The Molecular Evolution of Cytochrome P450 Genes within and between *Drosophila* Species

Robert T. Good[†], Lydia Gramzow[†,1], Paul Battlay, Tamar Sztal[2], Philip Batterham, and Charles Robin*

Department of Genetics, University of Melbourne, Australia

[1]Present address: Department of Genetics, Friedrich Schiller University Jena, Philosophenweg 12, Jena, Germany

[2]Present address: School of Biological Sciences, Monash University, Australia

*Corresponding author: E-mail: crobin@unimelb.edu.au.

[†]These authors contributed equally to this work.

Accepted: April 14, 2014

Abstract

We map 114 gene gains and 74 gene losses in the P450 gene family across the phylogeny of 12 *Drosophila* species by examining the congruence of gene trees and species trees. Although the number of P450 genes varies from 74 to 94 in the species examined, we infer that there were at least 77 P450 genes in the ancestral *Drosophila* genome. One of the most striking observations in the data set is the elevated loss of P450 genes in the *Drosophila sechellia* lineage. The gain and loss events are not evenly distributed among the P450 genes—with 30 genes showing no gene gains or losses whereas others show as many as 20 copy number changes among the species examined. The P450 gene clades showing the fewest number of gene gain and loss events tend to be those evolving with the most purifying selection acting on the protein sequences, although there are exceptions, such as the rapid rate of amino acid replacement observed in the single copy *phantom* (*Cyp306a1*) gene. Within *D. melanogaster*, we observe gene copy number polymorphism in ten P450 genes including multiple cases of interparalog chimeras. Nonallelic homologous recombination (NAHR) has been associated with deleterious mutations in humans, but here we provide a second possible example of an NAHR event in insect P450s being adaptive. Specifically, we find that a polymorphic *Cyp12a4/Cyp12a5* chimera correlates with resistance to an insecticide. Although we observe such interparalog exchange in our within-species data sets, we have little evidence of it between species, raising the possibility that such events may occur more frequently than appreciated but are masked by subsequent sequence change.

Key words: cytochrome P450, Cyp12a4, *phantom*, Cyp6a20, *Drosophila* Genetic Reference Panel, nonallelic homologous recombination.

Introduction

Comparative genomics between closely related species affords an evolutionary context by which we can begin to understand functions of genes in multigene families and their role in the adaptation of organisms to their ecological niche (Claudianos et al. 2006; McBride and Arguello 2007; Sackton et al. 2007; Low et al. 2007; Shah et al. 2012). A central concern in the analysis of multigene family diversification is the extent to which it is driven by adaptation to species-specific environmental niches. Few would dispute that the gain and loss of genes have played a major role in the adaptation of organisms to their environments. However, nonadaptive processes such as "concerted evolution" may affect the evolution of at least some multigene families (Coen et al. 1982). Thus, while gene duplication, particularly when accompanied by sufficient sequence divergence provides one of the most obvious of candidates for adaptive divergence between species (Prince and Pickett 2002), selectively neutral explanations may explain gene number differences between species, and these may be proffered as null hypotheses that need to be rejected. For example, in analyzing genes of the P450 superfamily, Feyereisen (2011) notes that stochastic gene birth death models are sufficient to explain their proliferation among arthropods, and he therefore posits that it is not necessary to invoke adaptation to explain P450 gene number change. The suggestion is that some arthropod P450 genes may be functionally redundant.

Although birth–death model and other models may satisfactorily describe changes in gene number over evolutionary time (Reed and Hughes 2004; Novozhilov et al. 2006; Hahn 2009; Ames et al. 2012), they do not explicitly address the role of adaptation in gene family proliferation. Such models focus

on gene numbers and ignore the fact that each gene has a sequence that is subjected to the forces of molecular evolution. Scrutiny of these sequences (in contrast to the flux in gene numbers) can provide a strong delineation between adaptive and selectively neutral expectations. Nonfunctional sequences with homology to protein-coding sequences will accumulate many mutations (such as those appearing as hypothetical frameshifting mutations) that will not occur in functional sequences. Furthermore, in the *Drosophila* genus nonfunctional sequences are lost quickly, with the half-life of pseudogenes being estimated to be 18 Myr (Petrov and Hartl 1998; Robin et al. 2000). Thus, *Drosophila* genes that have maintained their potential to code for proteins, despite significant divergence from homologs are unlikely to be considered as "redundant" with respect to fitness. Rather their divergence from their paralogs suggests they have evolved their own selectively favored function that may only be apparent in the context of the ecological niche of the organism.

One caveat to the logic is that many divergent pairs of duplicate genes appear to fulfill complementary subfunctions of an ancestral gene, and in this way genetic flux may be accompanied by phenotypic stasis (Hughes 1994; Force et al. 1999). The most powerful tests of subfunctionalization require a detailed investigation of biological and molecular function of the gene products and their effect on phenotypes that may only manifest in one of many environments (Hillenmeyer et al. 2008). Although there are some elegant genetic experiments illustrating the subfunctionalization process (van Hoof 2005), they do not discount the possibility that subfunctionalization itself could have been adaptive, perhaps in subtle ways.

Another caveat to the logic that argues genes with substantial divergence are likely to have their own function, is that substantial sequence divergence could arise in redundant gene sequences via occasional interparalog exchange that would not necessarily introduce frameshift, and other inactivating mutations. That nonallelic homologous recombination (NAHR) events are mutationally possible is demonstrated by the presence of chimeric genes segregating within populations, including those of humans (Dumont and Eichler 2013). A pertinent example comes from the moth *Helicoverpa armigera* where a chimera between two cytochrome P450 paralogs (*Cyp337b1* and *Cyp337b2*) called *Cyp337b3* has been found. The *Cyp337b3* haplotype segregates with the *Cyp337b1–Cyp337b2* haplotype in natural populations (Joussen et al. 2012) and appears to be adaptive as it is associated with greater levels of resistance to a widely used insecticide, esfenvalerate.

A common phenomenon observed in multigene families in comparative genomic data sets is the occurrence of lineage-specific gene amplification of paralogs- or "phylogenetic blooms" (Ranson et al. 2002; Feyereisen 2011). Feyereisen (2011) cites multiple examples of cytochrome P450 gene blooms including the 15 *Cyp2c* genes in mice, the 19 *Cyp4ab* genes in the wasp *Nasonia vitripennis*, and the 12 *Cyp6a* genes in *Drosophila melanogaster*. To understand the relative roles of selective and neutral processes in such phylogenetic blooms and in multigene families more generally, it is necessary to focus on recent evolutionary events, in multigene families where functional analyses are tractable. The genomic data sets currently available for species within the *Drosophila* genus have divergence times ranging from <0.5 to ~50 Ma (*Drosophila* 12 Genomes Consortium et al. 2007). It is therefore possible to observe molecular evolution at unprecedented resolution, such that 1) the age of gene gain events can be accurately mapped to a species phylogeny and 2) many gene loss events can be observed as pseudogenes. Furthermore, when the divergence of nonfunctional DNA has not reached saturation, those sequences can be used to normalize rates of sequence change thereby allowing tests for adaptive evolution to be performed (Yang 2007). In *Drosophila*, there is the added benefit of the availability of population genomic data sets for *D. melanogaster* (Langley et al. 2012; Mackay et al. 2012).

Here, we examine within and between species copy number variation (CNV) through the lens of the large and highly divergent cytochrome P450 gene family among species within the *Drosophila* genus. In insects, this multigene family encodes enzymes that catalyze a variety of molecular reactions, typically hydroxylations, on endogenous and exogenous substrates (Feyereisen 2005). They have diverse biological functions that are best characterized in the model insect *D. melanogaster*, which also has extensive transcriptomic data sets that informs functional analyses. Particular P450s have been associated with detoxification of insecticides, whereas others have key developmental roles and many of them have been partially characterized in reverse genetic RNAi screens (Chung et al. 2009). Previously, gene duplication and loss have been studied for particular *Drosophila* P450 genes (Sztal et al. 2007; Schmidt et al. 2010; McDonnell et al. 2012; Harrop et al. 2014) and the P450 multigene family has been included in larger studies (Wu et al. 2011). Here, we examine the patterns of P450 gene duplication within and between *Drosophila* species and ask: 1) Are there lineage effects, such as phylogenetic blooms, among *Drosophila* species? 2) Is there any evidence for nonadaptive molecular evolutionary processes shaping the divergence of paralogs? 3) Are there signs of adaptive evolution in the divergence patterns of P450 genes, and if so which ones, in which lineages? and 4) What insight can be gained into the function of those genes whose function is currently uncharacterized?

Materials and Methods

Annotation of P450 Genes

Iterative BLAST (Altschul et al. 1990) searches using *D. melanogaster* P450 gene sequences as queries were used to identify contigs containing P450 genes in the other species. Later on,

also newly identified P450 genes from the other species were used as queries to ensure discovery of the whole set of P450 genes. In case the contigs that were identified did not contain the whole P450 gene or deviated in structure from orthologous contigs, we tried to improve the identified contigs by searching the trace archives and reassembling the corresponding contig. To identify the putative gene structures in these contigs, we used the automated annotation program Phat (http://bioinf.wehi.edu.au/Phat/, last accessed April 30, 2014). The automated annotations were adjusted in Artemis (http://www.sanger.ac.uk/resources/software/artemis/, last accessed April 30, 2014) using orthologous P450 genes as a guide. The coding sequences of the final annotations are provided in the supplementary data files, Supplementary Material online.

Phylogenetic Trees

An alignment of all Cytochrome P450 enzymes identified in the 12 *Drosophila* species was created using ClustalW (Thompson et al. 1994) and a neighbor-joining phylogeny was reconstructed based on this alignment. From this tree, 77 clades were identified and named as follows: If a clade has one-to-one orthologs to *D. melanogaster* in all *Drosophila* species, it was named after the *D. melanogaster* enzyme. If a clade contained homologs to more than one *D. melanogaster* P450 enzyme, its name is a concatenation of the names of the *D. melanogaster* proteins. For instance, the homologs to the three *D. melanogaster* enzymes *Cyp4p1*, *Cyp4p2*, and *Cyp4p3* form one clade and thus the clade was named *Cyp4p1/2/3*. To count the number of P450 proteins for each *Drosophila* species functional genes as well as pseudogenes were taken into account. Functional enzymes of a clade were aligned using ClustalW. Protein alignments were used as template to create nucleotide alignments using the program MRTRANS or translatorX (http://www.translatorx.co.uk, last accessed April 30, 2014). Phylogenetic trees were generated using the Mobyle server (http://mobyle.pasteur.fr, last accessed April 30, 2014). Phylogenetic trees shown in the figures were rendered using Figtree vs1.4 (http://tree.bio.ed.ac.uk/software/figtree/, last accessed April 30, 2014).

Locating Duplication Events

Neighbor-joining trees for 30 clades containing more than one copy of a gene in one or more species were created using ClustalW version 1.83. The trees were rooted using the midpoint method as implemented in PHYLIP version 3.66. The protein tree was then compared with the species tree using the Forester algorithm to locate duplication events. Some duplications that were predicted by Forester (Zmasek and Eddy 2001) seemed unlikely and were ignored or placed at a different branch of the tree. These cases include instances where:

- There is exactly one gene of each species in a subclade of the tree but the topology is different to the species tree. This phenomenon was explained by incomplete lineage sorting in the case of differences in the topology in *D. erecta*, *D. yakuba*, and *D. melanogaster* or an accelerated rate of evolution in one lineage. It might also be caused by long-branch attraction where long branches are grouped together although they are separated by short branches in reality.
- One species had two copies of a gene and one of these copies was an outgroup to the other genes in the subclade. In this case, Forester predicted a duplication at the root of the subclade. The duplication was relocated to the species that has two copies of the gene.

