (Li et al., 2018). The lack of CYP314 is not related to the presence or absence of *Pseudomocaridia* actinobacterial symbiosis. Other ecdysteroidogenic P450s are present in Attini. Fungus farming has also evolved in some termites, but the genome of *Macrotermes natalensis* has a CYP314 gene. In aphids, we found two CYP314 genes in *M. persicae* and three in *Acrystosiphon pisum* (Supplementary File S2).

3.6.1.6. "Missing" genes. Barnacles (Thecostraca, Cirripedia) are particularly startling cases of “missing” genes, which highlight the limitations of the “Halloween” gene denomination and the need for further searches of ecdysteroidogenic P450s. These sessile organisms clearly molt, and clearly have 20-hydroxyecdysone and ecdysone. The CYP314 gene is also missing from the genome of eight ants of the Myrmicinae subfamily, notably *Atta* (3 species), Trachymyrmex (3 species), *Cyphomyrmex costatus* and *Acromyrmex echinatior* although it is found in all ants and 23 other species of Myrmicinae. The ants lacking CYP314 are all fungus farming ants of the Attini tribe which evolved around 50 MYA (Li et al., 2018). The lack of CYP314 is not related to the presence or absence of *Pseudomocaridia* actinobacterial symbiosis. Other ecdysteroidogenic P450s are present in Attini. Fungus farming ants of the Attini tribe which evolved around 50 MYA (Li et al., 2018).

3.6.1.7. P450 integration in arthropod ecdysteroid metabolism and origin of the pathway. Because ecdysteroidogenic P450s do not form a monophyletic clade and are instead dispersed in at least two clans of the CYPomes, their identification is most rigorous by phylogenetic analysis of the conserved synteny of CYP306 and CYP18 is maintained since the original duplication despite genomic rearrangements (inversions) seen in different lineages. At least in *Bombbyx* and *Drosophila*, the two genes have opposite effects on ecdysteroid titers, so perhaps the risk of hormonal imbalance has favored their inheritance as a linked locus. Species lacking CYP18 (*A. gambiae* and some cockroaches) rely on fewer, less redundant inactivation mechanisms (C3 oxidation, C22 conjugation) and this may point to potential target sites for biorational insecticides.

The conserved synteny of CYP306 and CYP18 is maintained since the original duplication despite genomic rearrangements (inversions) seen in different lineages. At least in *Bombbyx* and *Drosophila*, the two genes have opposite effects on ecdysteroid titers, so perhaps the risk of hormonal imbalance has favored their inheritance as a linked locus. Species lacking CYP18 (*A. gambiae* and some cockroaches) rely on fewer, less redundant inactivation mechanisms (C3 oxidation, C22 conjugation) and this may point to potential target sites for biorational insecticides.
Several insect lineages within Hymenoptera and Hemiptera cannot dealkylate C28 or C29 sterols, and therefore make predominantly C24-methyl- or -ethyl-ecdysteroids. Nothing is known of the substrate specificity of the ecdysteroidogenic P450s in those species. Even *Drosophila* makes C28 ecdysteroids depending on its fungal sterol supply. Nothing is known of the enzymes that hydroxylate at unusual positions, or that hydroxylate usual positions in the absence of some “Halloween” genes. For instance, the centipede *Lithobius forficatus* lacks 2-hydroxylase activity (Descamps and Lafont, 1993), but appears to have 25-, 22- and 20-hydroxylase activities. Both ecdyson and 20-hydroxyecdysone were observed by HPLC coupled with immunnoassays, thus identified with reasonable confidence in *L. forficatus* (order Lithobiomorpha). This biochemical evidence is paradoxical, as we found two CYP315 (probable 2 hydroxylase if the specificity is conserved) in the *S. maritima* genome (order Geophilomorpha) and we also found CYP315 transcripts in the TSA of *Scelopocryptops sexspinatus*, *Cryptops hortensis*, *Scoleoptrorhina* and *Craterostigmus tasmanianus* (orders Scolopendromorpha, Scutigeromorpha and Craterostigmomorpha). Yet we found no CYP302 (probable 22-hydroxylase) in *S. maritima*, although 3–400 MY separate these various orders of Chilopoda. A recent study in *N. lugens* (Zhou et al., 2020) shows that RNAi of CYP306 and CYP302 has no effect on molting and development, and the functional expression of CYP302 failed to show 22 hydroxylase activity. Yet this rice planthopper has 20E as principal molting hormone, raising questions as to the identity of the C25 and C22 hydroxylases. Such accumulations of discrepancies (Supplementary Table S8) do not strongly support the classical scheme of ecdysteroid biosynthesis. While this classical scheme is certainly very important in view of the overall conservation of the genes (Fig. 12), it is probably not the exclusive pathway of ecdysteroid biosynthesis and future research may reveal the function of other P450s and the existence of alternate or grid pathways.

