Prediction of sites under adaptive evolution in flavin-containing monooxygenases: Selection pattern revisited

HAO DaCheng^{1*} & XIAO PeiGen^{2*}

¹Biotechnology Institute, Dalian Jiaotong University, Dalian 116028, China;

²Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100193, China

Received December 8, 2010; accepted January 5, 2011

Flavin-containing monooxygenase (FMO), like cytochrome P450 (CYP), is a monooxygenase that uses the reducing equivalents of NADPH to reduce one atom of molecular oxygen to water, while the other atom is used to oxidize the substrate. Recently, it was shown that some CYP isoforms have been subject to positive selection. However, it is unknown whether the highly conserved phase I detoxification enzyme, FMO, has undergone similar positive Darwinian selection. We used maximum-likelihood models of codon substitution, evolutionary fingerprinting, and cross species comparison to investigate the occurrence of adaptive evolution in FMO sequences. We used recent genomic data from a range of species, including vertebrates and invertebrates. We present the evidence for the occurrence of adaptive evolution in mammalian FMO 3, 4, 5, and fugu FMOs but not in mammalian FMO 1, FMO 2, frog FMOs, other fish FMOs and invertebrate FMOs. The sites under adaptive evolution were significantly associated with the insertion domain in mammalian FMO 5. We identified specific amino acid sites in FMOs 3–5 that are likely targets for selection based on the patterns of parallel amino acid change. The most likely role of adaptive evolution is the repair of mutations that permitted optimal NADP⁺ binding and improved catalytic efficiency. The occurrence of positive selection during the evolution of phase I detoxification enzymes such as FMOs 3–5 and fugu FMO suggests the occurrence of both high selection pressure acting on species within their unique habitats and significant changes in intensity and direction (forms of xenobiotics and drugs) resulting from changes in microhabitat and food.

flavin-containing monooxygenase, adaptive evolution, positive selection, insertion domain, evolutionary fingerprint, maximum likelihood

Citation: Hao D C, Xiao P G. Prediction of sites under adaptive evolution in flavin-containing monooxygenases: Selection pattern revisited. Chinese Sci Bull, 2011, 56: 1246–1255, doi: 10.1007/s11434-011-4380-8

Flavin-containing monooxygenases (FMOs, E.C. 1.14.13.8) oxygenate the nucleophilic O, N, S and Se atoms of a wide range of substrates, including amines, amides, thiols and sulfides. FMO, like cytochrome P450 (CYP), is a monooxygenase that uses the reducing equivalents of NADPH to reduce one atom of molecular oxygen to water, while the other atom is used to oxidize the substrate. Similarly, FMO and CYP also share similar tissue and cellular distributions, molecular weights, substrate specificities, and both exist as multiple enzymes that are under developmental control. The human FMO functional gene family is much smaller

(5 families each with a single member) than the CYP family. Furthermore, FMO does not require a reductase to transfer electrons from NADPH and the catalytic cycle of the two monooxygenases also differs significantly. The FMOs associated with liver microsomes are divided into 3 classes: FMOs, *N*-hydroxylating monooxygenases, and Baeyer-Villiger monooxygenases (BVMOs). BVMO and FMO may be distinguished by their signature sequences FXGXXXHXX-XW(P/D) and FXGXXXHXXX(Y/F), respectively. In general, CYP is the primary contributor to oxidative xenobiotic metabolism. However, an increasing number of drugs and xenobiotic substances are known to be metabolized by FMOs [1,2]. FMO and CYP have overlapping substrate specificities,

*Corresponding authors (email: hao@djtu.edu.cn; xiaopg@public.bta.net.cn)

but often yield distinct metabolites with potentially significant toxicological/pharmacological consequences.

An increasing number of researchers study gene function and regulation using population genetics and phylogenomics techniques. This approach is aided by the rapidly expanding effort towards genome sequencing and genotyping. The evolution of a gene's sequence among human populations or species reflects, in part, adaptations to changing physiology or environment. Recently, Qiu et al. [3] investigated the evolution of the *CYP3* genomic loci using genomic sequences from 16 species, and found 2 recent episodes of particularly strong positive selection acting on primate *CYP3A4* and *3A7* protein-coding sequences. Zawaira et al. [4] used maximum-likelihood models of codon substitution to investigate the role of adaptive evolution in the evolution of CYP sequences, and found evidence for the occurrence of adaptive evolution in the evolution of rat CYP2C, rabbit CYP2C, rat CYP2D, human CYP3A, and rabbit CYP4A. In addition, Chen et al. [5] conducted a comprehensive study of the nucleotide diversity and haplotype structure of the *CYP3A* locus and concluded that *CYP3A4* and *3A7* had recently undergone, or were undergoing, a selective sweep in all 3 populations under study. However, *CYP3A43* and *3A5* were undergoing a selective sweep in non-Africans and Caucasians, respectively.