A total of 49 exceptions from the duplications located by Forester were made.

PAML Analysis

Saturation of synonymous sites was studied in P450 genes as saturation leads to an overestimation of the ω ratio. The method of Nei and Gojobori as implemented in PAML version 3.14 was used to predict synonymous substitution rates between pairs of genes in a clade. Pairs of species were ordered according to their divergence times. A curve was fitted to rate data derived from all clades with one-to-one orthologs using locally weighted polynomial regression as implemented in the statistical package R. The curve was used to determine at which evolutionary distance saturation occurs in P450 genes.

To avoid false detection of positive selection, PAML analyses were restricted to genes from species in the *D. melanogaster* group. Looking at this subset of species allowed to break up certain clades into two clades. A total of 81 clades were tested for lineage- and site-specific effects. For the study on evolution after gene duplication, genes from the *D. obscura* group were included additionally to the *D. melanogaster* group. More information on evolution in background branches was obtained by loosening the conservative approach that was used before. MRTRANS alignments and species trees with duplications inferred as described above were used as input for the codeml program of PAML version 3.14. Where applicable, three different trees were used, one for each possible topology of *D. erecta* and *D. yakuba* in relation to *D. melanogaster*. Each PAML analysis was repeated with three different start values for ω (0.5, 1, and 2) to identify the global minima.

Branch-Specific Models

Three different tests were conducted to identify lineage-specific effects in the evolution of clades. The free-ratio model was compared with the one-ratio model. The free-ratio model allows different ω values for each branch while the one-ratio model assumes a single ω value for all branches in the tree. Twice the difference of the log-likelihood values for these models was compared with the χ^2 distribution with degrees of freedom equal to the number of branches in the

tree minus one. Bonferroni correction was used to determine whether these LRTs were significant. The following two studies were conducted on the topology of *D. erecta* and *D. yakuba* in relation to *D. melanogaster* that had the highest log likelihood in the test above. Two-ratio models were compared with the one-ratio model. A two-ratio model allows one ω value for 1) specified branch/es (called foreground branch/es) and 2) another ω value for the rest of the branches in the tree (background). Each branch in a tree of a clade was used as foreground branch once resulting in as many two-ratio models for a clade as there are branches in the tree. Each two-ratio model was compared with the corresponding one-ratio model using an LRT as described above with one degree of freedom. Bonferroni correction was applied twice, first to account for multiple testing within a clade and second to account for multiple testing having 81 clades. Two- and three-ratio models were used to study change of selective pressure after gene duplication. The three-ratio model has one ω ratio for branches ancestral to the duplication, one ω ratio for the two branches immediately following the duplication event and a third ω ratio for subsequent branches. If a duplication had occurred in a terminal branch, the third ω ratio was not applicable and a two-ratio model was used. Correction for multiple testing was applied using the Bonferroni method and taking into account that 44 duplications were studied.

Site-Specific Models

To identify positive selection among sites models M0, M1a, M2a, M3, M7, and M8 were used. Model M0 is equivalent to the one-ratio model described above. Models M1a to M8 classify sites into two or more classes with different ω values. Model M1a defines two site classes of which one evolves neutrally and the other one is under purifying selection. M2a has an additional site class that allows sites to evolve adaptively. Model M3 assumes a general discrete distribution of ω ratios whereas M7 assumes a beta distribution of ω values over sites. As the beta distribution is limited to the interval (0, 1), M7 does not allow sites to evolve adaptively. In contrast, M8 allows an additional site class that can have an ω value of >1. LRTs were performed to compare M3 with M0, M2 with M1, and M8 with M7 as defined above with degrees of freedom 4, 2, and 2, respectively. The LRT comparing M3 with M0 is a test of variable selective pressure among sites whereas the other two LRTs are tests of positive selection among sites. The Bonferroni method was used to correct for testing of 81 clades. Clades with a significant result in the LRT comparing M8 and M7 were analyzed to identify sites under positive selection. Posterior probabilities for each site to belong to the site class with an ω value >1 were extracted from the PAML results.

Structural Model for Cyp318a1

The structure of *Cyp318a1* was modeled using MMM model (Rai et al. 2006). The nearest structural neighbor to the *D. melanogaster* enzyme as stated in the NCBI protein database is the structure of the human microsomal *CYP3A4* (PDB 1TQN) and was used as a template for modeling.

Sequencing

Cyp6a16 alleles were PCR amplified using primers (TCACACT GCTGCTGCTGAC-3′ and AGGTTAGTTTCCCGTGCTTG-3′) with a touch-down PCR protocol with annealing temperature reduced from 70 to 55 °C over 15 cycles followed by 30 cycles of 55 °C. The alleles were isolated from isochromosomal lines generated from natural populations of *D. melanogaster* spanning the eastern Australia latitudes (Schmidt et al. 2010). The PCR products were purified using Qiaquick columns and sequenced using BigDye terminator technology.

Insecticide Bioassays

Ten DGRP lines identified with the *y*; *cn bw sp*; reference genome arrangement of *Cyp12a4* and *Cyp12a5* (*426*, *45*, *239*, *639*, *101*, *40*, *491*, *440*, *42*, and *228*), and eight DGRP lines with the *Cyp12a4/5* chimeras arrangement (*358*, *399*, *217*, *365*, *129*, *443*, *705*, and *357*), along with the lufenuron-resistant strain *NB16* (Bogwitz et al. 2005), were raised on rich media and placed in mass-bred cages. First instar larvae were collected from laying plates and placed in vials containing screening media at a density of 50 larvae per vial. Three replicates were performed for each fly line, at doses of 0.25, 1.5, and 3.5 μg/ml lufenuron. Vials containing larvae were incubated at 25 °C for 14 days, after which time-eclosed adults, both alive and dead were scored as having survived to adulthood. Proportions surviving were calculated by dividing the mean number of eclosed adults from each dose, with the mean number of eclosed adults from control treatments.

Results

P450 Gene Gain and Loss among Species

We have identified and annotated a total of 975 P450 sequences in 11 *Drosophila* species in addition to the 90 P450 sequences known from *D. melanogaster* (supplementary data set S1, Supplementary Material online; Tijet et al. 2001). The annotation process required extensive curation that included some reanalysis of genes previously thought to be pseudogenes and the identification of new start codons of some P450 genes of *D. melanogaster*. The identified sequences include 928 putatively functional genes (i.e., these sequences appear to encode complete P450 proteins without reading frame disruption) and 47 pseudogenes.

P450s are classified by family (e.g., CYP6, CYP4, CYP307; originally defined as having >40% amino acid sequence identity) and then by subfamily (e.g., CYP4d, CYP4ae;

Nelson 2006). A phylogeny of the P450s shows broad agreement with expectations set by the P450 nomenclature system and by previous studies (fig. 1; Feyereisen 2005; Nelson 2006; Strode et al. 2008). As is well established, multiple families contain mitochondrial target sequences (e.g., CYP12, CYP315, and CYP49), and they all fall within a deeper monophyletic group. The denser sampling provided here resolves some family level relationships showing that the CYP9, CYP317, and CYP310 families are all nested within the CYP6 family, and the CYP312 family is nested within the CYP4 family. Most of the subfamilies group within a single family-specific clade (e.g., all the CYP12s form a clade, and the same is true for the CYP9s, CYP28s, and the CYP313s).

The phylogeny reconstruction shown in figure 1 depicts 77 clades that we trace back to the Most Recent Common Ancestor of the 12 *Drosophila* species studied ($MRCA_D$). Hereafter, these clades will be referred to as *AncD* (for ancestral *Drosophila*) clades. The P450 genes were assigned to these clades based on the species phylogeny of the 12 *Drosophila* species (*Drosophila* 12 Genomes Consortium et al. 2007; Stark et al. 2007). In the majority of cases, the recapitulation of the species phylogeny in these genes means that we can be highly confident of these assignments. However, there are some clades that we are less certain of, particularly the "dynamic" clades that exhibit many gene duplications and losses. The *AncD* clades are listed in table 1, which also indicates those for which gene gain and loss is more difficult to ascertain (we have named the clades after the *D. melanogaster* gene contained within them, or by the most accurate Flybase annotation if they do not contain a *D. melanogaster* gene). Even if the phylogeny and its interpretation are absolutely accurate, 77 genes are the minimum number of genes that occurred in the MRCA of *Drosophila*, because of the possibility that ancestral genes have been lost in all 12 genomes studied here.

Thirty of the *AncD* clades are evolutionarily "stable" (Thomas 2007) meaning that they have only one gene from each of the 12 species (table 1). The phylogenetic relationship within these stable groups frequently showed slight deviations from that expected from the species tree, and we attributed these to shortcomings in phylogeny constructions (Pamilo and Nei 1988; Pollard et al. 2006) rather than invoke complex gain and loss events of P450 genes. The stable genes include those involved in ecdysteroid synthesis (*Cyp302*, *Cyp314*, *Cyp306*; Gilbert 2004; Rewitz et al. 2006b), ecdysone modification (*Cyp18a1*; Guittard et al. 2011), bristle development (*Cyp303a1*; Willingham and Keil 2004), and cuticular hydrocarbon metabolism (*Cyp4g1*; Qiu et al. 2012). However, there are many stable P450 genes with unknown function (e.g., *Cyp4s3*, *Cyp6v1*, and *Cyp4ad1*).

Seventeen *AncD* clades have lost but not gained P450 genes since the $MRCA_D$. Reconciliation of the gene trees and species trees within these clades shows that six of these have lost a single P450 gene in a terminal species-limited branch (i.e., *Cyp4c3* and *Cyp313a4* are missing from the *D. grimshawi* genome, *Cyp6a13* from the *D. mojavensis* genome, *Cyp6t3* from the *D. ananassae* genome, *Cyp12b2* from the *D. yakuba* genome, and *Cyp310a1* from the *D. sechellia* genome). The "loss" of these genes in particular may have technical explanations such as poor assemblies, sequencing errors, sequencing gaps, or loss of function alleles in the sequenced strains (see below), rather than genuine losses fixed within a species. Three clades exhibit a single gene loss inferred to have occurred in an internal branch of the species tree and therefore they are absent from multiple genomes and technical explanations for their absence are less likely (*Cyp307a1*, *Cy4d2*, *Cyp_Dvir\GJ21722*). The remaining eight clades exhibit multiple independent losses of the same gene across the *Drosophila* radiation (*Cyp6a16*, *Cyp308a1*, *Cyp12c1*, *Cyp6d2*, *Cyp_Dvir\GJ21709*, *Cyp_Dmoj\GI21254*, *Cyp4d21*, and *Cyp4e3*). The gene we refer to as *Cyp_Dvir\GJ21709* (temporarily named after an automatic genome annotation of the *D. virilis* genome) is a previously unidentified gene that does not have any orthologs in *D. melanogaster* but is upstream of *Cyp4e2* in the species where it occurs. It can be distinguished from *Cyp4e2* by a distinct exon–intron boundary and a distant 5′-exon. This gene was lost independently three times: In *D. grimshawi*, in the ancestor of *D. persimilis* and *D. pseudoobscura*, and in the ancestor of the *D. melanogaster* group (supplementary data S3, Supplementary Material online).