Onychophora are the accepted sister group of arthropods. Nelson (2018) analyzed the CYPome of *Euperipatoides rowelli*, and reported the presence of a CYP51 (sterol 14-demethylase) that is absent from nematodes and arthropods, a CYP20 and a CYP26, along with representatives of the four usual CYP clans. We used the TSA of *E. kanangrensis* to verify and complete this set, and in a few cases used the *E. kanangrensis* full length ortholog in place of the partial *E. rowelli* sequence. We found no conserved CYP302, CYP314, CYP306, but two distant CYP315-like sequences and a CYP18-like sequence. Hoffmann (1997) reported the presence of ecdysone in *E. leuckarti*, identified by HPLC-RIA, and their origin (endogenous or food-derived) is unknown. Indeed, beyond arthropods, the presence of ecdysteroids has been repeatedly reported (see Lafont and Koolman 2009 for review) without evidence for an endogenous (biosynthetic) origin. For instance, the leech *Hirudo medicinalis* is a molting lophotrochozoan reported to contain ecdysteroids (Sauber et al., 1983). But if the leech does indeed make ecdysteroids, then this would be at best a case of convergent evolution, as we found no trace of arthropod-type ecdysteroidogenic genes in its genome. Although many plants and Cnidaria clearly make a variety of ecdysteroids, it is therefore in non-arthropod “protostomians” that their presence, let alone biosynthesis, remains doubtful. This is a common limitation of studies beyond arthropods. The lack of a complete set of known ecdysteroidogenic enzymes beyond arthropods (see also Schumann et al., 2018) leads us to conclude that while there are many aspects of ecdysteroidogenesis still unclear in arthropods, ecdysteroidogenesis itself beyond arthropods, i.e. in other protostomians, is still unproven. This conclusion echoes that of Lafont and Koolman (2009) and answers their plea for phylogenomic approaches to this question.

3.6.2. Other conserved genes

3.6.2.1. CYP15. The CYP15 sequences do not form a monophyletic clade in our analysis. The founder CYP15A1 from *D. punctata* is the highly specific methyl farnesoate (MF) epoxidase that makes juvenile hormone III (methyl farnesoate 10R,11-epoxide) in the corpora allata of this cockroach (Helvig et al., 2004). It is a part of a strongly supported clade (93/93.8/72) of CYP15A clade of lepidoptera, related to CYP15B1 of mosquitoes and to the CYP15C clade of Lepidoptera. CYP15C1 of *B. mori* is the allatinal farnesic acid epoxidase (Daimon et al., 2012), while CYP15A1 of *T. castaneum* is a mixed farnesic acid/MF epoxidase (Minakuchi et al., 2015). The CYP15 A/C clade therefore represents the bona fide enzymes producing the epoxide characteristic of juvenile hormones (JH). The CYP nomenclature places several CYP2 clan sequences in the CYP15 family, including two CYP15H from locusts, six from *M. harbei*, three from *C. splendens*, eight from *C. aquilonaris* and three from *S. curviseta*, but these do not form a single well supported clade and their function is unknown. In the termite *Reticulotermes flavipes*, CYP15F1 is induced by JH and facilitates JH-dependent caste differentiation but its precise biochemical function is unknown (Tarver et al., 2012). The three *C. splendens* CYP15 sequences are also outside the 15A/C clade. This raises the question of the precise origin of the CYP15 A/C clade and when during evolution of insects the JH function switched from the “ancestral” MF to its epoxidized congener, JH III. RNAi of CYP15A1 in *T. castaneum* does not cause precocious metamorphosis (Minakuchi et al., 2015), suggesting that the physiological function of MF was not abolished by the emergence of JH III, although other interpretations are possible, such as an extra-allatal epoxidation by other P450 enzymes.