In contrast to the depth of information on the CYP family, little is known about the evolution of FMOs. Allerston et al. [6] analyzed the genetic diversity within FMO 3 and found an excess of intermediate-frequency SNPs and haplotypes, a ragged pairwise mismatch distribution, and an excess of replacement polymorphisms, providing evidence that FMO 3 has been the subject of balancing selection. Hao et al. [7] reconstructed a phylogenetic tree for the FMO family and characterized the long-term evolution and functional divergence followed by members of this family. The authors noted that there is extensive silent divergence at the nucleotide level suggesting that this family has been subject to strong purifying selection at the protein level. However, recently published genomic data from a range of species, including vertebrates and invertebrates, have not yet been fully evaluated with respect to determining the evolution of

FMOs. The whole genome sequencing of a rapidly increasing number of species facilitates studies of the evolution of gene families using a comparative genomic approach. Given the similarities between FMO and CYP, we hypothesize that FMO is also subject to positive Darwinian selection. We test this hypothesis using a maximum-likelihood approach and cross species sequence comparison.

1 Methods

1.1 Orthologous sequences

We retrieved 106 sequences from Ensembl (<http://www.ensembl.org>) (Dataset S1 and Table 1).

1.2 Paralogous sequences

We used 46 sequences, including 9 *Strongylocentrotus purpuratus* (purple sea urchin) sequences (Dataset S1 and Table S1).

The protein database predicted from the *S. purpuratus* genome (release 2.1) [8] was searched using hidden Markov models (HMMs) of FMOs constructed with Hmmer 2.3 [9]. The HMM was constructed with FMOs from humans, mice, rats, and fish. Both global and local multidomain HMMs were constructed and used to search the predicted protein database. Predicted proteins were aligned with known FMOs using Clustal W2 [10]. Examination of the *S. purpuratus* genome assembly using Blast searching confirmed the FMO protein sequence predictions. Genome sequences were obtained from GenBank, Ensembl, and the Joint Genome Institute. Some preliminary gene predictions were performed using Genewise [11] and Genscan [12].

1.3 Sequence alignment and phylogeny reconstruction

DNA sequence and codon alignments were performed with RevTrans [13] (<http://www.cbs.dtu.dk/services/RevTrans/>) and Clustal W2. For amino acid sequences, we used the neighbor-joining (NJ) method [14] to reconstruct the phylogenetic tree with MEGA4 [15]. The best model, JTT + G,

Table 1 Likelihood ratio statistics and AICc scores for tests of positive selection of orthologous FMO sequences^{a)}

Gene family (no. of genes)	M8 vs. M8a (df=1)		MEC vs. M8a		M8 vs. M7 (df=2)	
	Log-likelihood	P	Log-likelihood	AICc	Log-likelihood	P
FMO 1(24)	-12840.2/-12840.2	Not significant	-12741.7/-12840.2	25493.4/25688.4	-12840.2/-12846.1	0.002739
FMO 2(23)	-12555.8/-12579.4	0.000	-12515.5/-12579.4	25041.0/25166.8	-12555.8/-12586.7	0.000
FMO 3(18)	-10488.1/-10491.4	0.010198	-10417.9/-10491.4	20845.8/20990.8	-10488.1/-10500.5	0.000004
FMO 4(19)	-12228.4/-12241.4	0.000	-12115/-12241.4	24240.0/24490.8	-12228.4/-12255.6	0.000
FMO 5(22)	-13188.5/-13206.1	0.000	-13022.5/-13206.1	26055.0/26420.2	-13188.5/-13224.1	0.000

a) $-2\Delta\ln L = 2(\ln L_{\text{alternative hypothesis}} - \ln L_{\text{null hypothesis}})$, χ^2 distribution. $\text{AICc} = -2 \log L + 2pN/(N-p-1)$; L , likelihood; P , no. of free parameters; N , the sequence length. The smaller the AICc value, the better the model explains the data. For FMOs 1–5, the likelihood ratio test (LRT) between the models MEC and M8a indicates a significance level of: AICc score of MEC is lower than M8a, significance test passed. Sites under positive selection identified by M8 and MEC are listed in Table 2. The occurrence of positive selection in FMO 2 was not supported by the evolutionary fingerprint analysis (Figure 2) and parallel amino acid substitution (Table S3).

Table 2 Positively selected amino acid sites in orthologous mammalian FMOs^{a)}

	FMO 1	FMO 2	FMO 3	FMO 4	FMO 5
SLAC	243L	None	92I 116H 236V 350I	311V	253A 255K 407A 410S 504V 88V 232A 240L 253A 255K 263Q 343V 382S 407A 410S 412A 415E 504V
FEL	243L 411A 417P 509S	206N 489 ^{b)}	92I 111S 116H 236V 318I 350I 360L 378A 459M 506Q 512F 527V	121E 218T 244A 278K 296I 311V	
REL	243L 411A 417P	None	35N 92I 93A 109F 114N 121T 164N 181V 235L 236V 318I 332S 360L 417R 432V 459M 500R 503G 504R 512F 527V	89D 121E 221L 226D 231Y 235V 240C 244A 256Q 259K 281F 311V 323V 354Q 474M 501I 517A 537L 543K 545Q 556 ^{b)} 553V 554S 555L 556W 558G 89D 221L 226D 231Y 235V 259K 281F 323V 354Q 537L 545Q 554S 561Q (platypus) ^{c)} 85E 89D 121E 136N 154H 221L 226D 231Y 235V 259K 279A 281F 323V 354Q 374K 474M 501I 537L 540V 542D 543K 545Q 555V(platypus) ^{c)} 553V 554S 555L 556W 560Q(platypus) ^{c)} 561Q (platypus) ^{c)}	132S 182G 232A 242H 248C 250Q 253A 255K 256Y 263Q 297L 410S 412A 471I 504V 506R 511T 513T 182G 232A 235L 236F 241T 242H 244I 245W 248C 250Q 252L 253A 255K 256Y 260K 297L 492D 510M 513T
M8	131H	240R	360L 504R		
MEC	131H 200A 411A 417P 430S 434T 442Y 458H	180D 252V 441L 449F 450C 506G(platypus) ^{c)}	35N 74F 93A 135K 235L 360L 417R 432V 459M 498Q 504R		106R 182G 232A 235L 241T 242H 244I 245W 248C 250Q 252L 253A 255K 256Y 260K 297L 492D 510M 513T

