Engineering Biocatalysts for the C-H Activation of Fatty Acids using Ancestral Sequence Reconstruction

Bethan S. Jones[1], Connie M. Ross[2], Gabriel Foley[2], Nico Kress[1], Raine E. S. Thomson[2], Yosephine Gumulya[2], Martin A. Hayes[3], Elizabeth M. J. Gillam[2]*, and Sabine L. Flitsch[1]*

[1] Department of Chemistry, The University of Manchester, Manchester Institute of Biotechnology, 131 Princess Street, Manchester M1 7DN (UK).

[2] School of Chemistry and Molecular Biosciences, The University of Queensland, St. Lucia, Brisbane, 4072 (Australia).

[3] Compound Synthesis and Management, Discovery Sciences, R&D, AstraZeneca, Gothenburg (SE).

* Corresponding authors: e.gillam@uq.edu.au, sabine.flitsch@manchester.ac.uk

Abstract

Selective, one-step C-H activation of fatty acids from biomass is an attractive concept in sustainable chemistry. Biocatalysis has shown promise for generating high-value hydroxy acids but to date enzyme discovery has relied on laborious screening and produced limited hits, which predominantly oxidise the sub-terminal positions of fatty acids. Here we show that ancestral sequence reconstruction (ASR) is an effective tool to explore the sequence-activity landscape of a family of multi-domain, self-sufficient P450 monooxygenases. We resurrected eleven catalytically active CYP116B ancestors, each with a unique regioselectivity fingerprint that varied from sub-terminal in the older ancestors to mid-chain in the lineage leading to the extant, P450-TT. In lineages leading to extant enzymes in thermophiles, thermostability increased from ancestral to extant forms, as expected if thermophily had arisen *de novo*. Our studies show that ASR can be applied to multi-domain enzymes to develop active, self-sufficient monooxygenases as regioselective biocatalysts for fatty acid hydroxylation.

Selective hydroxylation of fatty acids offers many opportunities for generating a wide range of pharmaceuticals, flavours, and fragrances from the limited starting materials that are available from biomass[1]. For instance, hydroxylation at each of the 19 C-H bonds in the highly abundant starting material, decanoic acid, would provide 17 unique products that can only be accessed by kinetic control over reaction conditions[2]. The challenge is to find catalysts that can distinguish chemically similar positions so as to selectively modify each of these C-H bonds. Biocatalytic methods offer the potential for exquisite control over the position of hydroxylation, but despite considerable efforts, only a limited number of effective biocatalysts for selective fatty acid hydroxylation have been identified to date. Most of these biocatalysts are from the superfamily of cytochrome P450 monooxygenases (P450s, CYP family) and have been largely limited to terminal or subterminal hydroxylases (Fig. 1)[3]. Many different P450 families hydroxylate fatty acids[3], *e.g.,* CYP2[4], CYP4[2], CYP52A[5], CYP102[6–8], CYP116[9], CYP147G[10], CYP152A[11], and CYP153A[12]. Despite the potential of these enzymes for use in organic synthesis, difficulties with recombinant, heterologous expression, and low stability have limited their implementation. In addition, the identification and development of new biocatalysts is time-consuming; intensive screening of large numbers of natural and synthetic sequences is typically needed to identify a very small number of viable biocatalysts. Efforts to identify novel P450 biocatalysts from metagenomic databases have yielded very few P450s that are capable of selective mid-chain hydroxylation[2,3,13–15], although much of the metagenomic sequence space remains unstudied, identifying forms that have the necessary specificity and kinetic selectivity is challenging. Directed evolution and rational protein design has improved the activity and selectivity of some terminal fatty acid hydroxylases, *e.g.,* CYP102A1 (P450-BM3)[8,16,17] and altered the regioselectivity of others, *e.g.,* CYP152A (P450-Bsβ)[18,19], but engineering P450s, like metagenome mining, is time-consuming, and both labour- and resource-intensive, due to the need to screen large numbers of potential biocatalysts to identify one with the desired properties.

We recently identified CYP116B46 (P450-TT) from a panel of self-sufficient thermostable wild-type, extant CYP116B enzymes as being capable of selective mid-chain hydroxylation of decanoic acid to (*S*)-5-hydroxydecanoic acid[13]. While related enzymes from the same family (P450-AT, -TB, -AX, and -JT, Fig. 2) were either not regioselective or favoured ω-1 hydroxylation[13]. To further understand what made P450-TT selective for mid-chain hydroxylation, and to develop other P450s with similar capacities for mid-chain hydroxylation, we performed ancestral sequence reconstruction (ASR) on the CYP116B family.

ASR was employed as a tool to explore the sequence space of the CYP116B family, which enabled us to analyse a comparatively small number of potential biocatalysts with a greater chance of successfully identifying catalytically active enzymes with the desired properties of thermostability, self-sufficiency, and selective fatty-acid hydroxylation. ASR allows the generation of new proteins based upon a sequence alignment and inferred evolutionary tree that link a set of extant sequences; the most probable common ancestor and other evolutionary intermediates are inferred from the initial input[20]. The method has previously been used to identify and engineer enzymes with improved properties: greater thermal stability, enhanced expression, broader pH, and solvent tolerances[21–33]. This technique can identify novel sequences leading to functional diversity that otherwise may not have been investigated *via* conventional enzyme engineering and, unlike mutagenesis, most ancestors are functional[34,35]. ASR has also been used to improve the thermostability and solvent tolerance of animal P450s[36–39], which can improve the heterologous expression and durability of biocatalysts[40]; typically, inferred ancestors have been found to be more thermostable than their extant counterparts. However, the evolution of thermostable enzymes within a largely mesophilic family had not been previously explored. As the extant CYP116B family has both mesophilic and thermophilic branches, and several previously characterised CYP116B sequences (*e.g.,* P450-TT) originate from thermophilic (rather than mesophilic) bacteria we were interested in what evolutionary trends would emerge in the ASR of the CYP116B enzymes.