Two lineages in insects have a different type of JH, the higher Diptera which make a JH “biseopoxide” (methyl farnesoate 6S,10R-diepoxide) and bugs, such as *Rhodius prolixus* (in which JH was first discovered) and several other Hemiptera which make a JH “skipped biseopoxide” (methyl farnesoate 2R,10R-diepoxide) (Villalobos-Sambucaro et al., 2020; Kotaki et al., 2009). *R. prolixus* has a single CYP15A1 gene, suggesting that this P450 can epoxidize at one if not at both sites. Aphids have duplicated the CYP15 gene with three copies in the peach aphid.

However, CYP15 is missing in higher Diptera, namely Cyclorrhapha. But is found in other Diptera such as *Empidioidea* (partial TSA sequence in *Heteropilus ingens*), Asiloidea (*Proctacanthus coquilletti*), Bibiornomorpha (*Mayetiola destructor*) (Supplementary File S2), Tipulomorpha, Psychodomorpha and Culicomorpha. The identity of the epoxidase in Cyclorrhapha is therefore unknown. CYP66G2 is highly expressed in the allatinal portion of the *Drosophila* ring gland, and its RNAi knockdown is lethal (Chung et al., 2009; Ou et al., 2016; Christesen et al., 2017). The biochemical function of CYP66G2 is unknown, but many CYP6 enzymes have a broad substrate specificity, and house fly CYP6A1 is highly active as 6- and 10-epoxidase of MF (Andersen et al., 1997).

Next to CYP15 A/C in our phylogeny is the CYP305 clade. Synteny relationships indicate that CYP15 and CYP305 are neighboring genes and that this synteny is recognizable from termites to mosquitoes (Supplementary Fig. S5). The function of the CYP305 genes is still unknown, but they are found in most Neoptera as a single gene. In gregarious locusts, one of the paralogs, CYP305M2, controls biosynthesis of the defense compound phenylacetonitrile (Wei et al., 2019), suggesting that it plays a regulatory role in phase determination.

3.6.2.2. CYP303. CYP303A1 is a strongly supported clade (100/100/100) with generally a single gene for each species, but it is duplicated in the Argentine ant, *Linepithema humile*, and in the carpenter ant *Camponotus floridanus* where the two genes are in a tandem array (Supplementary File S2). It is also duplicated in the damsel fly *C. splendens*. CYP303 was not found in genomes or TSA beyond winged insects.

The CYP303A1 are highly conserved with a 498 ± 4 amino acids length, yet the hymenopteran orthologs are much longer (*A. mellifera* 562 aa, *N. vitripennis* 587 aa). The difference is a single long insertion, confirmed by TSA of a variety of species. This insertion is predicted to be located between helices D and E, thus on the outside of the globular P450 structure and away from the ER membrane surface.

With possibly more exceptions as noted above, CYP303A1 is mostly a...
single copy gene, “stable” in insects, yet there is a bloom of CYP303 genes in fireflies (Fallon et al., 2018). The genome of P. pyralis carries 11 genes and two pseudogenes that are all paralogs of CYP303A1. This appears to be related to the biosynthesis of defensive compounds (lucibufagins), because related fireflies that do not make lucibufagins have only a single CYP303 gene (Fallon et al., 2018). Making these polyhydroxylated steroids from cholesterol may require 6–7 P450 actions. While the enzymatic details of lucibufagin biosynthesis are unknown, the firefly case suggests that the original CYP303A1 is already able to metabolize a sterol or terpenoid structure. Its endogenous substrate is still unknown, but the conserved regulatory function of the gene in Drosophila and locusts also point to a signal molecule, possibly a hormone (Wu et al., 2019, 2020b).

3.7. Unusual P450 sequences

3.7.1. Sequences lacking the conserved Cys

Our analysis provides a glimpse of the sequence space occupied by arthropod P450 sequences. Different genes can be identical at the amino acid level, reflecting a very recent duplication event. Conversely, as noted previously (Sezutsu et al., 2013), some may also lack the invariant residues that are the signature of this gene family. P450 sequences lacking the conserved Cys codon providing the thiolate ligand to the heme have evolved at least six independent times in arthropods. The first type evolved about 400 MYA and can be found as a full-length single sequence in 40 surveyed species of Polyneoptera and Condylognatha (Supplementary File S2). Fig. 13 shows its distribution. It includes the CYP408 sequences discussed previously (Sezutsu et al., 2013). This first type is from the CYP3 clan, closely related to CYP360 sequences. It is most noticeable by the lack or recognizable Cys pocket motif with in some cases Ser at that position (Fig. 13).