a) Sites under positive selection detected by various models are listed according to the human sequence numbering. Positively selected sites in FMOs 3–5 detected by M8 model are in bold type, some of which are also detected by other methods. b) Alignment position; no amino acid for human FMOs aligned on this position. c) No amino acid for human FMOs aligned on this position (according to the platypus sequence numbering); the amino acid for platypus FMOs at this position is shown.

was identified using ProtTest [16]. To determine whether the outcome was dependent on this choice, we conducted a phylogenetic inference analysis by reconstructing a Bayesian tree (MrBayes 3.1.2) [17] with 4 Markov chain Monte Carlo chains run for 1×10^6 generations. For the nucleotide sequences, we used a Bayesian analysis and the maximum likelihood (ML) method (GARLI) [18] to infer the phylogenetic trees. The best model, GTR + I + G, was selected using ModelTest 3.8. Bayesian probabilities were obtained under this model using a Markov chain Monte Carlo simulation (MCMC: 4 chains, each runs for 4×10^6 generations). We used random trees as the starting point and sampling every 500th generation. To test the reliability of the obtained topologies, we calculated bootstrap probability (BP) and posterior probability (PP) values for each internal branch, assuming that $BP \geq 80\%$ and $PP \geq 95\%$ were statistically significant. *S. pombe* FMO [19] and the FMO from *Saccharomyces cerevisiae* [20] were assigned as outgroups in the reconstructions.

1.4 Detection of positively selected sites and evolutionary fingerprint analyses

We tested for evidence of positive selection by comparing the nonsynonymous substitution rate (d_N) to the synonymous substitution rate (d_S). If a gene is evolving neutrally, $\omega = d_N/d_S$ is expected to equal 1, whereas $\omega > 1$ is considered strong evidence that a gene has experienced positive selection. We used several ML approaches to test for evidence of positive selection on FMOs. The first approach, developed by Yang et al. [21] (referred to as Yang models), involves comparisons of a neutral codon substitution model with ω constrained to be ≤ 1 to a selection model where a class of sites has $\omega > 1$. Because neutral models are nested within the

corresponding selection models, a likelihood ratio test (LRT) can be used to compare the two. The test statistic $-2\ln L$ ($\Delta \ln L$ = the difference in log likelihoods of the 2 models) follows the χ^2 distribution with degrees of freedom (df) equal to the difference in number of parameters between models. In the specific models implemented, ω varies between codons as a beta distribution (neutral: M7, M8a; selection: M8). We implemented models M7, M8a, and M8 with the codeml program in PAML4 [22]. Because Yang models are based on theoretical assumptions and ignore the empirical observation that distinct amino acids differ in their replacement rates, we also implemented the MEC (Mechanistic Empirical Combination) model [23]. This takes into account not only the transition-transversion bias and the nonsynonymous/synonymous ratio, but also the different amino acid replacement probabilities as specified in empirical amino acid matrices. Because the LRT is applicable only when two models are nested and is therefore not suitable for comparing MEC and M8a models, we used the second-order AIC (AICc) for comparisons [23]. Those sites that are most likely to be in the positive selection class ($\omega > 1$) are identified as likely targets of selection.

Although the Yang models allow for variation in the nonsynonymous substitution rate, the synonymous rate is fixed across the sequence. A number of methods have been proposed for detecting positive selection that allow for variation in the synonymous rate (e.g., fixed effects methods and random effects methods). The fixed effect likelihood (FEL) method [24] estimates ω on a site-by-site basis, uses ML estimation, and treats shared parameters (branch lengths, tree topology, and nucleotide substitution rates) as fixed. The random effects likelihood (REL) method is similar to the Yang model M3. However, both nonsynonymous and synonymous rates vary as gamma distributions with 3

rate classes [24]. We implemented the REL, FEL, and SLAC methods using the web interface DATAMONKEY [25].

Over time, natural selection molds every gene into a unique mosaic of sites that evolve rapidly or resist change, an “evolutionary fingerprint” of the gene. Pond et al. [26] developed a novel model for coding sequence evolution that uses a general bivariate discrete parameterization of the evolutionary rates. This approach provides a better fit to the data using a smaller number of parameters than existing models. We used the probability distributions generated using this method to represent the evolutionary fingerprints of the FMOs under consideration.