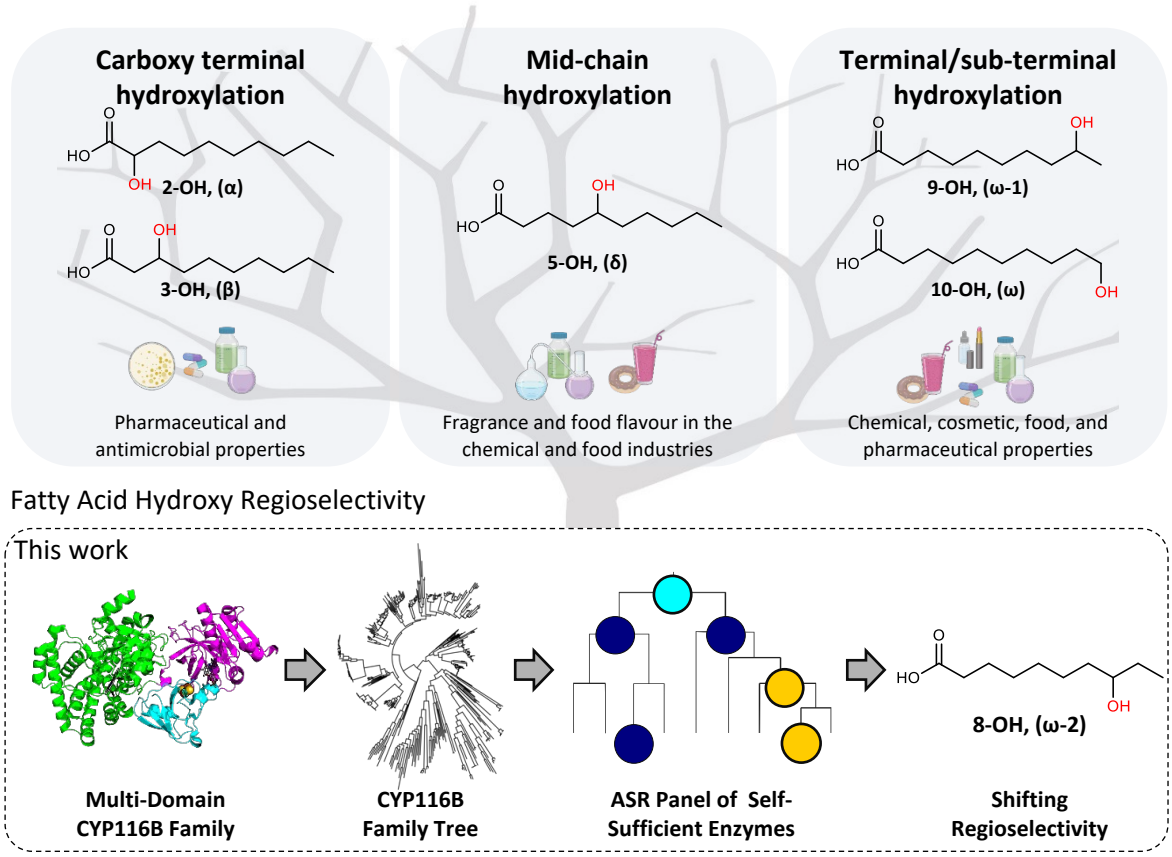

Fig. 1. Hydroxy acid regioselectivity classification: carboxy terminal (α, β), mid-chain (δ/ω-5), and terminal/sub-terminal (ω, ω-1) positions, and their properties. The previously identified self-sufficient extant CYP116B enzymes displayed activity towards fatty acids with a range of mid-chain and sub-terminal regioselectivities[13]. P450-TT uniquely favoured mid-chain hydroxylation of decanoic acid. Therefore, the CYP116B family was further explored through ancestral sequence reconstruction, with particular focus on the lineage leading to P450-TT. Extant sequences with a sequence similarity >45% were gathered to generate a phylogenetic tree for the CYP116B family using IQ-TREE 2 (v2.1.2). Ancestral sequences from within the tree were then resurrected to create a panel of ancestral self-sufficient biocatalysts, where the nodes shown depict a simplified representation of the differing regioselective of decanoic acid hydroxylation from mid-chain (ω-5; orange) to different sub-terminal positions (ω-1, dark blue; ω-2, cyan). A shift in regioselectivity was identified in the oldest ancestor (ω-2), differing from both the mid-chain selectivity of P450-TT, and the predominant sub-terminal ω-1 regioselectivity observed for other members of the CYP116B family. This highlights the usefulness of ancestral sequence reconstruction to enable the exploration of the sequence space within the CYP116B family and identify different clades with differing regioselectivity.