The second type evolved about 360 MYA in Heteroptera (“true bugs”), forming a monophyletic group of 16 sequences including CYP394A1 of Cimex lectularius, CYP394B1 of R. prolixus and CYP3221A1 of Murgantia histrionica belonging to the mito clan and closely related to CYP334 enzymes (Supplementary File S2). These sequences lack the WxxxR motif but have ExxR and PERW motifs, yet lack a Cys pocket recognizable motif.

The third type is restricted to the aphid lineage (Aphidoidea - Adelgidae and Aphididae) and therefore evolved about 250MYA (Li et al., 2017). The aphid CYP6YC1 was initially called CYP6UN1 as a fragment in A. gossypii. While it has homologies with the CYP6 family, it is an odd member of the CYP3 clan outside the main (polyplyhetic) CYP6 family. These sequences, found in single copies in 16 species (Supplementary File S2) are 30–40 codons shorter than most P450s. They have an N-terminal membrane-spanning region but miss parts of alpha helices B and C (no WxxxR motif), alpha F’ alpha G and H. They have ETLM in helix K, as well as the PxFR motif. The Cys pocket motif is truncated, and instead of the CYP3 clan typical PxFxxGXRxCxG they have a P(H/Y)xxGxxXYIG motif with Tyr replacing Cys. This odd type of P450, while possibly tethered to the membrane, has a wide open active site with perhaps a large hydrophilic substrate. Mutants of CYP119 with Cys replaced by other amino acids can accommodate heme iron coordination (Machotshi et al., 2015), and the aphid CYP6YC1 may therefore use either Tyr or the adjacent Ser (Goel et al., 2013) instead of the usual Cys as heme ligand. It is difficult to predict what type of reaction would be catalyzed by this P450.

The fourth type, derived from the CYP6 family, evolved about 220 MYA in a coleopteran clade that includes Tenebrionoida, Cucujoida and the “Phytophaga” (Chrysomeloidea and Curculionoidea). (Supplementary File S2). It includes CYP393A1 of D. ponderosae and CYP348A1 of T. castaneum. These short sequences have the ETLM and PDRF motifs and a conserved I helix lacking the usual Thr, and also lack the Cys in the heme binding pocket.

The fifth type is restricted to Orthoptera, more specifically to the family Acrididae, including L. migratoria CYP3136A1 (Wu et al., 2020a) and TSA sequences from four other species (Supplementary File S2). It is also derived from the CYP6 family and has a PFGxxxxNDTG consensus with Asp replacing Cys.

A sixth type is found in Psocodea, and is related to the mito clan CYP334 enzymes, but appears to be different from the heteropteran type. This type has a conserved Cys, but displaced 6 codons 5’ from the consensus (Supplementary File S2). The human body louse CYP375A1 belongs to this type.

3.7.2. Other sequences with unusual Cys pocket ExxGxxC motif

Such very unusual sequences as described above are also found punctually throughout the phylogeny. S. curvifera CYP3720A1 (CYP2 clan) lacks the conserved Cys and has a Ser instead. The other CYP3720 genes have the usual sequence with Cys. It also lacks the conserved WxxxR motif or I helix conserved Thr. This is fully supported by TSA data (GGYGO1006102), and is not an error from incorrect intron junctions, yet no such gene was found in other Collembola. The CYP301 orthologs are conserved in Neoptera but they differ in their Cys pocket motif. Whereas CYP391B genes all have the usual PFXxGxxxCxxG motif, all CYP901A genes have a PXxxGxxxCxxG motif instead. Two more examples are described below.