The remaining 30 *AncD* clades show gene duplication in one or more *Drosophila* species. Twenty of these orthologous groups have gene loss and gene gain. The most dynamic of the *AncD* clades is the *Cyp4p* clade (fig. 1). According to the reconstruction in figure 2, this clade has experienced 20 gene duplications and 3 gene losses since the $MRCA_D$, although the confidence of some of the nodes in the tree is low so perhaps a scenario involving 19 gains and 1 loss is more parsimonious (table 1). The next most "dynamic" P450 clades since the $MRCA_D$ are the *Cyp313a1/2/3/5* and *Cyp6a2*s, each of which exhibits eight duplications and two losses. Although *Cyp6a2* is in the large *Cyp6a* gene subfamily in *Drosophila* (fig. 1), it is not within the largest gene cluster, which encodes other *Cyp6a* subfamily genes (fig. 3).

Considering all the 77 *AncD* clades, we estimate a total of 114 duplications and 74 losses (table 1). Our estimation is conservative as we applied a parsimonious approach rather than a strict reconciliation between the gene tree and species tree. Although inevitably these interpretations introduce a level of subjectivity, we believe they represent a more accurate depiction of the true phylogeny rather than objective computational reconciliations (e.g., those done with Forester; Zmasek and Eddy 2001). In supplementary data set S2, Supplementary Material online, we supply a full phylogeny for comparison. Of the 74 lost genes, 47 are still recognizable as pseudogenes whereas the remaining 27 losses were inferred from

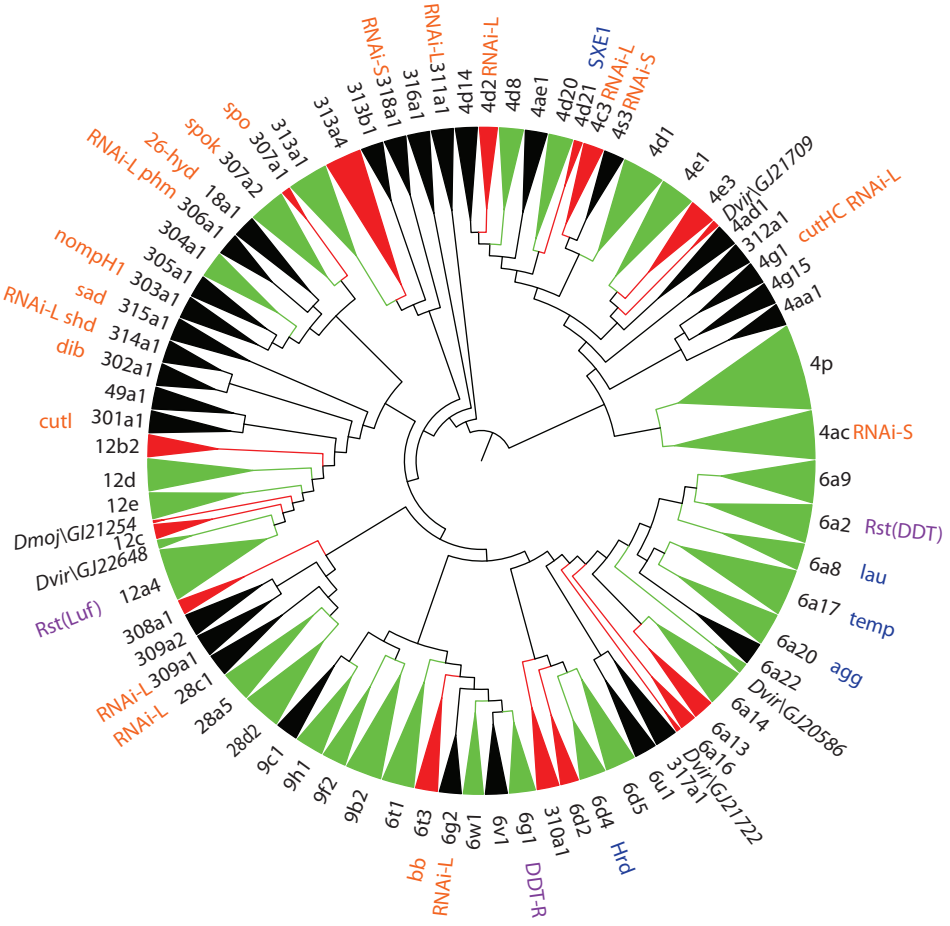

Fig. 1.—Phylogeny of cytochrome P450 genes in *Drosophila*. An unrooted circular cladogram of a neighbor-joining tree of ~1,000 P450 proteins from the *Drosophila* genus. The tree has been collapsed down to 70 clades representing those that are inferred to be present in the ancestor of all *Drosophila*. The "stable" clades are shown in black, the clades with only gene loss are shown in red, the clades with gain (and possibly loss) of genes are shown in green. Genes with specific functions in development are noted in orange (*Hwn* Halloween: Gilbert 2004, Namiki et al. 2005, Rewitz et al. 2006a, *nompH1*: Willingham and Keil 2004; *bb*: Rewitz and O'Connor 2011; *cutl*: Sztal et al. 2012 *cutHC*: Qiu et al. 2012) those associated with insecticide resistance are in purple (*DDT-R*; Daborn et al. 2002, *Rst*(*DDT*): Amichot et al. 2004, and *Rst(luf)*: Bogwitz et al. 2005) and others that have been the focus of publications are represented by blue lettering (*Hrd*: Hardstone et al. 2006, *lau*: Helvig et al. 2004, *temp*: Kang et al. 2011, *agg*: Dierick and Greenspan 2006, and *SXE1*: Fujii et al. 2008). RNAi-L refers to genes shown to be lethal in an RNAi screen of Chung et al. (2009) and RNAi-S are sublethal in that screen.

phylogenetic reconstructions alone. No traces of the nucleotide sequences of these 27 genes have been identified from the corresponding genomes suggesting they have been deleted or mutated beyond recognition.

P450 Gene Gain and Loss within *D. melanogaster*

An early annotation of P450 genes in the *D. melanogaster* genome identified 90 sequences, 7 of which were thought to be pseudogenes (*Cyp307a2*, *Cyp6t2*, *Cyp6a16*, *Cyp6a15*, *Cyp9f3*, *Cyp49a1*, and *Cyp313a1*; Tijet et al. 2001). We have previously found that *Cyp307a2* is not a pseudogene (Sztal et al. 2007) and neither is *Cyp49a1* or *Cyp313a1* (more recent Flybase annotations). As alluded to above we have also found that genes can be misannotated as pseudogenes because the reference genome has carried inactivating mutations that are not present in other alleles. We refer to these as null alleles to distinguish them from pseudogenes that are fixed in the population. To verify whether *Cyp6a16* was a pseudogene, we sequenced 1 kb around the 11 nt frameshifting deletion observed in the *y; cn bw sp* reference strain (position 2L: 5622861 and 5622862 of genome release = r5.56: CTCAG G.....CGGAAAAGGACT) from eight Australian isofemale lines and failed to find this inactivating mutation. However, the 11 nt deletion is unlikely to be a sequencing error as it is also observed in other sequenced lines (e.g., *Drosophila* Genetic Reference Panel [DGRP]-136 strain). Furthermore, none of the other 13 nt polymorphisms that

Table 1
The AncD clades

Stable	Unstable	Duplications	Deletions	Confidence in Reconstruction
Cyp18a1	Cyp4c3	0	1	[a]
Cyp314a1	Cyp4d2	0	1	[a]
Cyp4g1	CYP9F2	1	0	[a]
Cyp6v1	Cyp310a1	0	1	[b]
Cyp303a1	Cyp12b2	0	1	[b]
Cyp4ad1	Cyp6a13	0	1	[a]
Cyp301a1	Cyp6t3	0	1	[b]
Cyp305a1	Cyp307a2/3	1	0	[a]
Cyp4g15	Cyp9h1	1	0	[a]
Cyp313b1	Dvir\GJ21722	0	1	[a]
Cyp4aa1	Cyp307a1	0	1	[a]
Cyp302a1	Cyp313a4	0	1	[b]
Cyp49a1	Cyp6d2	0	2	[a]
Cyp317a1	Cyp304a1	2	0	[a]
Cyp309a2	Cyp4d8	1	1	[a]
Cyp6g2	Cyp12e1	2	0	[a]
Cyp6a22	Cyp12c1	0	2	[a]
Cyp9c1	Cyp4d20	1	0	[a]
Cyp6u1	Cyp4d21	0	4	[c]
Cyp4d14	Cyp308a1	0	2	[a]
Cyp309a1	Dmoj\GI21254	0	2	[b]
Cyp318a1	Cyp4d1	1	0	[a]
Cyp28c1	Dvir\GJ21709	0	3	[a]
Cyp312a1	Cyp4e3	0	2	[b]
Cyp315a1	Cyp6g1	3	0	[a]
Cyp311a1	Cyp6w1	1	2	[a]
Cyp4ae1	Cyp6a16	0	2	[b]
Cyp316a1	Dvir\GJ22648	1	2	[b]
Cyp306a1	Cyp4e1/2	5	1	[c]
Cyp4s3	Dvir\GJ20586	3	2	[b]
	Cyp6a8/18	2	3	[b]
	Cyp12d1/2	2	3	[b]
	Cyp6d5	4	1	[a]
	Cyp28a5	4	1	[b]
	Cyp6d4	3	3	[a]
	Cyp6a17/23	6	0	[b]
	Cyp4ac1/2/3	3	3	[c]
	Cyp9b1/2	3	3	[c]
	Cyp6a9/21	5	2	[b]
	Cyp28d1/2	6	1	[b]
	Cyp12a4/5	7	0	[b]
	Cyp6t1/2	2	4	[c]
	Cyp6a19/20	4	5	[b]
	Cyp6a14/15	5	4	[c]
	Cyp6a2	8	2	[b]
	Cyp313a1/2/3/5	8	2	[c]
	Cyp4p1/2/3	19	1	[c]

[a]Highly confident.
[b]Fairly confident.
[c]Low confidence.

we did observe were obviously disabling, 6 were replacements, and 7 were silent ($R/S = 0.86$) and a McDonald–Kreitman test (using the divergence data of 12 replacement and 21 synonymous fixations $R/S = 0.57$) suggested that the pattern of polymorphism between synonymous and nonsynonymous sites was not different to the pattern observed in the divergence between *D. melanogaster* and *Drosophila simulans* ($G = 0.36$, $P > 0.05$). Thus, the *Cyp6a16* allele of the genomic reference strain (*y; cn bw sp*) seems to be a null allele. In contrast, polymorphism data confirm *Cyp6t2* and *Cyp6a15* are genuine P450 pseudogenes in *D. melanogaster*.

To analyze P450 CNV within a species, we analyzed genomes of *D. melanogaster* lines from the DGRP (Mackay et al. 2012). Three bioinformatic analyses were performed to identify and assess CNV. Firstly, a coverage-based screen relying on read-depth variation was used to identify putative P450 gene CNV. Secondly, the distance between Illumina paired end reads for each strain was examined and compared with the reference genome. We sought paired-end violations replicated across multiple DGRP strains. Thirdly, some of the DGRP strains have also been sequenced with 454 sequencing and so single reads spanning CNV breakpoints were identified. Ten P450 genes exhibiting CNV among the DGRP were found in more than one of the 162 DGRP lines (fig. 4). All ten come from the "dynamic/unstable" clades, for which gene copy varies between *Drosophila* species. Among the ten is a duplication of the *Cyp9f2* gene, which was previously identified as *Cyp9f3* and assigned pseudogene status as it occurs in the *y; cn bw sp* genome reference strain. The previously characterized structural variation at the *Cyp12d1* and *Cyp6g1* locus were observed at high frequency (Schmidt et al. 2010).