1.5 Parallel amino acid substitutions

Parallel and convergent evolution refer to independent acquisitions of the same character state on more than one occasion during evolution. The distinction between parallelism and convergence is that the former refers to the situation in which the ancestral states were identical among independent lineages, whereas the latter requires different ancestral states [27]. To identify parallel amino acid substitutions, we performed ML reconstructions of ancestral sequences and individual mutation events using PAML4 (baseml and pamp). The marginal reconstruction approach [28] compares the probabilities of different character assignments to an interior node at a site and selects the character that has the highest PP. We then calculated the number of independent changes, Grantham’s distance [29] between the starting and ending amino acid, the universal evolutionary index (EI) [30], and the possible alternative amino acid substitutions for the identified parallel amino acid substitutions.

1.6 Statistical significance of the overlap between sites under adaptive evolution and defined insertion domain

S. pombe FMO (447 aa) consists of two structural domains [19]. Residues 176–291 form a small structural domain, i.e., the insertion domain (ID), with the remainder of the polypeptide chain forming a larger single domain. The prokaryotic FMO from *Methylophaga* sp. strain SK1 is made up of 2 distinct domains [31], a larger FAD-binding domain (residues 1–169 and 281–461) and a smaller NADP-binding domain (residues 170–280, corresponding to the ID of the yeast FMO). The positions of the ID in mammalian FMOs were assigned according to the sequence alignment with yeast and bacterial FMOs (Figure 1). In addition, we defined the active sites (ASs) of mammalian FMOs by extending the yeast and bacterial FMO ASs by 3 amino acid residues on both sides (Figures 1 and S1).

We constructed a model for analyzing the statistical significance of the overlap between the sites under adaptive evolution and the ID as follows. Let X be the total number of sites (for analysis of a given gene or set of genes as listed in the first column of Tables 1 and S1) predicted to be under

adaptive evolution that lie in the ID. We used the binomial distribution to model X . To test the statistical significance of the overlap between the model-predicted sites and the ID, we tested the null hypothesis that the probability of a predicted site lying in the ID (which we refer to as P) is half, versus the alternative hypothesis that P is greater than half, that is:

$$H_0: P = 0.5 \text{ vs. } H_1: P > 0.5.$$

If the null hypothesis is accepted, the observed proportions support the conclusion that the model-predicted sites are equally likely to lie inside or outside the ID, implying a weak association between the predicted sites and the ID. In contrast, rejection of the null hypothesis supports the conclusion that there is an association between the predicted sites and the ID, i.e., the observed overlap between sites under adaptive evolution and the ID is greater than that expected from chance alone (where the adaptive evolution sites are equally likely to occur inside or outside the ID). If n is the number of predicted sites in a given analysis then under the H_0 the probability that X takes the value r [denoted to by $P(X=r)$] is given by the binomial distribution as follows:

$$P(X=r) = \binom{n}{r} \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{n-r} = \binom{n}{r} \left(\frac{1}{2}\right)^n \\ = \frac{n!}{(n-r)!(r)!} \left(\frac{1}{2}\right)^n.$$

We performed a one-tailed test with a significance level of 10%. This involved finding an integer (a) such that $P(X \geq a) < 0.1$. For a given study, the value of a defines the rejection/acceptance criteria as follows: if $X < a$, then the integer a does not lie in the critical region and the null hypothesis is accepted, otherwise it is rejected and the alternative hypothesis accepted.

2 Results

2.1 Maximum-likelihood analysis of orthologous and paralogous FMO sequences

The positively selected sites of the orthologous and paralogous sequences obtained from the maximum-likelihood inferences are shown in Tables 1, 2 and S1, respectively. We identified positively selected sites in the FMOs 3, 4, and 5 orthologs and in the fugu FMO paralogs. Based on the evolutionary fingerprint analysis, both FMOs 4 and 5 have 5 evolutionary rate classes, fugu FMOs have 4 rate classes, and FMO 3 has 3 rate classes (Figure 2). The presence of one rate class with $\omega (\beta/\alpha) > 1$ in these 4 groups suggests the occurrence of positive selection. The SLAC, FEL, REL, M8 and MEC models suggested that one or several FMO 1 sites may have been subject to positive selection. However, the evolutionary fingerprint analysis did not yield evidence of

we found no evidence for positive selection at sites in the sea urchin, medaka, tetraodon, and other non-mammalian data sets, except *fugu* FMOs (Table S1).

The pattern of amino acid change based on ML ancestral sequence reconstruction provides further evidence that FMOs 3–5 evolved under positive selection. Twenty seven amino acid sites in FMO 3, 34 sites in FMO 4, and 32 sites in FMO 5 changed independently to the same amino acid in 2 or more mammalian species (Table S3). For example, site 360 in FMO 3 changed from the polar residue glutamine (Q) to the nonpolar leucine (L) in tarsier and the lesser hedgehog tenrec (Figure 3(a)). In FMO 4, site 235 changed from the nonpolar valine (V) to the negatively charged aspartate (D)