Although ASR has been successfully used for single domain proteins, less work has been carried out on multi-domain enzymes. CYP116B enzymes are multi-domain, Class VII P450s: electrons from NAD(P)H are transferred to the P450 haem *via* a fused reductase domain and an iron-sulfur ferredoxin. These are 'self-sufficient' enzymes, where the expressed proteins can be directly assayed for bioactivity without the need for additional redox partners[19]. The proximity of the reductase and haem domains facilitates electron shuttling and enables a more effective transfer, enhancing their potential as biocatalysts[41]. Therefore, we wanted to investigate whether ASR was also an effective engineering approach when applied to the three-domain protein of the CYP116B family, and whether the individual domains would retain folding and activity when inferred together.

Using ASR to identify important mutational paths that lead to changes in enzymatic activity can provide insight into the mechanisms that control substrate specificity and activity[31–33,42–51], as demonstrated in the recent work on stereoselective C-H activation by flavin-dependent monooxygenases[51]. We were interested in how the regioselectivity towards fatty acid hydroxylation arose within the ancestral CYP116B family, especially in the pathway toward the unique mid-chain selectivity of P450-TT. In particular, we wondered whether some of these ancestors would allow us to access new regioselectivities, such as ω-2, which is not commonly observed in other biocatalysts.

We found ASR to be an effective method to generate eleven, self-sufficient, ancestral CYP116B P450 enzymes. A range of stabilities was observed, with a trend towards increasing thermostability along the branches leading to the more thermostable extant forms, suggesting that thermophilicity arose recently in the P450-TT and other extant thermophilic lineages, rather than being maintained from a highly stable ancestor. The inferred CYP116B ancestors were all catalytically active and showed unique regioselectivity profiles towards the hydroxylation of decanoic acid, including ancestors with novel selectivity for the ω-2 position. Finally, by examining the existing CYP116B structural information[52–55] and comparing the ancestral sequences, we were able to identify two active site residues that were partially responsible for changes to the regioselectivity of the CYP116B ancestors *vs*. the wild-type P450-TT from terminal to mid-chain hydroxylation.

Results

Inference of the phylogenetic and ancestral intermediates. The CYP116B family is highly diverse and CYP116B enzymes vary substantially in their thermostability, fatty acid hydroxylation regioselectivity[13,41], and sequence identity (Supplementary Tables 1 and 2). To generate an enzyme panel with diversified catalytic range, CYP116B sequences were collected by database searches and curated. We obtained 467 CYP116B sequences that included all three domains (P450/haem, iron-sulfur, and reductase). The sequences were mostly derived from two main phyla, Actinobacteria and Proteobacteria (Supplementary Fig. 1)[56,57] and covered twenty different bacterial orders, encompassing sequences that were >45% identical to P450-TT. In contrast to the previously characterised thermophilic panel of extant CYP116B sequences, most of the sequences were isolated from mesophiles (Supplementary Fig. 2)[41], although specific lineages appeared to be thermophilic, such as P450-TT and its closest homologues from *Caldimonas taiwanensis* and *Schlegelella thermodepolymerans,* which have been isolated from hot springs and activated sludge, respectively[58,59]. To explore the sampled sequence space, we inferred the ancestors from each major phylogenetic group (N374, N1, N213, N104, N118, N23, N70, and N3), as well as specific ancestors of the P450-TT lineage (N106 and N115) and the node (N0) connecting the major clade (Proteobacteria) with the outgroup (Actinobacteria). A total of eleven full-length multi-domain CYP116B ancestors (Fig. 2, the highlighted nodes within the tree) were inferred using the inference program Graphical Representation of Ancestral Sequence Predictions (GRASP), *via* a joint reconstruction maximum likelihood approach[20]. These self-sufficient ancestral enzymes were selected to explore the thermostability and fatty acid hydroxylation regioselectivity trends in the P450-TT lineage and to investigate nodes on divergent CYP116B branches corresponding to mesophilic wild-type sequences. All ancestors were expressed as holoenzymes with characteristic P450 spectra, with two forms reaching micromolar concentrations in culture (Supplementary Table 3). N106 and N115 showed higher expression at 25 °C than all other CYP116B enzymes including their direct descendant, P450-TT, which was the most highly expressed extant form.

Thermostability profiles. Enhanced thermostability has been repeatedly observed in reconstructed ancestral proteins, including other P450s, compared to the extant forms from which they are inferred, where sequences have been derived from mesophilic organisms. However, to the best of our knowledge, no reconstructions have been performed to trace the evolution of stability in protein families from thermophiles. As the CYP116B family contains proteins from both mesophiles and thermophiles (Supplementary Fig. 2), we were interested in whether thermostability in CYP116B family enzymes is a recent development and a specific adaptation in CYP116B sequences from thermophilic organisms. We assessed the thermostability of the CYP116B ancestors by incubating the proteins at a range of temperatures (25-80 °C) and assessing the loss of the P450 holoenzyme by changes to the characteristic P450 spectrum. Two incubation times were used (15 and 60 minutes), which allowed us to determine both the $^{15}T_{50}$ and $^{60}T_{50}$ stabilities for each ancestor (Table 1; Supplementary Fig. 3 and 4). A wide range of thermostabilities ($^{15}T_{50}$ 38-54 °C and $^{60}T_{50}$ <30-55 °C) was observed among different ancestors across the CYP116B tree. As expected, the $^{60}T_{50}$ values were generally lower than the $^{15}T_{50}$ values, indicating a lower tolerance for prolonged heating, particularly in the ancestors and wild-type enzymes with lower stability *e.g.*, $^{15}T_{50}$ <45 °C (Fig. 2, Table 1).