Are There Lineage Effects in the Patterns of Gene Gain and Loss among *Drosophila* Lineages?

The number of P450 genes per species ranges from 74 putatively functional genes and 14 pseudogenes in *D. sechellia* to 94 putatively functional genes and 8 pseudogenes in *D. willistoni*. Strong lineage-specific effects were observed in the number of duplications and losses (fig. 5). The number of duplications in a lineage roughly correlates with divergence time. For instance, gene duplications have been particularly numerous along the branch leading to *D. willistoni* (25 gene gains) which is one of the longest branches on the species tree. In contrast, the striking observation about lineage-specific gene loss relates to one of the shortest branches on the species tree: *D. sechellia* has lost 14 P450 genes (listed in supplementary data set S4, Supplementary Material online). No losses were detected in the sister lineage leading to *D. simulans* and thus relative rate of gene loss down these sibling lineages is highly significant (Tajima's 1*D* relative rate test, Fisher's exact test, χ^2 value $= 9.3$, $P < 0.01$). All of the *D. sechellia* losses are detectable as pseudogenes that harbor frameshift and nonsense mutations.

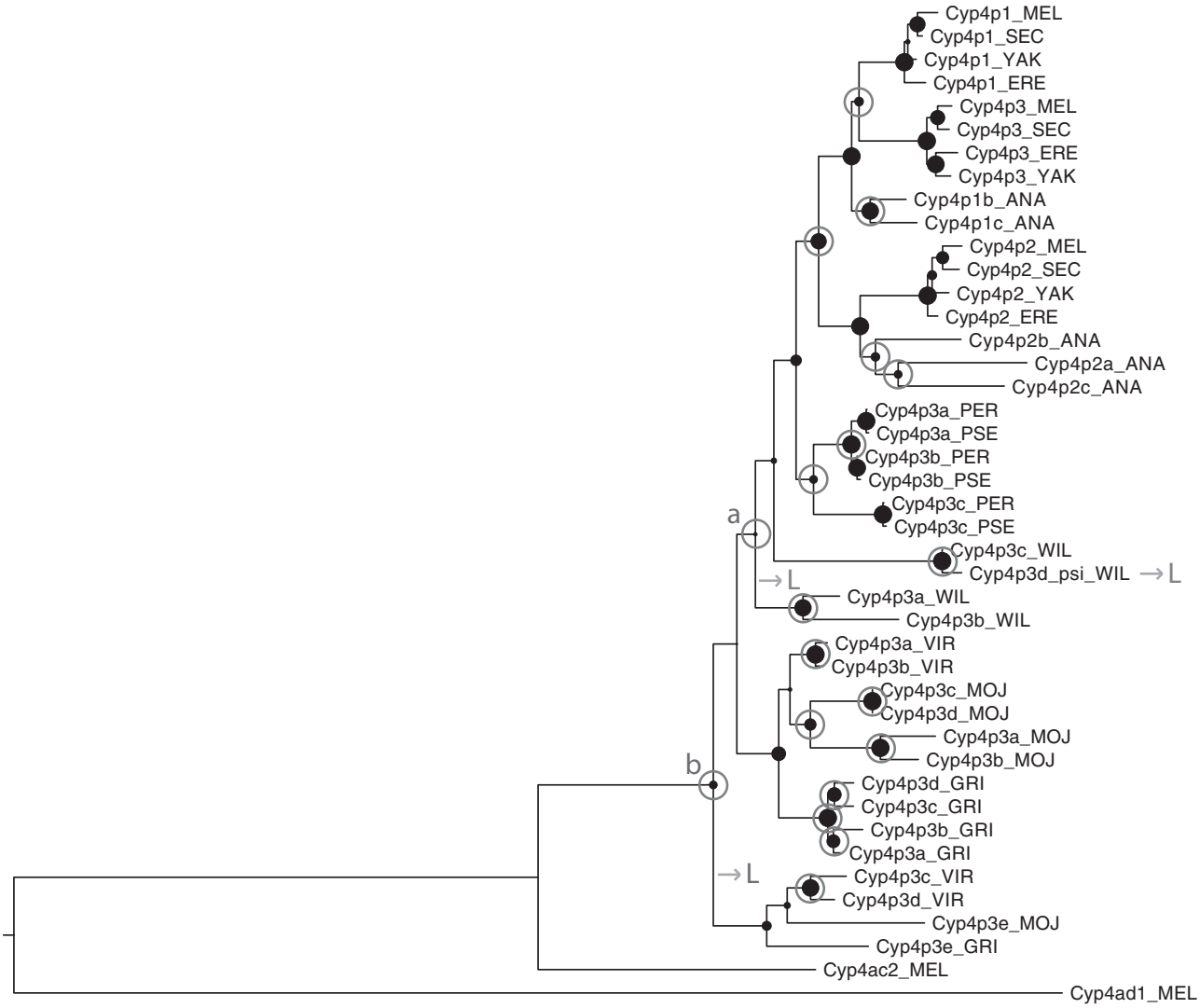

Fig. 2.—A phylogenetic tree of the *Cyp4p* genes of the *Drosophila* genus. The maximum-likelihood tree was generated using protein sequences using the phyML algorithm. Full length sequences of *D. simulans* were not available and so they have not been included in the analysis. The size of the black circles at the nodes represents the bootstrap confidence scores and the 19 nodes that have a gray circle around them represent inferred gene duplication events. The three gene loss events inferred by this tree are indicated by gray Ls. The node marked with an "a" suggests that there was a gene duplication before the divergence of the *D. willistoni* from the other Sophophorans, which consequently would require a gene loss in the rest of the Sophophorans. However, this node has a very low bootstrap support (46%) and perhaps a more parsimonious solution would be if the duplication happened in the willistoni lineage (as no loss is necessary). Similarly, if the gene duplication indicated at node b (with bootstrap support of 63%), actually occurred after the divergence of the *Drosophila* and Sophophoran subgenera then the gain before the divergence of the *Drosophila* species and the loss in the Sophophora subgenus (as indicated by this tree) could be replaced with a single gene gain in the *Drosophila* subgenus.

What Molecular Evolutionary Processes Affect the P450 Multigene Family?

The overwhelming majority of gene duplicates are at adjacent locations suggesting they originated by unequal recombination. For example, all of the 19 gene duplications occurring in the *Cyp4p* lineage resulted in adjacent genes, all of which contain introns, strongly suggesting unequal recombination as their mechanism of origination. Over evolutionary time adjacent genes have become separated by secondary events such as inversions. A clear example of these processes is observed in some of the *Cyp6a* genes. In *D. melanogaster*, there is a cluster of

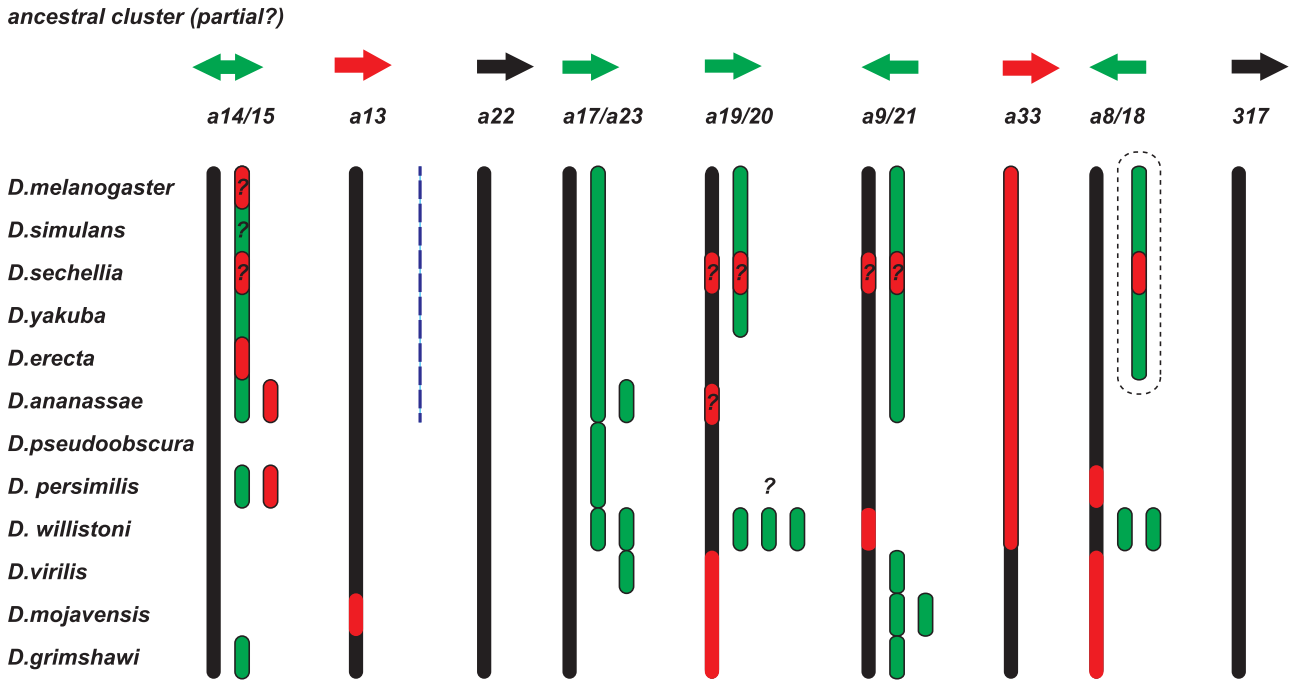

Fig. 3.—The *Cyp6a* cluster. The inferred composition of the ancestral *Cyp6a* cluster is shown with arrows representing genes and their direction of transcription. Obviously this does not include any genes for which there is no recognized descendants in the species examined and therefore the figure may represent only a partial version of the cluster. The gene order may also have been different in the ancestral species. *Cyp6a14* and *Cyp6a15* are divergently transcribed in *Drosophila melanogaster*, and since it is not clear which direction the ancestral gene was transcribed, it is represented as a double-headed arrow. The genes that have not changed in copy number during the divergence of the 12 species are indicated in black. The genes for which there has only been gene loss are shown in red, whereas those with gain or gain and loss, are shown in green. The distribution of the *Cyp6a* genes of this gene cluster throughout the 12 species are represented by vertical rectangles with rounded edges. Green rectangles represent gain, red rectangles loss. A red rectangle superimposed on a larger rectangle represents a loss of a gene that has representatives in other species. A red rectangle on its own represents a gene that has been gained but is in the genome as a nonfunctional copy. If a gene is gained or lost in internal branches of the species tree, it is represented as a longer rectangle that aligns to the species descended from that branch. The vertical dashed blue line indicates a break in microsynteny in the melanogaster group species. *Cyp6a18* is circled by a dotted line, as it is not part of the *Cyp6a* cluster but it is included in the diagram because it appears to have arisen from a duplication of *Cyp6a8* after the divergence of the *Drosophila* species.