Figure 2 Phylogenetic relationships among FMO proteins and the evolutionary fingerprint of each FMO lineage. The reconstruction was carried out using the JTT + G model. The tree/branch style is shown as a circle. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances are in the units of the number of amino acid substitutions per site. The concept of an evolutionary fingerprint is formulated based on the probability distribution of site-to-site of synonymous (α) and nonsynonymous (β) substitution rates in an alignment [26]. Methods that exploit the $\omega = \beta/\alpha$ ratio as an indicator of positive or negative selection have become a mainstay of modern evolutionary analyses. An accurate estimate of the strength and extent of selective forces acting on a gene, represented by a distribution of α and β , is undoubtedly more informative than a simple binary test for any kind of positive selection. The panel beside each FMO lineage shows the approximate sampling distributions, which are drawn on the log scale that is consistent between all 10 data sets, with α on the x axis and β on the y axis, and the diagonal line corresponding to values $\alpha = \beta$. A proxy for neutral evolution. The color intensity reflects the density assigned to a particular square of the rate distribution. The ellipses are centered on the approximate sampling means of the corresponding rate estimates, and each axis is drawn as $1.96\times$ the sampling standard deviation for α (horizontal) or β (vertical) rates.

in bushbabies and dolphins (Figure 3(b)). This change was classified as radical based on Grantham's distance, which takes into account amino acid size, hydrophobicity, charge, and polarity. This classification was supported by the universal evolutionary index (EI, Table S3). Changes from a noncharged to a charged residue were found in FMOs 4 and 5. Changes at 2 sites in FMO 3, 6 sites in FMO 4, and 2 sites in FMO 5 were not conservative or moderately conservative, as defined by changes in charge or by Grantham's distance. There was no significant difference in the distribution of the four types of amino acid change among FMOs 3–5 (chi square test, $\chi^2=3.86$, $P=0.695$). Such nonconservative changes occur much less frequently than is expected under neutrality [35]. Thus, the nonconservative changes observed are more likely to have consequences for enzyme structure and/or function. Parallel evolution at the amino

acid sequence level can be interpreted as evidence of adaptive evolution [27]. Thus, sites that have changed in parallel are likely targets of selection, in addition to those sites identified using the ML approach. Our results also suggest that the number of available pathways of adaptive evolution may be constrained.

2.3 Assignment of predicted sites under adaptive evolution to the map of the insertion domain and active sites

The predicted sites under adaptive evolution (listed in Tables 2 and S1) were projected onto the map of the ID and AS (i.e., onto the structure of yeast and bacterial FMOs) [31] using the profile alignment strategy. The projections shown in Figure 1 were carried out for FMOs 3–5 and fugu FMO where adaptive evolution was observed (Figure 2).

Figure 3 Mapped codon substitutions for site 360 of FMO 3 (a) and site 235 of FMO 4 (b). Substitutions by branch are shown. The scale (substitutions per site) for each gene is indicated in each tree. Tree branches are color coded by the type of substitutions as follows: both synonymous and non-synonymous (yellow), only synonymous (blue), only non-synonymous (red). (a) Site 360 in FMO 3 changed from the polar residue glutamine (Q) to the nonpolar leucine (L) in tarsier and lesser hedgehog tenrec; (b) site 235 in FMO 4 changed from the nonpolar valine (V) to the negatively charged aspartate (D) in bushbabies and dolphins.

2.4 Statistical significance of the overlap between sites under adaptive evolution and the insertion domain of the FMOs

We used the binomial distribution to model the total number of sites predicted to be under adaptive evolution that lie either inside (Figure 1) or outside the ID. The statistical analysis is summarized in Table S2. A significant number (14 of 16) of positively selected sites in FMO 5 were located in the ID. Similarly, a significant number (4 of 4) positively selected sites in the fugu FMOs detected by MEC model were located in the ID. These sites may have contributed to the fugu's adaptation to specific aquatic environments. None of the positively selected sites in FMOs 3 and 4 were located in ASs 1–6. In addition, only one positively selected site (465L) in FMO 5 was in AS-6 (Figure S1). In contrast, Zawaira et al. [4] found a significant association between sites under adaptive evolution and Gotoh's substrate recognition sites (SRSs) in rat and rabbit CYP2C, human CYP3A, and rat CYP2D. Furthermore, 2 of 3 positively selected sites in the primate CYP3A were in SRS-6 [3]. The method for assessing the statistical significance of the overlap between the sites under adaptive evolution and the ID assumes that all positions in the non-ID region are functionally relevant, either for substrate specificity, binding NADP⁺, or enzyme kinetics. This assumption should be supported or modified based on experimental evidence. Our map of the ID and ASs is complete given current data, but should be updated when information regarding substrate-binding sites encoded in FMO-substrate complexes and mutagenesis become available.

3 Discussion and conclusions

FMOs play an important role in mediating interactions between organisms and their chemical environment due to their involvement in the metabolism of xenobiotics and endobiotics. Thus, they are likely targets for natural selection. Previously, we suggested that FMOs are subject to strong purifying selection [7]. Allerston et al. [6] provided population genetic evidence that human FMO 3 has been subject to balancing selection, but not positive selection. To further our understanding of the evolutionary history of the *FMO* loci, we analyzed mammalian and non-mammalian DNA and protein sequence variation within a phylogenetic framework. For the first time, we unambiguously identified a number of positively selected sites in FMOs 3–5 and showed that the loci have been subject to adaptive evolution.