In contrast to most other reconstructions, particularly of P450 enzymes, we did not observe a general trend towards greater thermostability in the deeper nodes. Instead, specific lineages appeared to have lost and then sometimes developed thermostability *de novo* during CYP116B evolution (Figure 2, Supplementary Fig. 5). P450-AX, -TT, and -TB were the most thermostable extant forms, in agreement with previous studies[41] with $^{15}T_{50}$ values of 54, 53, and 52 °C, respectively, over 10°C higher than those for P450-JT and -AT, forms that were markedly less thermostable and derived from organisms with generally lower optimal growth temperatures[41]. Amongst the ancestors, N3, which theoretically gave rise to mesophilic and uncharacterised extant CYP116B sequences, was the most stable with a $^{15}T_{50}$ of 51 °C (Table 1, Supplementary Fig. 5).

A closer look at the sequence comparisons in relation to three dimensional structures showed that variation was concentrated in the loops of the ancestors (Supplementary Fig. 6). Greater flexibility in the protein scaffold may explain the reduced thermostability of the ancestors relative to P450-TT[60]. Often proteins from thermophiles are shorter than those from mesophiles, by reduction in the length of loops and linkers; however, no such difference was apparent here between P450-TT and the ancestral CYP116B forms. A general comparison of amino acid composition (Supplementary Table 4) showed a high degree of similarity between ancestors and extant forms, with a slight shift towards more hydrophobic amino acids in the ancestors compared to P450-TT, as noted previously for the CYP3 family reconstruction[36].

The ancestors were all catalytically active, *i.e.,* maintained electron transfer across the inferred domains. In multi-domain enzymes such as the self-sufficient CYP116B forms, loop and linker variation across domains, such as seen here among in the ancestors (Supplementary Fig. 7), can influence the dynamics and interactions between the domains. In particular, efficient electron transfer across the domains to the haem domain is vital for hydroxylation activity. The ability of the ancestors to hydroxylate fatty acids was analysed using purified enzyme (10 and 50 μM; Supplementary Figures 8-9), with decanoic acid (0.5 mM) as the substrate. The biotransformations utilised glucose dehydrogenase for concomitant regeneration of NADPH, the preferred cofactor of the reductase domain of P450-TT[41]. At the lower enzyme concentration (10 μM), N70 displayed 69% conversion of decanoic acid, whereas the conversion with the other ancestral enzymes was low or negligible (0-37%, Supplementary Table 5). Therefore, the ancestral regioselectivity fingerprints (Fig. 2) were examined using a five-fold higher enzyme loading. Under these conditions, all self-sufficient ancestors were catalytically active (Supplementary Tables 5 and 6), implying that electron transfer was maintained in all the ancestors, and attesting to the ability of ASR to facilitate the engineering of active electron transport chains. Interestingly, the highest turnover among ancestral forms at both P450 loadings was achieved with N70, N106, and N115, forms which also showed the best expression.