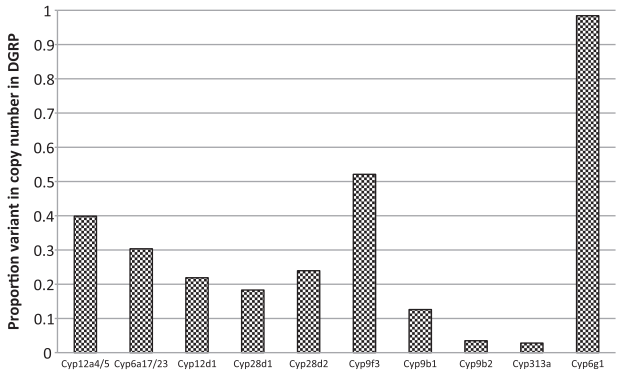

Fig. 4.—Copy number variation with *Drosophila melanogaster* P450 genes. P450 genes where CNV occurs in more than one DGRP line have their frequency in the DGRP represented.

nine adjacent P450 genes at position 10.7 Mb on chromosome arm 2R (Muller element C; fig. 3). Another cluster of three 6a genes exists on 2R position 4.4 Mb. The separation of these two clusters seems to have occurred in the ancestor of the *melanogaster* group species, as the orthologs of the two sets of genes are in one cluster in the other species such as *D. willistoni* that has a cluster of 14 *Cyp6a* genes (fig. 3).

There are some examples of gene origination by retrotransposition. One example is that of the Halloween genes *spook* (*Cyp307a1*) and *spookier* (*Cyp307a2*; Sztal et al. 2007). Another example is *Cyp6t1* that is derived from *Cyp6t2*. *Cyp6t2* is located on chromosome arm 2L (Muller element B) and contains one intron. An analysis including the additional eight genomes currently available on Flybase (http://flybase.org/blast/, last accessed April 30, 2014) suggests that the duplication generating the intronless *Cyp6t1*, occurred after the divergence of *D. eugracilis* from the *D. melanogaster* subgroup,

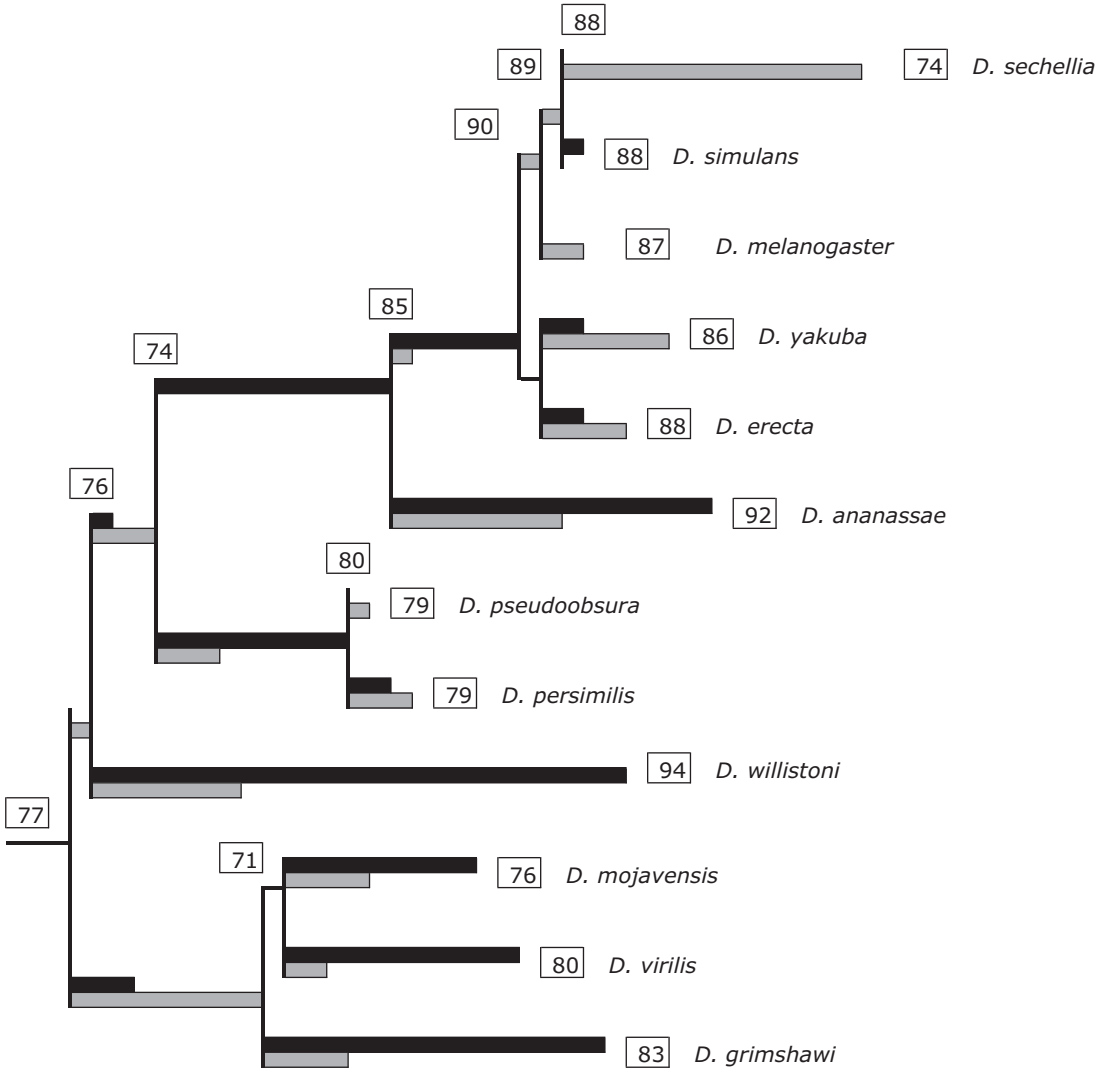

Fig. 5.—P450 gene gain and loss across the species phylogeny. P450 gene gain and loss is shown on the topology of the species phylogeny of the 12 *Drosophila* species analyzed. The length of the black bars is proportional to the number of gene gains in a lineage and the length of the gray bars is proportional to the number of gene losses. The number of functional P450 genes in each of the genomes is boxed next to the species name, as are the number of P450 genes inferred to be in the ancestor of each species.

and the new copy was retrotransposed to chromosome X. A nonfunctional copy of the gene on chromosome 2 (*Cyp6t2*) is in the genomes of *D. melanogaster*, *D. simulans*, *D. sechellia*, and *D. yakuba* whereas the intronless gene on chromosome X (*Cyp6t1*) is conserved in all descendent species.

There are also two notable examples where gene structure has changed since the divergence of the *Drosophila* species. The first potentially provides a novel example of subfunctionalization. *Cyp4d1* is the only P450 gene in *D. melanogaster* that exhibits alternate splicing. The two alternate first exons are conserved throughout all *Drosophila* species, except *D. mojavensis*, where the gene, except the most distal first exon is duplicated. It appears that the alternate first exons now exist in separate genes (fig. 6). The second notable example of gene structure change is in the *Cyp4e* subfamily. The same phase 0 intron has apparently been lost independently three times: In the *Cyp4e3* clade in *D. willistoni*, the *Cyp4e1/2* clade in *D. willistoni*, and the *Cyp4e1/2* clade in the obscura group species (supplementary fig. S1, Supplementary Material online). Perhaps this is evidence for interparalog exchange between *Cyp4e1/2* and *Cyp4e3* in *D. willistoni* (or between *Cyp4e1/2* and *Cyp4e3* in the *melanogaster* subgroup in which case it would be seen as intron gain), however the location of the genes suggests that such exchange would need to have occurred between genes on different

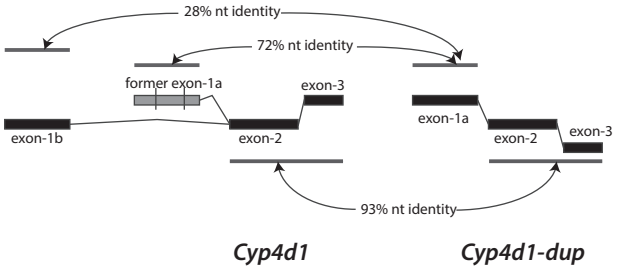

Fig. 6.—A gene duplication separating alternate splice forms into individual genes. The *Cyp4d1* gene has two splice forms in most of the examined *Drosophila* species that differ by the first exon used (exon 1a or exon 1b). In *D. mojavensis* a gene duplication appears to have involved exon 1a, exon 2, and exon 3. The presumed ancestral copy (*Cyp4d1*; in the left of the figure) has not retained a functional exon-1a (gray box with vertical lines representing multiple inactivating mutations).

chromosomes. Independent loss of the introns, without interparalog exchange, seems more likely.

The most striking examples of interparalog exchange occur in the analysis of structural variants within *D. melanogaster*. Two forms of chimera between the neighboring paralogs *Cyp6a17* and *Cyp6a23* were observed; one was observed in 8% of lines and the other in 16% of lines examined (fig. 7*a*). In both cases, the chimeric genes seem to replace both parental genes. In contrast a chimera of *Cyp12a4* and *Cyp12a5* is clustered with the two parental genes (fig. 7*b*) with 40% of the *D. melanogaster* lines. The *Cyp12a4* and *Cyp12a5* genes have previously been associated with resistance to the insecticide lufenuron (Bogwitz et al. 2005), so to test whether the CNV affects lufenuron resistance we compared the egg with adult viability of ten DGRP strains with the two gene "reference" haplotype to eight DGRP strains with the more complex three gene haplotype reared on lufenuron laced food. The difference between the two classes was significant (two tailed *t*-tested with unequal variance, $P = 0.034$) with the five most resistance lines having the three-gene haplotype (supplementary fig. S2, Supplementary Material online).

There is also one analogous case of chimeric genes in the nonmelanogaster data sets. This involves a recent polymorphic duplication in the *D. simulans* lineage. *Dsim_Cyp4ac1a* is found in multiple strains contributing to the original composite assembly of the *D. simulans* genome (Begun et al. 2007) and is similar to the *Cyp4ac1* gene over most of its length except for a small patch of 66 nt in which it is most similar to *Cyp4ac2* (supplementary fig. S3, Supplementary Material online).

Are There Signs of Adaptive Evolution in the Divergence Patterns of P450 Genes?

Relative Rates and Patterns of Coding Sequence Evolution

The 12 genome sequences allow us to examine the relative rates of sequence change among the orthologous groups of P450s. In supplementary table S1, Supplementary Material online, the orthologous groups are ranked by the number of amino acid substitutions observed per unit of time. For each orthologous set, we have calculated the tree length from a maximum-likelihood estimate using the program RAxML with the JTT matrix as substitution model. For our time estimates, we use the branch lengths of the species tree derived from whole-genome analysis as our proxy (Stark et al. 2007). If a P450 is missing from a branch or branches then those branches were not included in our estimate of "time." If a gene is duplicated in a particular clade then the time attributed to that clade is doubled in our calculation of rate. Eight of the ten slowest evolving genes are stable genes. This correlation between gene gain/loss events and divergence rates can be generalized across the data set as a whole (fig. 8). A notable outlier in this analysis is *Cyp306a1*, which exhibits no gene duplications or losses but which is one of the fastest evolving proteins since the $MRCA_D$. This gene is also known as *phantom*, it performs the second of many hydroxylations in the ecdysone synthesis pathway, and perhaps the relatively high divergence suggests it acts on multiple substrates.