FMOs belong to the flavoenzyme class of single component flavoprotein monooxygenases. They use equivalents from NADPH to reduce the FAD cofactor which, in turn, becomes capable of reacting with molecular oxygen to yield the C4a-hydroperoxy FAD intermediate [31]. A defining characteristic of this reaction is that the presence of NADP⁺ is essential for intermediate stabilization. Furthermore, NADP⁺ remains consistently bound to the insertion domain of the enzyme throughout the catalytic cycle, being the last product to be released. Interestingly, we observed a significant overlap between the FMO 5 sites under adaptive evolution (PAML-predicted) and the insertion domain. The most likely role of adaptive evolution is to repair the mutations that permitted optimal NADP⁺ binding and improved

catalytic efficiency toward the expanding substrate repertoire during long-term evolution of FMO 5. Compared to FMOs 1–4, FMO 5 is ancient in origin [7]. FMO 5 is the most prominent form of FMO in the fetal liver. *FMO 5* and *FMO 3* transcripts are both prevalent in human adult liver. FMO 5 is also the most abundant FMO transcript present in the human small intestine and may contribute to intestinal first-pass metabolism [36]. FMO 5 does not oxygenate typical FMO substrates (i.e., methimazole, ranitidine, or cimetidine) [37,38], but has been reported to S-oxygenate thioethers with a proximal carboxylic acid, a somewhat unique FMO substrate activity [39]. The dietary anticarcinogen quercetin enhanced the action of genes involved in phases I and II metabolism, including *FMO 5* [40]. E7016, an inhibitor of poly(ADP-ribose) polymerase, undergoes FMO 5-mediated Baeyer-Villiger oxidation in liver microsomes [41]. The specificity of FMO 5 toward catalyzing this Baeyer-Villiger oxidation was confirmed and may be mediated by the amino acid residues in the ID that are under adaptive evolution as these positively selected sites are adjacent to the well-known signature motif FXGXXXHX-XX(Y/F) which is very similar to that of the active site of the BVMOs. As more specific substrates are discovered, FMO 5 may show selective functional activity. Thus, we speculate that FMO 5 differentiated at an early stage, following the split from the common ancestor of FMOs 1–4, to fulfill a critical role in the liver and intestine. The evolutionary fate of FMO 5 is consistent with the neofunctionalization model that predicts that one of the duplicated genes acquires a new function because of positive Darwinian selection. Whether the positively selected sites found in this study are related to the distinct monooxygenation profile of FMO 5 for a large repertoire of xenobiotic agents requires further experimental study.

Seven out of 13 positively selected FMO 4 sites are located in the ID, 5 in the C-terminal region and only 1 in the highly conserved N-terminal region. This is reasonable biologically because genetic variability at the *FMO 4* locus has presumably evolved to detoxify novel types of xenobiotics and drugs. FMO 4 is the least studied FMO isoform. The loci for *FMOs 1–4* diverged from the ancestral *FMO* locus a long time ago and now have different functions and metabolize different substrates [7]. Whereas other FMOs displayed a significant, dominant tissue-specific mRNA profile (i.e., FMO 1 in kidney, FMO 2 in lung, FMOs 3 and 5 in adult liver), FMO 4 mRNA was observed more broadly at relatively comparable levels in the liver, kidney, lung, and small intestine [36]. Real time quantitative RT-PCR confirmed that 20 $\mu\text{mol/L}$ p-NO-ASA (NO-donating aspirin) upregulated FMO 4 by 4.5 ± 1.67 -fold [42]. Taken together, these observations illustrate FMO subfunctionalization which, in addition to neofunctionalization and nonfunctionalization [34] following *FMO* gene duplication, increases the complexity of the evolutionary scenario.

Although neofunctionalization may not be important for

the evolution of FMOs 1 (the predominant FMO in the fetal liver) and 2 (tend to nonfunctionalization), it is reasonable to postulate that, given a constantly changing environment, some positively selected sites were involved in the neofunctionalization of FMOs 3–5. Following the relatively recent split of FMO 4 and the common ancestor of FMOs 1 and 3 (Figure 2), these detoxification enzymes are currently in the process of specialization. Due to limited knowledge regarding substrate specificity and the lack of information on the 3-D structures of mammalian FMOs, we are unable to adequately explain why positively selected sites were observed in FMOs 3–5, but not in FMOs 1 and 2.

Mammalian FMO 3 oxygenates a variety of nucleophilic primary, secondary, and tertiary amines, as well as sulfur and other heteroatom-containing chemicals and drugs [1]. Because FMO 3 is not readily induced or inhibited, variation in the functional activity is derived from genetic variability arising largely from common SNPs (Figures 1 and S1). Conversely, individuals with rare FMO 3 mutations that cause defective TMA N-oxygenation suffer from trimethylaminuria (TMAU). Allerston et al. [6] found no evidence for positive selection when only considering the intraspecific FMO 3 sequence variation. Conversely, we found two positively selected sites using the M8 model by comparing interspecific sequence variation, which suggests that adaptive evolution has had a role in the long-term, but not short-term, evolution of FMO 3. One of positively selected sites is 360Leu, a polymorphic site. A small number of individuals (i.e., 1%) possessed the Pro360-FMO 3 variant, which is catalytically more efficient than the wild-type FMO 3 [43]. P360-FMO 3 oxygenated mercaptoimidazole, TMA, and 10-(*N,N*-dimethylaminopentyl)-2-(trifluoromethyl)phenothiazine approximately 3-, 5-, and 2-fold more efficiently, respectively, than wild-type FMO 3. For these substrates, the K_m values for P360-FMO 3 and the wild-type enzyme were similar, but the $V_{\max}$ values were greater. It is possible that being adjacent to the insertion domain, P360-FMO3 is able to facilitate desorption of NADP^+ or dehydration of FAD pseudobase water and, thus, speed up the FMO reaction [44]. Another possibility is that P360-FMO 3 modifies the structure of FMO 3. It is known that substitution of Pro into an enzyme can change protein structure. Analogously, the change in amino acid from Q to L, or other cases on site 360, may have adjusted the spatial relationship and/or modified the affinity between the insertion domain and NADP^+ , thus optimizing FMO 3 reactions with newly encountered substrates in the environment.