3.7.2.1. CYP316A1. CYP316A1 of Diptera has an unusual Cys pocket motif, YSLGPRLCPPAR in D. melanogaster and WGLPGRPNCV in C. capitata (XP_004519952) (Supplementary Fig. S6). This gene is poorly understood and expressed at very low levels in D. melanogaster. Good et al. (2014) consider this gene to be “stable” in the Drosophila lineage but while it is found in Drosophila relatives and in Tephritidae, it is not found in Calyptrata. The gene is a CYP4 clan member located head to head with CYP4D8 in D. melanogaster on 3L at 66A2, and tail to head with CYP4D8 in Tephritidae. In Calyptrata, CYP4D4 is present as tandem duplicates with the more usual conserved FSAGPRNGCQG motif. The Cys-Pro dipeptide of the drosophilids is relatively rare in P450s. In Mycobacterium tuberculosis CYP121, a Pro at that position causes a kink in one of the pyrrole rings of the heme because of their close interactions (Leys et al., 2013). In the tephridid CYP316A1, the conserved Phe is replaced by an unusual Trp, which, in view of the close proximity of the usual Phe to the heme probably also affect heme conformation and redox properties as shown for P450 BM3 Phe mutants (Chen et al., 2004). CYP316A1 probably catalyzes an unusual reaction or has an unusual type of substrate. While this gene is a fast evolving distant member of the CYP4 clan, it is not a CYP19 clan P450 as suggested by Kawashima and Hatta (2014).

Fig. 13. Distribution of the P450s related to CYP360 lacking the conserved Cys. Orders in which such sequences are found (including locust CYP408B1 and body louse CYP358A1) are marked in bold. Their consensus sequence around the missing Cys is shown as weblogo and is aligned with the consensus CYP360 sequence below.
3.7.2.2. CYP428A1. CYP428A1 in Lepidoptera is an unusual member of the mito clan, but predicted to have a mitochondrial presence and to be destined to the inner mitochondrial membrane. Its sequence is highly conserved, but very distinctive from other mito clan P450s. The consensus sequence around the Cys pocket motif has a three codon deletion, ASMPFG - - - xxCPxG when compared with the consensus ASLPGFGPRMCIGRR. Also, the CYP428A1 sequences have just one of the three positively charged residues known to favor binding of the adrenodoxin electron donor and have an unusual I helix, so they may use a different electron donor or may be independent of an external source of reducing equivalents (e.g. if acting on an endo-hydroperoxide). CYP428A1 is not found in Trichoptera or in non-Ditrysian Lepidoptera, although genomic data in those groups are scant.

3.8. Whole Genome Duplications (WGD) and ohnologues

Whole genome duplications, as in plant, vertebrate or fish lineages, are often quickly followed by the loss of most gene duplicates, or ohnologues. The remaining duplicates can provide evidence for a WGD event. Evidence for WGD in arthropods has only recently been presented. Horseshoe crabs (Xiphosura) have undergone one, or possibly two WGD events, estimated 230–310 MYA and 450–600 MYA (Nossa et al., 2014; Kenny et al., 2015; Roelofs et al., 2020). In chelicerates, one common WGD apparently occurred prior to the split between Scorpiones and Aranea, at least 430 MYA (Schwager et al., 2017).

We analyzed these CYPomes for traces of ohnologues. In L. polyphemus, the distribution of P450s is quite atypical, as we saw no blooms, but an excess of P450s in pairs. Pairs of P450s constitute fully half the CYPome and only nine genes are singletons. We next analyzed the genomic localization of the P450s to detect recent tandem duplications. Although the scaffold sizes are relatively small and this may bias the analysis, we found only two pairs of tandem duplicates, suggesting that the majority of pairs are ohnologues. The four copies of CYP315 appear as two pairs of ohnologues. Of the conserved orthologous groups, only CYP302, 314, 20 and 16 have cleared their ohnologues after WGD. Horseshoe crabs as “living fossils” are therefore unique in multiple respects, and WGD has not led to CYPome diversification. Whether the lack of P450 blooms is related to the WGD, or whether another mechanism limits the retention of CNV as gene duplicates is unclear. It has been hypothesized that retained and dosage-balanced ohnologs and their neighboring genes in vertebrates are refractory to CNV (Makino et al., 2013).

The situation is different in spiders and scorpions. In the spider P. tepidariorum, the distribution pattern of P450 is more typical, with a few blooms (e.g. 22 CYP41 genes and 20 CYP3304 genes) and many singleton genes, but no obvious ohnologues. In the scorpion C. sculpturatus, only six pairs of P450s appear to be ohnologues, including CYP315, CYP18 and CYP307, although in the latter case the possibility of an allelic variant cannot be excluded. WGD has therefore left few traces in the CYPomes of spiders (see however Clarke et al., 2015) and scorpions, as opposed to horseshoe crabs.