The P450s exhibiting greater divergence in orthologous comparisons may have less selective constraint acting upon them or alternatively may have evolved for a period when natural selection favored amino acid change. To distinguish between these possibilities, we analyzed the ratio of nonsynonymous rate to the synonymous rate (ω) using the PAML software (Yang 2007). Our analysis of the P450 genes suggested the synonymous sites were saturated in comparisons between obscura and melanogaster groups so we limited these analyses to the six species of the melanogaster group.

Differences in ω Values between Different Groups of P450s

The simplest analysis of nonsynonymous to synonymous rates assumes that each P450 gene has a single ω value across all sites in the alignment and across all branches in the phylogenetic tree. Under this "one-ratio" model, the median ω values for all groups in all tests were well below one indicating purifying selection on all groups of genes. When the orthologous clades in the six species of the melanogaster group are ranked by their ω values, they largely concur when they are ranked by amino acid changes over the whole *Drosophila* phylogeny as calculated in the previous section (Spearman's rank correlation coefficient = 0.80). There are significant differences among P450 families (one-way ANOVA, $P = 4.0e\text{-}5$) and between stable and dynamic *AncD* clades (Wilcoxon $P = 1.92e\text{-}3$). We next looked for those orthologous groups that did not show a consistent rate of evolution across the melanogaster species group.

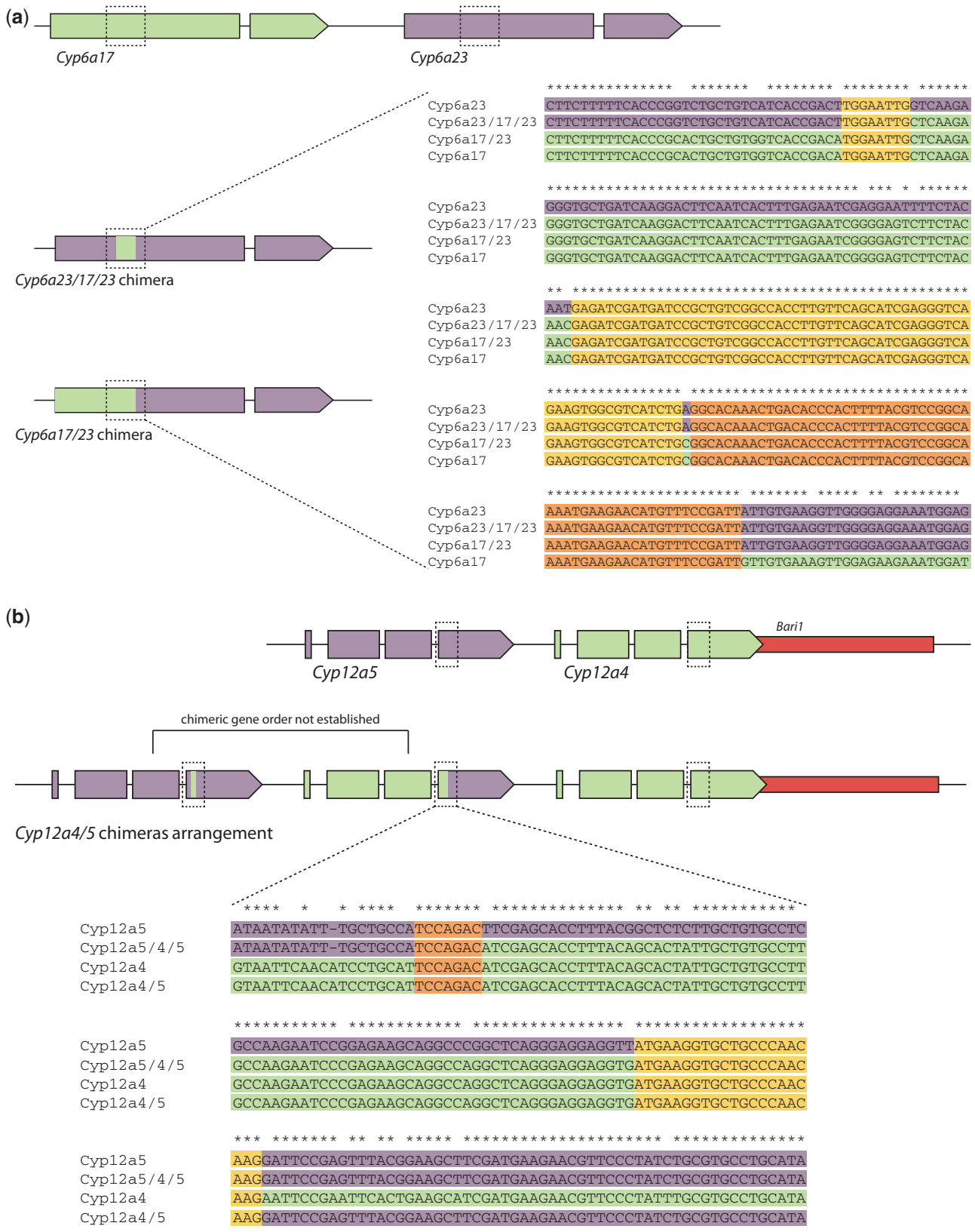

Fig. 7.—Structural polymorphisms in *Drosophila melanogaster* P450 genes (*a*). CNV at *Cyp6a17* and *Cyp6a23* in the DGRP. Two chimeras of *Cyp6a17* and *Cyp6a23* are observed in different DGRP lines. *Cyp6a23/17/23* is primarily *Cyp6a23* sequence, with a small portion of *Cyp6a17* sequence in the first

(continued)

Branch-Specific Models

To find genes that exhibit different rates in different branches a more sophisticated PAML analysis was performed which compares a model that allows the ω value to vary in different branches in a tree ("free ratio" model) to the "one ratio" model. In around one-third of the orthologous groups, the likelihood ratio test (LRT) comparing free-ratio and one-ratio model was significant using Bonferroni correction ($P < 0.0006$). These genes evolve at different rates in different lineages (supplementary table S2, Supplementary Material online). Notable among these are orthologous groups (*Cyp313a1/2/3*, *Cyp4d14*, *Cyp4d20*, *Cyp6a15*, *Cyp6t1/2*, *Cyp12d1*, and *Cyp28d2*) in which ω was greater than one in some lineages, suggesting that through at least part of the evolutionary history natural selection favored amino acid change in these orthologous groups in some lineages.

To reveal selective changes after gene duplication, we employed a model allowing different ω values before and immediately after duplication. Around 35% of the duplications studied showed a significant change in selective pressure after gene duplication (9 out of 26). In all but one of the duplications with a significant change, the ω ratio was increased immediately after duplication when compared with the ratio before duplication. The average ω ratio before duplication was 0.09 whereas the ratio immediately after duplication was 0.18. In six cases, the duplication occurred in an ancestral species rather than in a terminal branch of the *Drosophila* phylogeny. In these cases a third ω ratio could be calculated referring to the evolutionary rate after establishment of the two duplicates. The average for this ratio is 0.17, similar to the ratio immediately after duplication. Thus, selective constraints remained relaxed after establishment of duplicate P450 genes.

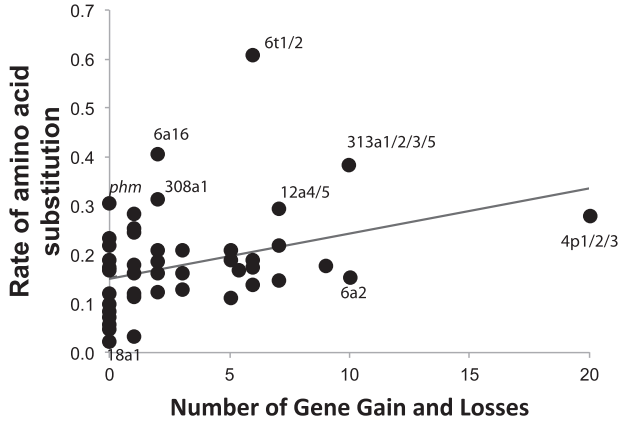

Fig. 8.—The relationship between gene gain/loss and rate of amino acid substitution. The *x* axis represents the number of gene gain and losses in the *AncD* clades. The *y* axis is the rate of amino acid substitution accounting for changes in evolutionary time due to gene gain and losses.

Site-Specific Models

We also compared models where ω was allowed to vary among sites. All but seven genes had a significant result testing variable selective pressure among sites (LRT comparing PAML models M3 and M0). However, there was only one orthologous group that showed evidence for positive selection acting on particular sites (LRT comparing PAML models M8 and M7); and that was *Cyp318a1*. The ω ratio for the site class allowing positive selection is 5.41. According to the Bayes Empirical Bayes method, three sites of *Cyp318a1* belong to the site class, which is under positive selection with a probability of >95%. These sites are in close proximity in the sequence at positions 442, 443, and 449 of the *D. melanogaster* protein. The structure of *CYP318A1* was homology modeled with respect to human *CYP3A4* (supplementary fig. S3, Supplementary Material online). These proteins have only 20.4% amino acid identity and so the modeling is tentative. The model suggests that the three sites most likely to be under positive selection are in a loop on the surface of the protein and are not in close proximity to the heme-binding site. Nor are these sites near the six recognized substrate recognition sites. Thus, it is unclear why changes at these sites may have been selectively favored.

Are There Molecular Evolutionary Correlates with Inducibility, Viability, Site of Expression, or Function of Cytochrome P450s?

A hypothesis that we wished to test is whether P450s that metabolize exogenous substrates are more likely to duplicate and evolve faster at more variable rates than those that metabolize endogenous substrates. Although some *Drosophila* P450s are believed to function on endogenous substrates (*spo-Cyp307a1*, *spok-Cyp307a2*, *phm-Cyp306a1*, *Cyp18a1*, *sad-Cyp315a1*, *shd-Cyp314a1*, *dib-Cyp302a1*, *nompH-Cyp303a1*, *Cyp301a1*, *Cyp6a17*, and *Cyp4g1*) and some have been linked to the metabolism of insecticides (*Cyp6g1*, *Cyp6a2*, and *Cyp12a4*) or environmental food substrates

exon. *Cyp6a17/23* has a *Cyp6a17*-derived 5′-end which shifts to *Cyp6a23* sequence in the first exon. The breakpoints of *Cyp6a23/17/23* (yellow highlights) and *Cyp6a17/23* (orange highlights) occur in regions of homology between parental and chimeric genes, suggesting conservation of frame and protein structure (*b*). CNV at *Cyp12a4* and *Cyp12a5* in the DGRP. In contrast to the reference genome arrangement (above), the alternate haplotype (below) contains two chimeric genes of *Cyp12a5* and *Cyp12a4*, and an intact copy of *Cyp12a4*. Both chimeras share a common breakpoint region (yellow), whereas a second breakpoint region (orange) is observed in the *Cyp12a5/4/5* chimera. The genomic order of *Cyp12a5/4/5* and *Cyp12a4/5* chimeras is not known, as this cannot be discerned from assembly of 454 reads.

(Bono et al. 2008), for most the substrates are unknown. However, we do know that RNAi-directed against nine P450s results in lethality (fig. 1; Chung et al. 2009), that 35 are transcribed in tissues implicated in xenobiotic metabolism (Chung et al. 2009), and that transcription can be induced in at least 21 with exposure to xenobiotics (Willoughby et al. 2006; Flybase). We examined these data sets with the rates and patterns of P450 evolution across the *Drosophila* phylogeny just described.