In conclusion, we present the first evidence for the occurrence of adaptive evolution in mammalian FMOs 3, 4, 5, and fugu FMOs. The adaptive evolution signal was not detected in mammalian FMOs 1, 2, frog FMOs, other fish FMOs, and invertebrate FMOs. Our results show distinct patterns of evolution for mammalian FMOs 1 and 3, both of which have a recent origin. Furthermore, the sites under adaptive evolution are significantly associated with the in-

sertion domain in mammalian FMO 5. Whether the identified positive selection and the parallel amino acid substitution alter the product profile in the respective species deserves further study. The finding of positive selection during the evolution of phase I detoxification enzymes such as FMOs 3–5 and fugu FMO is counterintuitive in the light of the supposed limited room for change in these molecules. Our results support expectations of both high selection pressure acting on the various species within their unique habitats and significant changes in intensity and direction (kinds of xenobiotics and drugs) resulting from changes in microhabitat and food. Future studies of enzyme structure, substrate specificity, and catalytic mechanism would benefit from site-directed mutagenesis based on our current results.

We gratefully acknowledge Dr. Jared V. Goldstone (Department of Biology, Woods Hole Oceanographic Institution) for his help in predicting sea urchin FMO sequences. We thank the two anonymous reviewers for their insightful comments. The work was supported by the National Science and Technology Major Project of the Ministry of Science and Technology of China (2008ZX10005-004), Educational Commission of Liaoning Province of China (2009A120) and China Postdoctoral Science Foundation (20080440019 and 200902069).