There is little support for WGD events in hexapods, but several lineages show tracts of large segmental duplications, including Collembola (Roelofs et al., 2020). We analyzed a series of recent duplications in S. curviseta, and found about equal numbers of tandem duplicates and non-tandem duplicates. The latter may represent markers of large segmental duplications, and this should be studied further when a chromosome-level assembly becomes available. Multiple, recent and independent duplication events involving hundreds of genes are reported in ticks (Van Zee et al., 2016), with only one paralog pair of P450 genes under positive selection. In aphidomorpha an ancient burst of duplications may have resulted from either WGD or large scale segmental duplications (Julca et al., 2020). These events led to several paralogs of conserved genes such as the CYP314, CYP307 or CYP15 in M. persicae and A. pisum.

3.9. A case of P450 lateral gene transfer?

The CYP3356A gene of Collembola (CYP3 clan) appears to have been transferred to the genome of the chive maggot Bradyysa odoriphaga (Diptera, Nematocera, Sciaroidea). The B. odoriphaga CYP3356A1 sequence (MF632275) is not found in any other species of Sciaroidea (six species with WGS or TSA data) or more generally Diptera. Its closest relatives are in Collembola, with 45% identity to the S. curviseta CYP3356A2 sequence, and to orthologous F. candida and O. cincta genomic sequences and other collembolean TSA sequences. The chive maggot P450 was reported to be expressed mainly in larval Malpighian tubules, and its RNAi suppression led to an increased susceptibility to several insecticides (Chen et al., 2019). In our phylogeny, the collembolan CYP3356 is found in a clade of chelicrate P450s, which is in itself remarkable.

3.10. Physiology and environmental response: a false dichotomy?

The function of some of the enzymes belonging to conserved orthologous groups described above are known or suspected. For the overwhelming majority of P450s however, no function is currently known. The pace of heterologous expression followed by biochemical characterization is slow and often driven by specific goals, such as the elucidation of a resistance mechanism. The commonly described dichotomy between P450s involved in detoxification and those involved in essential physiological reactions is often assimilated with the dichotomy between conserved (“stable” - see above) P450s and those in blooming clades. In Fig. 14 we tagged the P450s in both categories. As the function of some P450 is known for species that are not represented in our analysis, we tagged the closest BLASTp hit, and the figure should therefore be taken as broadly illustrative rather than as a precise catalog. The results show that both categories (hereafter called X for xenobiotic metabolizing and E for endogenous substrate, to simplify the discussion) are present in all four major clans. A key observation is the proximity of many E and X-P450s in this tree, i.e. there is no E clade and X clade. E-P450s are seemingly concentrated in the CYP2 and mito clans. This is

Fig. 14. Distribution of P450s with endogenous or xenobiotic substrates. Blue dots represent P450s with a known endogenous substrate in at least one species (E and S-P450s, see text). Red dots represent P450s known to metabolize a xenobiotic (X-P450s), or whose RNAi suppression leads to significant change in xenobiotic toxicity. Species not included in this tree are represented by their closest BLASTp match on the tree. The figure is therefore an illustration of the current state of knowledge. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)
biased by the ecdysteroidogenic P450s, whose importance is amplified by our choice to assign an E function to all those P450s, even though their function has been proven in just a handful of species. X-P450s are seemingly more abundant in the CYP3 clan, but this too is biased by the focus on mosquito and fly insecticide resistance work. The figure is therefore a snapshot of our current knowledge, or rather, of our ignorance about detoxification mechanisms in many major lineages of arthropods.

Gotoh (2012) introduced a third functional category beyond E and X, those involved in secondary metabolite synthesis and catabolism, called S. The need for an S category is most obvious for plant CYPomes, but those involved in secondary metabolite synthesis and catabolism, called by our choice to assign an E function to all those P450s, even though the two CYP337B genes of Melittta cinxia strongly associated with lifetime egg production (de Jong et al., 2014). While this would tend to classify this gene in the E category, unequal crossing-over between the two CYP337B genes of H. armigera can lead to a chimeric CYP337B3 which is an ambivalent xenobiotic metabolizer and cause resistance (Jousen et al., 2012), clearly an X category P450. Conversely, while the bloom of CYP6A genes in Diptera contains X-P450s (Sabourault et al., 2001; Tskirieli et al., 2019), it also contains CYP6A20 associated with male aggressive behavior in Drosophila (Dierick and Greenspan, 2006), possibly by modulating pheromone input (Wang et al., 2008), this would be an S-P450.