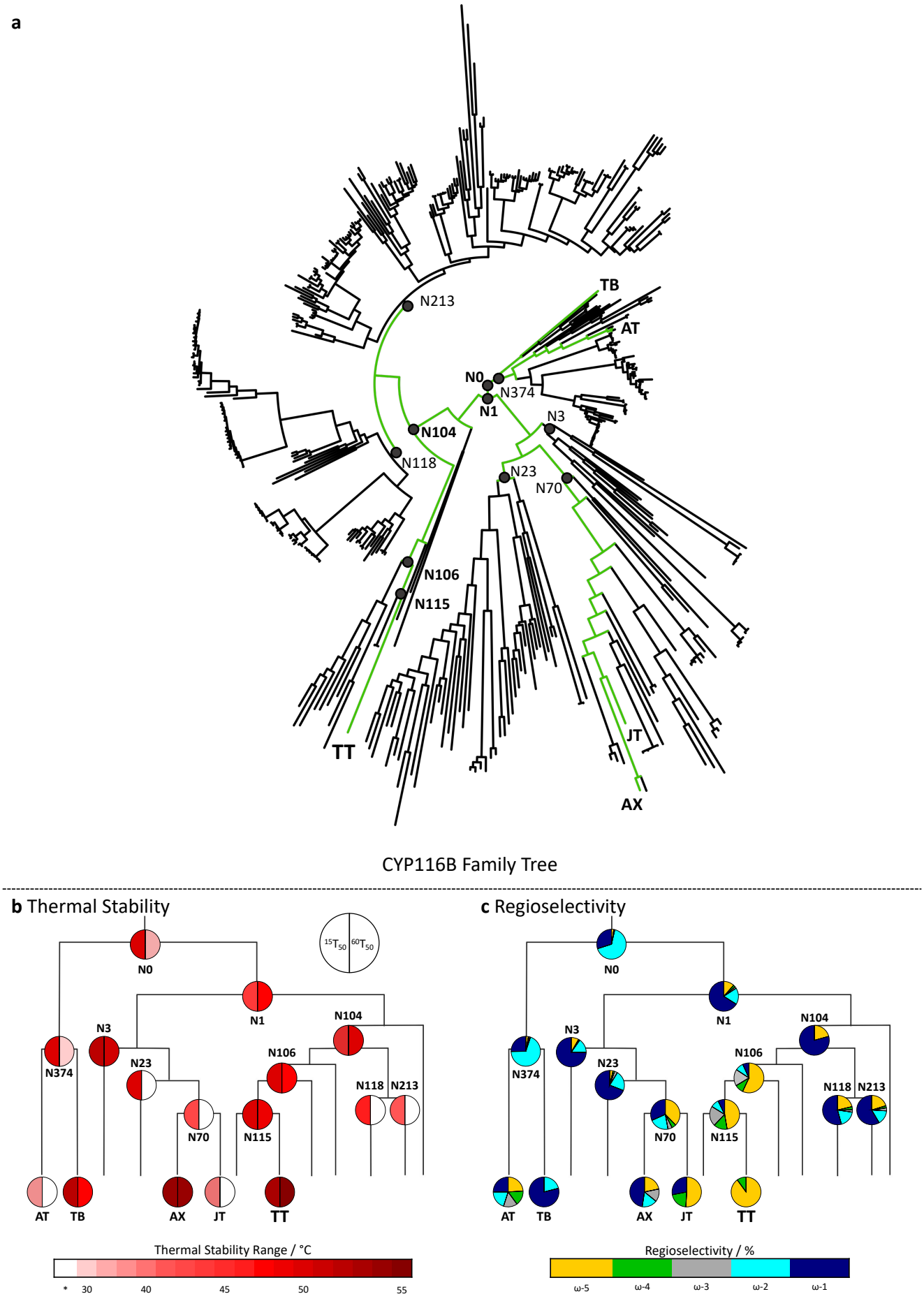

Fig. 2. Phylogeny of the 467 sequences from the CYP116B family used for ASR (a) and thermal stability profiles (b) and shifting regioselectivity pathways (c) observed within the characterised ancestral and extant CYP116B enzymes. The IQ-TREE 2 (v2.1.2) of the CYP116B family (top) depicting the 'evolutionary pathway' (in green) towards the enzymes in the extant wild-type panel and the resurrected nodes within the tree. The simplified tree representations (bottom) highlight the different thermal stability (b) and hydroxylation regioselectivity distribution (c) of the ancestral and extant enzymes at each node. The thermal stability ($^{15}T_{50}$) was determined from the residual P450 peak after incubating the enzyme at 25-80 °C for 15 min. A full sigmoidal denaturation curve was not obtained for 6 of the enzymes at 60 min, so the $^{60}T_{50}$ values are not shown (see Table 1); however, $^{15}T_{50}$ correlated with $^{60}T_{50}$ where both values could be determined (R^2 = 0.874). P450-TT and -AX were the most stable extant forms whereas N3 was the most thermostable ancestral enzyme. The regioselectivity of the ancestral and previously characterised extant P450s are shown in (c). The regioselectivity product distribution of the ancestral enzymes (50 μM) towards the hydroxylation of decanoic acid (0.5 mM) was investigated in biotransformations carried out with a glucose dehydrogenase recycling system, in 0.1 M potassium phosphate buffer, pH 8 and at 30 °C, with agitation at 200 rpm for 20 h. The ancestral enzymes in the pathway to P450-TT showed a change from the mid-chain selectivity in P450-TT towards sub-terminal selectivity in the older ancestors. In addition, there was a change in the ancestral nodes from ω-1 hydroxylation in the majority of the ancestral enzymes to ω-2 hydroxylation in N0 and N374. Data for the extant forms are from Manning *et al.*[13].

Table 1. Thermostability of the CYP116B ancestral enzymes. $^{15}T_{50}$ values for ancestral forms ranged from 40-51 °C compared to the 38-54 °C stability range from the extant CYP116B enzymes from thermophilic organisms. N3 was the most thermostable ancestor. Only five ancestral enzymes maintained their thermal stability after an hour incubation. Data are means ± standard error.

Thermal Stability	Ancestral CYP116B panel											Extant CYP116B panel				
	N0	N1	N3	N23	N70	N104	N106	N115	N118	N213	N374	TT	JT	AX	AT	TB
$^{15}T_{50}$ / °C	48±1	44±4	51±0.4	40±0.6	43±2	45±3	50±0.3	50±0.4	46±4	42±1	50±1	53±1	40±0.3	54±1	38±3	52±1
$^{60}T_{50}$ / °C	36±8	48±2	49±2	n.d.[a]	n.d.[a]	50±2	48±1	49±0.4	n.d.[a]	n.d.[a]	<30[a]	55±1	n.d.[a]	54±3	n.d.[a]	48±2

[a] *An accurate estimate of $^{60}T_{50}$ could not be determined (n.d.) since a full sigmoidal denaturation curve was not obtained.*