- 1 Krueger S K, Williams D E. Mammalian flavin-containing monooxygenases: Structure/function, genetic polymorphisms and role in drug metabolism. *Pharmacol Ther*, 2005, 106: 357–387
- 2 Phillips I R, Shephard E A. Flavin-containing monooxygenases: Mutations, disease and drug response. *Trends Pharmacol Sci*, 2008, 29: 294–301
- 3 Qiu H, Taudien S, Herlyn H, et al. CYP3 phylogenomics: Evidence for positive selection of CYP3A4 and CYP3A7. *Pharmacogenet Genomics*, 2008, 18: 53–66
- 4 Zawaira A, Matimba A, Masimirembwa C. Prediction of sites under adaptive evolution in cytochrome P450 sequences and their relationship to substrate recognition sites. *Pharmacogenet Genomics*, 2008, 18: 467–476
- 5 Chen X, Wang H, Zhou G, et al. Molecular population genetics of human CYP3A locus: Signatures of positive selection and implications for evolutionary environmental medicine. *Environ Health Perspect*, 2009, 117: 1541–1548
- 6 Allerston C K, Shimizu M, Fujieda M, et al. Molecular evolution and balancing selection in the flavin-containing monooxygenase 3 gene (FMO 3). *Pharmacogenet Genomics*, 2007, 17: 827–839
- 7 Hao D C, Chen S L, Xiao P G, et al. Molecular phylogeny, long-term evolution, and functional divergence of flavin-containing monooxygenases. *Genetica*, 2009, 137: 173–187
- 8 Sea Urchin Genome Sequencing Consortium. The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science*, 2006, 314: 941–952
- 9 Eddy S R. Profile hidden Markov models. *Bioinformatics*, 1998, 14: 755–763
- 10 Larkin M A, Blackshields G, Brown N P, et al. ClustalW and ClustalX version 2. *Bioinformatics*, 2007, 23: 2947–2948
- 11 Birney E, Clamp M, Durbin R. GeneWise and Genomewise. *Genome Res*, 2004, 14: 988–995
- 12 Burge C B, Karlin S. Finding the genes in genomic DNA. *Curr Opin Struct Biol*, 1998, 8: 346–354
- 13 Wernersson R, Pedersen A G. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res*, 2003, 31: 3537–3539
- 14 Saitou N, Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 1987, 4: 406–425
- 15 Tamura K, Dudley J, Nei M. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol*, 2007, 24: 1596–1599
- 16 Abascal F, Zardoya R, Posada D. ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics*, 2005, 21: 2104–2105
- 17 Ronquist F, Huelsenbeck J P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 2003, 19: 1572–1574
- 18 Zwickl D J. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Doctor dissertation. Austin: The University of Texas at Austin, 2006
- 19 Eswaramoorthy S, Bonanno J B, Burley S K, et al. Mechanism of action of a flavin-containing monooxygenase. *Proc Natl Acad Sci USA*, 2006, 103: 9832–9837
- 20 Zhang M, Robertus J D. Molecular cloning and characterization of a full-length flavin-dependent monooxygenase from yeast. *Arch Biochem Biophys*, 2002, 403: 277–283
- 21 Yang Z, Nielsen R, Goldman N. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 2000, 155: 431–449
- 22 Yang Z. PAML 4: A program package for phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, 2007, 24: 1586–1591
- 23 Doron-Faigenboim A, Pupko T A. Combined empirical and mechanistic codon model. *Mol Biol Evol*, 2007, 24: 388–397
- 24 Kosakovsky P S L, Frost S D. Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol*, 2005, 22: 1208–1222
- 25 Kosakovsky P S L, Frost S D. Datamonkey: Rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics*, 2005, 21: 2531–2533
- 26 Pond S L, Scheffler K, Gravenor M B, et al. Evolutionary fingerprinting of genes. *Mol Biol Evol*, 2010, 27: 520–536
- 27 Zhang J. Parallel functional changes in the digestive RNases of ruminants and colobines by divergent amino acid substitutions. *Mol Biol Evol*, 2003, 20: 1310–1317
- 28 Yang Z, Kumar S, Nei M. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, 1995, 141: 1641–1650
- 29 Grantham R. Amino-acid difference formula to help explain protein evolution. *Science*, 1974, 185: 862–864
- 30 Tang H, Wyckoff G J, Lu J, et al. A universal evolutionary index for amino acid changes. *Mol Biol Evol*, 2004, 21: 1548–1556
- 31 Alfieri A, Malito E, Orru R, et al. Revealing the moonlighting role of NADP in the structure of a flavin-containing monooxygenase. *Proc Natl Acad Sci USA*, 2008, 105: 6572–6577
- 32 Veeramah K R, Thomas M G, Weale M E. The potentially deleterious functional variant flavin-containing monooxygenase 2*1 is at high frequency throughout sub-Saharan Africa. *Pharmacogenet Genomics*, 2008, 18: 877–886
- 33 Nei M, Rooney A P. Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet*, 2005, 39: 121–152
- 34 Hernandez D, Janmohamed A, Chandan P. Organization and evolution of the flavin-containing monooxygenase genes of human and mouse: Identification of novel gene and pseudogene clusters. *Pharmacogenetics*, 2004, 14: 117–130
- 35 Li W H, Wu C I, Luo C C. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J Mol Evol*, 1984, 21: 58–71
- 36 Zhang J, Cashman J R. Quantitative analysis of FMO gene mRNA levels in human tissues. *Drug Metab Dispos*, 2006, 34: 19–26
- 37 Overby L H, Buckpitt A R, Lawton M P. Characterization of flavin-containing monooxygenase 5 (FMO 5) cloned from human and guinea pig: Evidence that the unique catalytic properties of FMO 5 are not confined to the rabbit ortholog. *Arch Biochem Biophys*, 1995, 317: 275–284
- 38 Cherrington N J, Cao Y, Cherrington J W. Physiological factors affecting protein expression of flavin-containing monooxygenases 1, 3 and 5. *Xenobiotica*, 1998, 28: 673–682
- 39 Ohmi N, Yoshida H, Endo H. S-oxidation of S-methylesonarimod by flavin-containing monooxygenases in human liver microsomes. *Xenobiotica*, 2003, 33: 1221–1231

- 40 Dihal A A, van der Woude H, Hendriksen P J, et al. Transcriptome and proteome profiling of colon mucosa from quercetin fed F344 rats point to tumor preventive mechanisms, increased mitochondrial fatty acid degradation and decreased glycolysis. *Proteomics*, 2008, 8: 45–61
- 41 Lai W G, Farah N, Moniz G A, et al. A Baeyer-Villiger oxidation specifically catalyzed by human flavin-containing monooxygenase 5 (FMO 5). *Drug Metab Dispos*, 2011, 39: 61–70
- 42 Nath N, Chattopadhyay M, Kodala R, et al. Modulation of stress genes expression profile by nitric oxide-releasing aspirin in Jurkat T leukemia cells. *Biochem Pharmacol*, 2010, 79: 1759–1771
- 43 Lattard V, Zhang J, Tran Q, et al. Two new polymorphisms of the *FMO 3* gene in Caucasian and African-American populations: Comparative genetic and functional studies. *Drug Metab Dispos*, 2003, 31: 854–860
- 44 Orbás T, Zhang J, Cerny M A, et al. Investigation of structure and function of a catalytically efficient variant of the human flavin-containing monooxygenase form 3. *Drug Metab Dispos*, 2006, 34: 1995–2002

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

Supporting Information

Table S1 Likelihood ratio statistics and AICc scores for tests of positive selection of paralogous *FMO* sequences

Table S2 Statistical significance of the overlap between sites under adaptive evolution and the insertion domain of FMOs

Table S3 Parallel amino acid substitutions in mammalian FMOs

Dataset S1 Orthologous and paralogous sequences used in this study

The supporting information is available online at csb.scichina.com and www.springerlink.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.