Most lethal and sublethal genes belonged to *AncD* clades with lower omega (ω) values than those that were not lethal when knocked down (Wilcoxon $P = 0.12$). P450s expressed in the detoxification (midgut: $\omega = 0.08$; Malpighian tubules: $\omega = 0.10$; midgut and Malpighian tubules: $\omega = 0.07$; midgut, Malpighian tubules, and fatbody: $\omega = 0.08$) and reproductive tissues ($\omega = 0.07$) have a significantly lower ω ratio than genes whose expression was not detected ($\omega = 0.13$) in the in situ studies of Chung et al. (2009). As noted by Chung et al. (2009), genes expressed in the hindgut have a significantly lower median ω ratio (0.04) than the genes belonging to the other categories. Of the 21 P450 genes that were found to be transcriptionally inducible by phenobarbital or caffeine (Willoughby et al. 2006) only one (*Cyp4d14*) is in the "stable" class which is fewer than expected if there was no relationship between inducibility and stability ($G = 11$, $P < 0.01$).

Finally, of the ten P450 genes for which we detected CNV within *D. melanogaster* all belong to the "dynamic/unstable" class of P450s (determined by interspecies comparison), 9 are expressed in the tissues where xenobiotics are metabolized and a higher than expected proportion are inducible by phenobarbital and/or caffeine ($G = 4.0$, $P < 0.05$). They also include two genes previously associated with insecticide resistance (*Cyp6g1* and *Cyp12a4/5*).

Discussion

In analyzing the rates and patterns of P450 gene gain and loss within and between *Drosophila* species, we are confronted with the question (that can be asked of many molecular evolutionary studies of multigene families): How much of the diversification in gene copy number that is observed is functional? At one extreme, copy number polymorphisms within a species may be targeted by natural selection, on the other extreme, copy number changes stochastically among species lineages and has no functional consequence. However, our analysis of the P450 genes of the *Drosophila* genus presented here suggests that the reality lies closer to the selectionist than the neutralist extreme and contrasts with an interpretation of phylogenetically deeper evolutionary comparisons of the same multigene family (Feyereisen 2011).

Although stochastic birth models may be found to adequately describe the process of gene proliferation over time, it is a mistake to argue that they discriminate clearly between models with and without selection. Furthermore, the correlation between function and duplicability of P450s that is observed is not expected under a model where divergence is predominantly driven by stochastic changes. With gene gain and loss events in the cytochrome P450 multigene family averaging ~0.006 events per million years in the *Drosophila* phylogeny (188 events/77 ancestral genes/400 Myr), it is more labile than the average *Drosophila* gene (0.0012 events per million years; Hahn et al. 2007). These events are not evenly distributed across the clades that are traceable to the ancestral *Drosophila* species, with 30 of the 77 ancestral clades having no gain or loss events. As has been noted before P450, genes currently associated with developmental functions have duplicated far less than those that have uncharacterized function or that have roles in xenobiotic metabolism (*Drosophila* 12 Genomes Consortium et al. 2007). We have extended upon this observation by showing that there is a correlation between the rate of amino acid replacement and the number of times a P450 has duplicated in the *Drosophila* phylogeny. These patterns are inconsistent with stochastic models that each gene is equally likely to have duplicated over evolutionary time.

Another argument against the redundancy model is that if a protein-coding gene was genuinely redundant then eventually, as it diverged from its functional paralog, it would accumulate an obvious inactivating mutation. The ratio of frameshifting mutations to nucleotide substitutions in a nonfunctional sequence has been estimated to be >1 in ten (Petrov and Hartl 1998; Robin et al. 2000) and the median number of nucleotide substitutions between *D. melanogaster* and *D. simulans* P450 orthologs is 27. So the chances are that if a P450 lacked a function it is likely to acquire an obvious inactivating mutation over this evolutionary time. *Drosophila simulans* has one P450 gene that is obviously inactivated (*Cyp6t2*) and *D. melanogaster* has three (*Cyp12d3*, *Cyp6t2*, and *Cyp6a15*) but each of these genes exists as conserved functional copies in other species; indicating they are not nonfunctional genes arising and disappearing without the influence of purifying and natural selection. In two of these cases (*Cyp6t2* and *Cyp12d3*), it seems that the loss is associated with a gain of another gene (*Cyp6t1* on the X chromosome and a *Cyp12d1* duplication that is polymorphic) that could be the complete functional replacement.

Patterns of P450 gene loss illustrate a striking and informative lineage effect in *D. sechellia*. P450s are not the only gene family to exhibit extensive loss in the *D. sechellia* lineage as the odorant receptor genes (McBride and Arguello 2007) and the glutathione *s*-transferases (Low et al. 2007) also show excessive loss in this lineage. *Drosophila sechellia* is a specialist species, found in a narrow ecological niche on the islands of the Seychelles (R'kha et al. 1991). So one hypothesis is that the gene loss can be attributed to a reduced chemical diversity in the narrow niche occupied by this island species and gene loss is neutral with respect to fitness. Indeed this has been proposed in a previous reports showing that P450 transcripts are

enriched among those that are downregulated in *D. sechellia* relative to *D. simulans* (Dworkin and Jones 2009; Wurmser et al. 2011). An alternative hypothesis is that the gene loss is associated with a severe reduction in population size in the history of this species, which has allowed slightly deleterious mutations, such as gene inactivating mutations, to become fixed in the population. This second hypothesis invokes the idea that the function of some genes is so minor that they are almost inconsequential, and they may be thought of as genes on the boundary of survival and extinction. If this were the case then perhaps the same genes would be lost multiple independent times across the phylogeny. In fact, of the 37 genes that have been lost somewhere on the phylogeny 22 have been lost more than once suggesting that they may be "genes on the boundary" of survival. This leads to the idea that repeated loss of a gene throughout a species radiation could be an indirect measure of the selective value of that gene. If it is readily dispensable, it would be of little value and inactivating mutations would be only slightly deleterious and so would be susceptible to population size fluctuations or genetic draft events (Gillespie 2001).

Previous comparative genomic studies of the P450 multigene family have noted lineage-specific amplification of particular genes, evocatively termed "blooms" (Feyereisen 2011). Such "blooms" are not a unique feature of P450s but are observed in many multigene families. However, in the taxonomically dense data set examined here, the P450 "blooms" previously identified as occurring in the *Drosophila* lineage (e.g., the *Cyp6a* genes) are no longer localized to a single branch in the species tree (four of the nine *melanogaster* genes in the large *Cyp6a* cluster arose after the ancestral *Drosophila*). In fact the most labile of *AncD* gene clades, the *Cyp4p*, is fairly unremarkable if the focus is on a particular branch, as the 19 duplications are distributed across the species tree. All of the *Cyp4p* duplications have arisen by unequal recombination and all are in the intron of the *hikaru genki* gene. Has this occurred because there is a mutational predisposition that has increased relative to other genes? If that were the case, then perhaps we would see CNV of the *Cyp4p*s within the *D. melanogaster* population—and yet we have detected none.

An alternate model to explain the *Cyp4p* phylogeny would be that there were far fewer gene duplications (maybe as few as three) and that recurrent subgene interparalog exchange (e.g., gene conversion) made genes within a species cluster together on the phylogenetic analysis. The frequency of such exchange events would be rare relative to sequence divergence as the *Cyp4p* paralogs within a species are substantially diverged across the whole length of the gene.

However, the patterns of copy number polymorphism suggest interparalog exchange does arise. We have identified multiple cases of polymorphic within-gene-family chimeras in the within-species data sets, not just in *D. melanogaster* (fig. 7) but also in *D. simulans* (supplementary fig. S3, Supplementary Material online). Such interparalog chimeras have been seen before segregating within *H. armigera* populations (Joussen et al. 2012) and there is some evidence for such exchange in other multigene families (Robin et al. 2009; Runck et al. 2009). However, the repeated observation of chimeras in our population data sets suggests that interparalog exchange may be more common than previously thought and that evidence for such nonallelic recombination may be obscured in more distant evolutionary comparisons by subsequent molecular events.

The occurrence of NAHR does not mean that the gene sequences involved are redundant or functionally equivalent. In humans such events are thought to be deleterious (Dumont and Eichler 2013). In the *Cyp337b* example of *H. armigera*, they can be adaptive (Joussen et al. 2012). Our finding of a chimera, at a locus previously implicated in lufenuron resistance, *Cyp12a4/5* (Bogwitz et al. 2005), motivated a preliminary experiment to show that there is a correlation between this chimera and lufenuron resistance. Lufenuron is an insecticide primarily used to control fleas and so it seems unlikely to be a selective agent; however, it demonstrates a functionality that may be selected in response to some other environmental toxin.

Another evolutionary mechanism proposed to explain the dynamics of multigene family evolution is the "subfunctionalization" model, in which daughter genes divide the functions performed by the parental gene in a complementary fashion (Force et al. 1999). We detected two types of events consistent with this model. Firstly, duplicate genes maintained increased d_N/d_S (ω) ratios for substantial periods of time after the gene duplication events, consistent with a relaxation of selective constraint. Secondly, a subfunctionalization event is suggested by the separation of alternate splice forms in *Cyp4d1* as seen in most *Drosophila* species in two separate genes in *D. mojavensis*.

Although the analysis of the P450 gene family evolution within the genus *Drosophila* has been informed by functional analyses of P450s (particularly in *D. melanogaster*), the reverse is also true, in that the evolutionary analyses informs us about P450 function. To some extent, the inactivated P450 genes in *D. sechellia* can be considered natural knockouts and that may motivate biological comparisons with closely related species. For example, a naturally occurring null allele of one of them, *Cyp6a20*, occurs at high frequency in *D. melanogaster* and is associated with male aggression (Dierick and Greenspan 2006; Robin et al. 2007; Wang et al. 2008). *Drosophila sechellia* may therefore be a useful species to include in studies trying to identify the substrate that the *Cyp6a20* enzyme works on. Similarly, the accelerated rate of amino acid change in the *phantom* gene should motivate studies on this important gene in ecdysteroid synthesis. Strengthening this motivation is molecular population genetic evidence that *phm* (*Cyp306a1*) has been the target of recent natural selection in the *D. melanogaster* lineage (Orengo and Aguade 2007).

The *Cyp12a4*/*Cyp12a5* chimera and the evidence of positive selection in *Cyp318a1* should also motivate studies into their function and their substrates.

In conclusion, our analyses suggest that the overwhelming majority of P450 paralogs in *Drosophila* have a *raison d'etre* based on a function determined by natural selection. We argue that when paralogs have diverged sufficiently from each other, selective neutrality should not be assumed without being demonstrated. Furthermore, it is clear that if we are to fully understand multigene family evolution, functional genomics needs to expand beyond evolutionary analyses to ecological analyses of gene function.

Supplementary Material

Supplementary tables S1 and S2, figures S1–S3, and data files S1–S4 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

R.T.G. retrieved the genomic sequences, identified, and annotated the P450 genes. L.G. performed the PAML analyses and with C.R. performed the other phylogenetic analyses. R.T.G. and P.B. characterized the copy number variation within *D. melanogaster*. P.B. performed the insecticide bioassays. T.S. characterized the *D. melanogaster* pseudogenes. C.R. performed analyses, supervised all components of the research, and wrote the paper with contributions from L.G. and P.B. This work was supported by Australian Research Council Discovery Project grant DP0557497 to C.R. and P.B.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Ames RM, Money D, Ghatge VP, Whelan S, Lovell SC. 2012. Determining the evolutionary history of gene families. Bioinformatics 28:48–55.