Evolutionary shift in regioselectivity. It was previously shown that P450-TT was highly selective for mid-chain hydroxylation of decanoic acid[13]. Across the ancestral panel, hydroxylation of decanoic acid at each position between ω-5 and ω-1 was detected; yet there was no observation of any mid-chain ω-6 hydroxylation or hydroxylation closer to the carboxylate. In N0 and N374, there was an unexpected shift in selectivity to favour the ω-2 position over the ω-1 position. This novel sub-terminal preference for the ω-2 position was not seen in any of the wild-type product distributions. By contrast, the other deep ancestors, N1, N104, N3 and N23, showed a preference for ω-1 hydroxylation (Fig. 2). The deeper ancestral enzymes tended to be more regioselective than the more recent ancestors or the extant enzymes that were examined, with the exception of P450-TT and -TB. Mid-chain hydroxylation arose independently in two different lineages (N70 and N106), from ancestors that were predominantly ω-1 hydroxylases (N1, N3 and N104). In the N70 lineage, a non-selective ancestor (N70: 38% mid-chain and 32% sub-terminal), gave rise to two non-selective wild-type enzymes P450-JT and -AX which displayed slight preferences for the ω-5 (51%) and ω-1 (48%) positions, respectively. In contrast, the P450-TT lineage developed a preference for mid-chain hydroxylation much earlier, in the N106 ancestor, and resulted in a switch to exclusively mid-chain hydroxylation in P450-TT. Indeed, the major product was hydroxylated at the ω-5 position, with 99% conversion, for both N115 and N106, although these ancestors also showed a greater tendency towards hydroxylation at the ω-3 and ω-4 positions, compared to P450-TT. Therefore, this lineage shows a prominent shift in the regioselectivity, compared with the CYP116B bacterial ancestors and suggests an evolutionary pathway towards the unusual regioselectivity of P450-TT. These results highlight the different ways in which selectivity might have evolved and provide biocatalysts for hydroxylation at alternative C-H positions in fatty acids.

Structural insights into the shifting regioselectivity. Having explored the range in regioselectivity in the ancestral CYP116B enzymes, we compared their sequences overall and within their active sites (Fig. 3) to identify determinants of selectivity. Ancestral sequence identities (Supplementary Table 2) ranged from 59 (earlier ancestors) to 77% (more recent ancestors) compared to the P450-TT holoenzyme. Most of the non-conserved residues were found in the regions that are predicted to be surface-exposed (Fig. 3a), with more conservation in the core regions of each domain. The only changes in the active site were two consecutive amino acids (Ala205Thr and Phe206Trp), which increased the active site polarity and were present in all ancestors apart from the two closest to P450-TT (N115 and N106). This suggested a more polar active site may favour sub-terminal hydroxylation in the CYP116B family. Only 17 of the 467 initial extant sequences in the alignment (Fig. 3c) lacked the Thr205-Trp206 motif present in most of the characterised ancestral sequences, instead having either Ala-Phe, Ile-Trp or Gln-Phe residues, suggesting these could be key active site positions governing regioselectivity towards fatty acid hydroxylation.

To better understand the switch in selectivity from the ω-5 to ω-2 or ω-1 products associated with the Ala205Thr and Phe206Trp conserved motif, Ala205 and Phe206 in P450-TT were mutated to obtain the corresponding single and double mutants. The double mutant was more active than P450-TT (Supplementary Figure 10), which may explain the prominence of these residues in other wild-type sequences. Although ω-5 hydroxylation was still the most favoured product, an increase in the proportion of ω-4 and ω-3 hydroxylated alkyl chains was observed (Fig. 3d), suggesting that the increased polarity in the active site has an impact on regioselectivity. Overall, ASR has enabled the identification of active site residues influencing the differing C-H regioselectivity in a panel of ancestral and wild-type CYP116B biocatalysts without the need for a large-scale mutagenic screen.