The E and X contrast can be useful to discern properties of each type of enzymes based on a number of criteria (Thomas 2007; Kawashima and Satta 2014; Scanlan et al., 2020), but can suffer from ad hoc exceptions to such criteria. Placing P450s in such boxes may be an oversimplification when the properties of the P450s are probably reflecting a continuum over evolutionary time. Kawashima and Satta (2014) considered the human CYP4 family (12 genes) to be in the X (“detoxification”) group, but eleven could well be considered in the E (“biosynthetic”) group instead (Guengerich, 2017). The situation in humans may be informative, as all 57 P450s have been functionally expressed and just four remain “orphans” with no known endogenous substrate. Yet of the 15 P450s generally thought to be xenobiotic metabolizers, 13 are also acting on endogenous substrates, and of the 38 P450s known to have endogenous substrates, 21 are also capable of metabolizing xenobiotics (Niwa et al., 2009; Durairaj et al., 2019). Another nonypical P450 classification would be needed for Arabidopsis CYP706A3, a highly promiscuous tertiary oxidase protecting flowers from insects and bacteria (Boachon et al., 2019), when plant P450s metabolizing terpenes are generally considered to be highly specific (“S”-type). Too little is known about arthropod P450s to document such ambivalent X-E-S enzymes, with perhaps time and site of expression as well as substrate availability playing an important role in our perception of the X, E and S categories. Drosophila CYP6G2 appears to be catalytically an X-P450 (Daborn et al., 2007; Denecke et al., 2017), but its expression profile suggests an E function as well (Christesen et al., 2017). The CYP4G enzymes may be another example (Feyereisen, 2020), usually not blooming and often involved in highly specific pheromone biosynthesis, yet having multiple substrates and being indirectly involved in toxicokinetics (Balabanidou et al., 2016).

3.11. Some remaining questions

This study provides a low resolution, first view of arthropod CYPome diversity and evolution. While it has focused on some aspects of P450 structure and function, questions remain and more carefully annotated CYPomes are needed to answer them. Assigning functions to P450 enzymes is a daunting task, that has not even been completed for the moderately-sized human CYPome (57 genes). In the insect model Drosophila, 60 P450 enzymes are still orphans but “xenobiotic”-metabolizing functions are postulated for about a third of the CYPome (Scanlan et al., 2020).

With 600 MY of evolution and a phylogeny of hundreds of paralogs, the arthropod P450 genes are a valuable dataset to study intron loss and gain. Conserved intron position and phase in P450 genes were already seen as a useful phylogenetic signal in Drosophila (Tijet et al., 2001). In some species, intron turnover is extremely high, as in M. occidentalis (Hoy et al., 2016). Indeed, CYP4EN1 and CYP4EN2 are recently duplicated genes (82% identity) with 0 or 7 introns, respectively. A focus on exon-intron structures may also facilitate the search and detection of gene conversion and alternative splicing events. Although our manual curation of P450 sequences has probably eliminated most obvious errors, only deep transcriptomes and high coverage genomes will help further improve the quality of our annotations, as well as shed some light on the question of alternative splicing. An interesting case was the very scorpion-like P450 structure (CYP367D3) found in the C. sculpturatus CYPome. BLASTn searches exon by exon revealed an overlapping sets of 24 exons, but a single transcript of 11 exons, with an intricate genomic structure (Supplementary Fig. S7). Partial genes or exons are a common feature of CYPomes, and future work will need to focus on the possibility of alternative splicing in species other than Drosophila where the alternative splice forms of CYP4D1 are known (Tijet et al., 2001; Chung et al., 2009; Good et al., 2014). Alternative splicing of human P450s has been well documented (Annalora et al., 2017).