Amichot M, et al. 2004. Point mutations associated with insecticide resistance in the *Drosophila* cytochrome P450 *Cyp6a2* enable DDT metabolism. Eur J Biochem. 271:1250–1257.

Begun DJ, et al. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol. 5: e310.

Bogwitz MR, et al. 2005. *Cyp12a4* confers lufenuron resistance in a natural population of *Drosophila melanogaster*. Proc Natl Acad Sci U S A. 102: 12807–12812.

Bono JM, Matzkin LM, Castrezana S, Markow TA. 2008. Molecular evolution and population genetics of two *Drosophila mettleri* cytochrome P450 genes involved in host plant utilization. Mol Ecol. 17:3211–3221.

Chung H, et al. 2009. Characterization of *Drosophila melanogaster* cytochrome P450 genes. Proc Natl Acad Sci U S A. 106:5731–5736.

Claudianos C, et al. 2006. A deficit of detoxification enzymes: pesticide sensitivity and environmental response in the honeybee. Insect Mol Biol. 15:615–636.

Coen E, Strachan T, Dover G. 1982. Dynamics of concerted evolution of ribosomal DNA and histone gene families in the melanogaster species subgroup of *Drosophila*. J Mol Biol. 158:17–35.

Daborn PJ, et al. 2002. A single p450 allele associated with insecticide resistance in *Drosophila*. Science 297:2253–2256.

Dierick HA, Greenspan RJ. 2006. Molecular analysis of flies selected for aggressive behavior. Nat Genet. 38:1023–1031.

Drosophila 12 Genomes Consortium, et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450:203–218.

Dumont BL, Eichler EE. 2013. Signals of historical interlocus gene conversion in human segmental duplications. PLoS One 8:e75949.

Dworkin I, Jones CD. 2009. Genetic changes accompanying the evolution of host specialization in *Drosophila sechellia*. Genetics 181:721–736.

Feyereisen R. 2005. Insect cytochrome P450. In: Comprehensive insect science. Pergamon. p. 406.

Feyereisen R. 2011. Arthropod CYPomes illustrate the tempo and mode in P450 evolution. Biochim Biophys Acta. 1814:19–28.

Force A, et al. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545.

Fujii S, Toyama A, Amrein H. 2008. A male-specific fatty acid omega-hydroxylase, SXE1, is necessary for efficient male mating in *Drosophila melanogaster*. Genetics 180:179–190.

Gilbert LI. 2004. Halloween genes encode P450 enzymes that mediate steroid hormone biosynthesis in *Drosophila melanogaster*. Mol Cell Endocrinol. 215:1–10.

Gillespie JH. 2001. Is the population size of a species relevant to its evolution? Evolution 55:2161–2169.

Guittard E, et al. 2011. *CYP18A1*, a key enzyme of *Drosophila* steroid hormone inactivation, is essential for metamorphosis. Dev Biol. 349: 35–45.

Hahn MW. 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. J Hered. 100:605–617.

Hahn MW, Han MV, Han S-G. 2007. Gene family evolution across 12 *Drosophila* genomes. PLoS Genet. 3:e197.

Hardstone MC, Baker SA, Gao JW, Ewer J, Scott JG. 2006. Deletion of Cyp6d4 does not alter toxicity of insecticides to *Drosophila melanogaster*. Pestic Biochem Physiol. 84:236–242.

Harrop TWR, et al. 2014. Evolutionary changes in gene expression, coding sequence and copy-number at the cyp6g1 locus contribute to resistance to multiple insecticides in *Drosophila*. PLoS One 9:e84879.

Helvig C, Tijet N, Feyereisen R, Walker FA, Restifo LL. 2004. *Drosophila melanogaster* CYP6A8, an insect P450 that catalyzes lauric acid (omega-1)-hydroxylation. Biochem Biophys Res Commun. 325: 1495–1502.

Hillenmeyer ME, et al. 2008. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. Science 320:362–365.

Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. Proc R Soc B Biol Sci U S A. 256:119–124.

Joussen N, et al. 2012. Resistance of Australian *Helicoverpa armigera* to fenvalerate is due to the chimeric P450 enzyme CYP337B3. Proc Natl Acad Sci U S A. 109:15206–15211.

Kang J, Kim J, Choi K-W. 2011. Novel cytochrome P450, cyp6a17, is required for temperature preference behavior in *Drosophila*. PLoS One 6:e29800.

Langley CH, et al. 2012. Genomic variation in natural populations of *Drosophila melanogaster*. Genetics 192:533–598.

Low WY, et al. 2007. Molecular evolution of glutathione-*S*-transferases in the genus *Drosophila*. Genetics 177:1363–1375.

Mackay TFC, et al. 2012. The *Drosophila melanogaster* genetic reference panel. Nature 482:173–178.

McBride CS, Arguello JR. 2007. Five *Drosophila* genomes reveal nonneutral evolution and the signature of host specialization in the chemoreceptor superfamily. Genetics 177:1395–1416.

McDonnell CM, et al. 2012. Evolutionary toxicogenomics: diversification of the *Cyp12d1* and *Cyp12d3* genes in *Drosophila* species. J Mol Evol. 74: 281–296.

Namiki T, et al. 2005. Cytochrome P450 *CYP307A1/spook*: a regulator for ecdysone synthesis in insects. Biochem Biophys Res Commun. 337: 367–374.

Nelson DR. 2006. Cytochrome P450 nomenclature, 2004. Methods Mol Biol. 320:1–10.
Novozhilov AS, Karev GP, Koonin EV. 2006. Biological applications of the theory of birth-and-death processes. Brief Bioinform. 7:70–85.
Orengo DJ, Aguade M. 2007. Genome scans of variation and adaptive change: extended analysis of a candidate locus close to the phantom gene region in *Drosophila melanogaster*. Mol Biol Evol. 24: 1122–1129.
Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. Mol Biol Evol. 5:568–583.
Petrov DA, Hartl DL. 1998. High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. Mol Biol Evol. 15: 293–302.
Pollard DA, Iyer VN, Moses AM, Eisen MB. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. PLoS Genet. 2:1634–1647.
Prince VE, Pickett FB. 2002. Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet. 3:827–837.
Qiu Y, et al. 2012. An insect-specific P450 oxidative decarbonylase for cuticular hydrocarbon biosynthesis. Proc Natl Acad Sci U S A. 109: 14858–14863.
Rai BK, Madrid-Aliste CJ, Fajardo JE, Fiser A. 2006. MMM: a sequence-to-structure alignment protocol. Bioinformatics 22:2691–2692.
Ranson H, et al. 2002. Evolution of supergene families associated with insecticide resistance. Science 298:179–181.
Reed WJ, Hughes BD. 2004. A model explaining the size distribution of gene and protein families. Math Biosci. 189:97–102.
Rewitz KF, Rybczynski R, Warren JT, Gilbert LI. 2006a. Identification, characterization and developmental expression of Halloween genes encoding P450 enzymes mediating ecdysone biosynthesis in the tobacco hornworm, *Manduca sexta*. Insect Biochem Mol Biol. 36: 188–199.
Rewitz KF, Rybczynski R, Warren JT, Gilbert LI. 2006b. The Halloween genes code for cytochrome P450 enzymes mediating synthesis of the insect moulting hormone. Biochem Soc Trans. 34: 1256–1260.
Rewitz KF, O'Connor MB. 2011. Timing is everything: PTTH mediated DHR4 nucleocytoplasmic trafficking sets the tempo of *Drosophila* steroid production. Front Endochronol. 2:108.
R'kha S, Capy P, David JR. 1991. Host plant specialization in the *Drosophila melanogaster* species complex—a physiological, behavioral, and genetic analysis. Proc Natl Acad Sci U S A. 88:1835–1839.
Robin C, Bardsley LMJ, Coppin C, Oakeshott JG. 2009. Birth and death of genes and functions in the beta-esterase cluster of *Drosophila*. J Mol Evol. 69:10–21.
Robin C, Daborn PJ, Hoffmann AA. 2007. Fighting fly genes. Trends Genet. 23:51–54.
Robin GC, Russell RJ, Cutler DJ, Oakeshott JG. 2000. The evolution of an alpha-esterase pseudogene inactivated in the *Drosophila melanogaster* lineage. Mol Biol Evol. 17:563–575.
Runck AM, Moriyama H, Storz JF. 2009. Evolution of duplicated beta-globin genes and the structural basis of hemoglobin isoform differentiation in Mus. Mol Biol Evol. 26:2521–2532.
Sackton TB, et al. 2007. Dynamic evolution of the innate immune system in *Drosophila*. Nat Genet. 39:1461–1468.
Schmidt JM, et al. 2010. Copy number variation and transposable elements feature in recent, ongoing adaptation at the *Cyp6g1* locus. PLoS Genet. 6:e1000998.
Shah N, Dorer DR, Moriyama EN, Christensen AC. 2012. Evolution of a large, conserved, and syntenic gene family in insects. G3 (Bethesda) 2: 313–319.
Stark A, et al. 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. Nature 450:219–232.
Strode C, et al. 2008. Genomic analysis of detoxification genes in the mosquito *Aedes aegypti*. Insect Biochem Mol Biol. 38:113–123.
Sztal T, et al. 2007. Two independent duplications forming the *Cyp307a* genes in *Drosophila*. Insect Biochem Mol Biol. 37: 1044–1053.
Sztal T, et al. 2012. A cytochrome p450 conserved in insects is involved in cuticle formation. PLoS One 7:e36544.
Thomas JH. 2007. Rapid birth–death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. PLoS Genet. 3:e67.
Thompson JD, Higgins DG, Gibson TJ. 1994. Clustal-W—improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.
Tijet N, Helvig C, Feyereisen R. 2001. The cytochrome P450 gene superfamily in *Drosophila melanogaster*: annotation, intron–exon organization and phylogeny. Gene 262:189–198.
van Hoof A. 2005. Conserved functions of yeast genes support the duplication, degeneration and complementation model for gene duplication. Genetics 171:1455–1461.
Wang L, Dankert H, Perona P, Anderson DJ. 2008. A common genetic target for environmental and heritable influences on aggressiveness in *Drosophila*. Proc Natl Acad Sci U S A. 105:5657–5663.
Willingham AT, Keil T. 2004. A tissue specific cytochrome P450 required for the structure and function of *Drosophila* sensory organs. Mech Dev. 121:1289–1297.
Willoughby L, et al. 2006. A comparison of *Drosophila melanogaster* detoxification gene induction responses for six insecticides, caffeine and phenobarbital. Insect Biochem Mol Biol. 36:934–942.
Wu D-D, Irwin DM, Zhang Y-P. 2011. Correlated evolution among six gene families in *Drosophila* revealed by parallel change of gene numbers. Genome Biol Evol. 3:396–400.
Wurmser F, et al. 2011. Population transcriptomics: insights from *Drosophila simulans*, *Drosophila sechellia* and their hybrids. Genetica 139:465–477.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.
Zmasek CM, Eddy SR. 2001. A simple algorithm to infer gene duplication and speciation events on a gene tree. Bioinformatics 17: 821–828.

Associate editor: Esther Betran