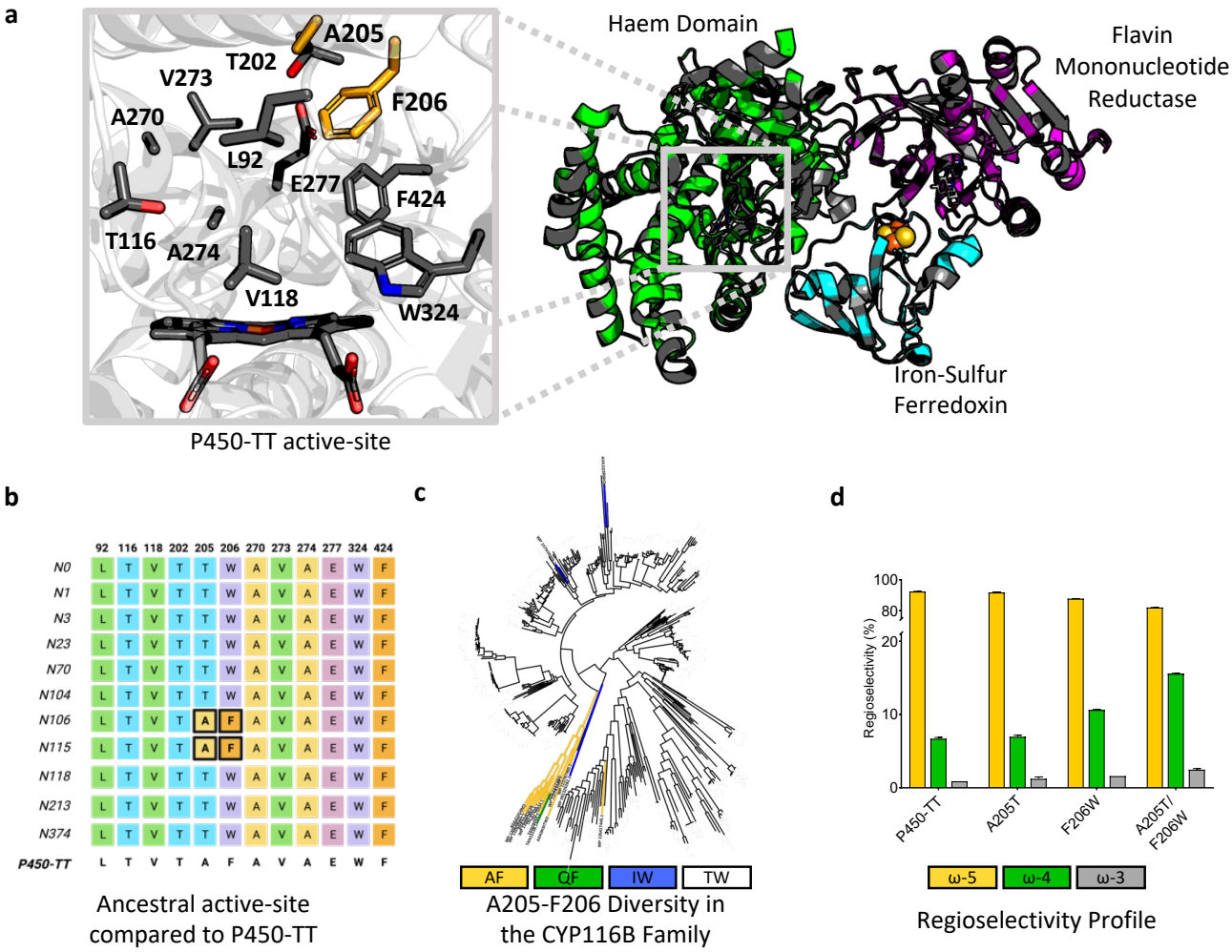

Fig. 3. Surprisingly, amino acids A205 and F206 were the only active site differences in the ancestral enzymes compared to P450-TT. a. Predominantly, across all the domains, most of the non-conserved amino acids were located on the enzyme surface. The cartoon representation of P450-TT (PDB ID: 6laa) shows the locations of all the residues that are not conserved between one or more of the ancestors and P450-TT (grey cartoon). The conserved haem domain residues are in green, conserved reductase residues in pink, and conserved ferredoxin residues in cyan. However, when comparing the haem domain active sites the only differences are A205 and F206 (orange sticks) between P450-TT and the ancestral sequences. b. The multiple sequence alignment of the active site residues in the ancestral sequences compared to P450-TT, where the order of residues corresponds to the chronological order in the active site of P450-TT. Almost all the residues were conserved, apart from A205 and F206, which were only in P450-TT and its closest ancestors, N115 and N106. c. Phylogeny highlighting the locations of the A205 and F206 diversity in the wild-type sequences, where only 17 sequences were not the consensus TW residues in the active site. Also, the branch containing P450-TT was the main region of divergence. d. Regioselectivity observed after introducing the A205T and F206W mutations separately and together into P450-TT; the product distribution was altered while the preference for mid-chain hydroxylation was retained.

Discussion

ASR has been widely used for single domain proteins[30,36,50,51], but enzymes composed of multiple domains linked by highly variable linkers, present a particular challenge since these regions influence conformational dynamics (*e.g.*, needed for efficient electron transfer) but are also inferred with least confidence in the ASR. Here we show that ASR can also be applied successfully to the three-domain family of self-sufficient CYP116B monooxygenases, generating self-sufficient single polypeptide ancestral proteins that are all properly folded, with electron transfer across the domains and enzymatic activity retained. Conducting ASR on the full length CYP116B system has several advantages, including that the activity of the P450 domain can be directly evaluated because all ancestors are self-sufficient and do not need to be reconstituted. Furthermore, one single series of experiments can yield new synthetic reductases as well as P450s, the activity of which is verified by the ability to support P450 activity in each ancestor. These synthetic reductases should be a useful addition to natural sequences available for engineering self-sufficient fused chimeric enzyme systems[61–67], an approach which has become a very popular for generating active P450s.

We were interested in tracing the evolutionary paths leading to the development of thermostable extant enzymes, such as P450-TT; many ASR studies using sequences from mesophiles and one thermophilic family of proteins have reported increases in thermostability in deeper ancestors, but it was not clear if the same would be true for enzymes from thermophiles within a larger, predominantly mesophilic tree[29,36,37,68]. In our studies we found a range of stability profiles for the haem domain, however, no clear overall trends were observed across the tree, and the oldest ancestors were not the most thermostable. Overall, there appeared to be considerable variability in the $^{15}T_{50}$ values of ancestors of various lineages, with intermediate nodes varying in $^{15}T_{50}$ between ~ 42 and 51 °C. However, ignoring the most basal node of the tree, N0, which joins the proteobacterial clade with the

actinobacterial outgroup so cannot be strictly considered as a potential historical ancestor, a general trend towards gain of thermostability over time can be seen in lineages terminating in a thermostable extant form (*i.e.,* P450-AX, -TT, or -TB). For example, in the P450-TT lineage, for which more intermediate nodes were characterised, the ancestors closest to P450 TT were also thermostable, with both N115 and N106 having a $^{15}T_{50}$ of 50 °C, suggesting a trend towards gain of thermostability in this lineage. These differences in thermostability may reflect a transition from meso- to thermophilic bacteria in some lineages, *e.g.*, leading to *Tepidiphilus thermophilus*[69], where there would have been selection pressure for enhanced thermostability. Characterisation of additional intermediate or descendant nodes in other lineages would allow better definition of trends in other branches of the tree, including the point at which thermostability arose in different CYP116B lineages. It would be of particular interest to assess descendants of N3 and N23, which showed relatively high $^{15}T_{50}$ values, and determine whether any of the uncharacterised source organisms of their descendants are thermophiles.