4. Conclusions

Our study shows that arthropods have widely different CYPomes
sizes and widely different CYP450 diversity and composition. Many arthropod lineages are still underexplored in terms of P450 diversity, particularly Myriapoda and Crustacea. Even in well represented groups such as Lepidoptera or Diptera new CYP families remain to be discovered. Only a handful of P450s are common to most but not all arthropods, and while these are presumed to be involved in molting hormone biosynthesis, this pathway is more diverse than generally described, and involves P450s that have not yet been discovered or characterized. Within arthropod lineages, often at the ordinal or subordinal level, small-sized conserved orthologous groups of P450 emerge. Only very few are currently known to catalyze reactions of importance to physiology, such as the CYP4G or the CYP15 in Neoptera. At the other extreme, multiple lineage-specific gene duplications or blocks are seen in all CYP clans and throughout arthropods. We owe Thucydidies in his account of Pericles’ funeral oration for contrasting the rule of the few and of the many. In the CYPomes we studied, only in the living fossil L. polyphemus did the few rule the distribution of sequences. This common distribution of many genes in few families and few genes in many families results from birth/death processes that have been modelled for gene families in general as a stochastic process. This should be considered the null hypothesis (Koonin, 2016), to be falsified before adaptation can be claimed. While evidence for adaptation has been presented in some cases and for some P450s, phylogenetic and demographic constraints have not yet been fully explored, and adaptation alone cannot be responsible for the observed diversity of numbers and sequences of arthropod P450s.

Arthropods have repeatedly explored the limits of P450 sequence diversity, with many variations on the consensus or signature sequences of the P450 superfamily. Studying the catalytic activity of these unusual P450s is likely to lead to the discovery of new types of reactions beyond classical monooxygenation. Arthropod CYPomes will therefore serve as an inspiration for green chemistry catalysts. The continued study of their evolution should also shed light on many aspects of arthropod physiology and inspire genomics-based ecotoxicology and pest management research.

Note added in proof: A CYP19 clan sequence was found in Thermobia domestica (D.R. Nelson & R. Feyereisen, unpublished results).

Acknowledgments

We thank David R. Nelson for the innumerable CYP names and his patience with our many requests, Toshinori Kozaki for help with B. mori sequences, and Jianzhong Zhang with L. migratoria sequences. This work was supported by the European Union’s Horizon 2020 research and innovation program [ERC grant 772026-POLYADAPT to TVL].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ibmb.2020.103490.

References


insect biochemistry and molecular biology. 127 (2020) 103490

Insect Biochemistry and Molecular Biology 127 (2020) 103490

21

Rong, Y., Fujii, T., Ishikawa, Y., 2019a. CYPs in different families are involved in the

Daphnia Halloween genes that encode cytochrome P450s mediating the synthesis of the arthropod molting hormone:

Roelofs, D., Zwaenepoel, A., Sistermans, T., Nap, J., Kampfraath, A.A., Van De Peer, Y.,
Bertone, M.A., Cassel, B.K., Bayless, K.M., Heimberg, A.M., Wheeler, B.M.,

Rong, Y., Fujii, T., Itoh, Y., 2019a. CYPs in different families are involved in the

Daphnia Halloween genes that encode cytochrome P450s mediating the synthesis of the arthropod molting hormone:

Roelofs, D., Zwaenepoel, A., Sistermans, T., Nap, J., Kampfraath, A.A., Van De Peer, Y.,
Bertone, M.A., Cassel, B.K., Bayless, K.M., Heimberg, A.M., Wheeler, B.M.,

Daphnia Halloween genes that encode cytochrome P450s mediating the synthesis of the arthropod molting hormone:

Roelofs, D., Zwaenepoel, A., Sistermans, T., Nap, J., Kampfraath, A.A., Van De Peer, Y.,
Bertone, M.A., Cassel, B.K., Bayless, K.M., Heimberg, A.M., Wheeler, B.M.,

Daphnia Halloween genes that encode cytochrome P450s mediating the synthesis of the arthropod molting hormone:

Roelofs, D., Zwaenepoel, A., Sistermans, T., Nap, J., Kampfraath, A.A., Van De Peer, Y.,
Bertone, M.A., Cassel, B.K., Bayless, K.M., Heimberg, A.M., Wheeler, B.M.,

Daphnia Halloween genes that encode cytochrome P450s mediating the synthesis of the arthropod molting hormone:

Roelofs, D., Zwaenepoel, A., Sistermans, T., Nap, J., Kampfraath, A.A., Van De Peer, Y.,
Bertone, M.A., Cassel, B.K., Bayless, K.M., Heimberg, A.M., Wheeler, B.M.,

Daphnia Halloween genes that encode cytochrome P450s mediating the synthesis of the arthropod molting hormone:
W. Dermauw et al.

Insect Biochemistry and Molecular Biology 127 (2020) 103490