These findings contrast with those from the ASR of the thermophilic myo-inositol-3-phosphate synthase family proposed to have descended from a thermophilic archaeal ancestor, where the resurrected ancestral forms were more stable than the extant enzymes[68]. However, in both cases, the trends in thermostability in the inferred ancestors *vs*. their extant counterparts reflect the expected changes based on the hypothesised niche of the ancestral organism; here, the ancestor is presumed to have been a mesophile since the majority of the extant forms containing a CYP116B are mesophilic, whereas the ancestral myo-inositol-3-phosphate synthase is proposed to have existed in an organism that was more thermophilic than its descendants.

It has been proposed that ASR using maximum likelihood methods is biased towards predicting ancestors with enhanced thermostability[70]. However, the trends seen here in lineages leading to thermophiles suggest instead that any such bias is likely to be small relative to the trends related to evolution of biological niche (*i.e.,* meso- to thermophile). This implies that it is not necessarily feasible to obtain a more thermostable enzyme simply by inferring and resurrecting ancestors of a group of extant sequences without considering the possible evolutionary pathway in the context of the biological niche of the extant enzymes. Rather, as shown by comparison with other studies[20,36–39], candidate families for which this strategy is most likely to succeed are those where there has been broad diversification of enzymes in a protein family from a common ancestor in mesophiles, leading to large extant families with diversified functions.

We have shown here that ASR can be used to obtain a panel of biocatalysts, each with a unique pattern of regioselectivity towards fatty acid hydroxylation. Notably, ancestor N374 would be a good starting point for engineering for efficient ω-2 hydroxylation, having shown ~ 70 % regioselectivity and 42% product conversion here, as well as relatively high $^{15}T_{50}$, although its longer term stability was not clear. Such predominant ω-2 hydroxylation had not been observed previously with other P450s and as such constitutes a new enzyme selectivity. Similarly, N70, N106, and N115 would be useful templates for engineering a higher efficiency biocatalyst for ω-5 hydroxylation, since they showed high regioselectivity, almost quantitative conversion of decanoic acid under the conditions used here, along with an overall expression yield that was comparable to or higher than the extant ω-5 hydroxylase, P450-TT. Directed evolution by recombination of these ancestors by CLADE may be particularly productive[37].

The distinctive pathway to the unique mid-chain selectivity of wild-type, P450-TT, highlights the utility of ASR to provide insight into the determinants of fatty acid hydroxylation regioselectivity and other properties. The shifts in regioselectivity seen here appear to occur steadily along specific evolutionary lineages, with the most dramatic being from the poor overall conversion and sub-terminal hydroxylation in N104 to the higher conversion and mid-chain ω-5 selectivity of N106, ancestors that are separated by a single node (N105) in the tree (which contains the AF motif at residues 205-206). Inspection of the sequence differences in this lineage compared to the rest of the ancestors may reveal sequence elements related to the enhanced overall turnover, thermostability or expression yield, not only regioselectivity, an investigation that could also be accelerated by a directed evolution approach[71].

In conclusion, facilitated by the steadily expanding pool of genome and metagenome sequences, ASR is a useful tool to both find biocatalysts with new reaction fingerprints, which can be accessed *via* considerably streamlined screening, as well as to understand trends in reaction selectivity and other properties. Fatty acid hydroxylation is an industrially important and chemically challenging target for regioselective C-H activation and ASR has been able to produce biocatalysts with new regioselectivities, compared to natural sequences explored previously. Our findings also provide insights into the structure-function relationships leading to the unique regioselectivity of P450-TT, *i.e.,* a shift towards mid-chain hydroxylation. This approach has provided a library of diverse artificial templates both for the P450 and the reductase domains and some landmarks in the sequence space around these enzymes that can be used as starting points in future protein engineering campaigns to achieve efficient and regioselective C-H hydroxylation.

Methods

Sequence gathering and ASR. Sequences were gathered from NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and UniProt (https://www.uniprot.org/blast/) databases based on sequence identity to the whole P450-TT sequence with a 37% cut-off. Duplicates were removed using SeqScrub[72]. Sequences with highly divergent regions suggestive of errors in annotation or calling of start and stop sites were removed in iterative rounds of sequence curation. The final total number of sequences was 467, and all had >45% identity compared to P450-TT. Sequences were aligned with MAFFT (v7.487) using the L-INS-i strategy[73]. A phylogenetic tree was inferred using IQ-TREE 2 (v2.1.2) with ModelFinder identifying LG+F+R10 as the best-fit model according to both the Akaike Information Criterion and the Bayesian Information Criterion[74,75]. The tree (Supplementary Fig. 1) was labelled with SH-like aLRT, approximate Bayes parametric test, and ultrafast bootstrap values[76]. The tree was rooted with the Actinobacteria clade separated from the Proteobacteria clade[56,57]; all curated sequences collected from the actinobacterial clade were included in the outgroup. Ancestral sequences were reconstructed using Graphical Representation of Ancestral Sequence Predictions (GRASP), from the alignment and the Newick tree files[20]. Ancestors were predicted *via* a joint, maximum likelihood reconstruction approach to facilitate comparison of multiple ancestors across the tree[77]. A pairwise sequence identity matrix

between the extant CYP116B panel and the ancestral enzymes (Supplementary Table 1) was calculated using the sequence manipulation suite (https://www.bioinformatics.org/sms2/ident_sim.html). Amino acid compositions (Supplementary Table 3) were calculated using the ProtParam program in the ExPASy proteomic server.

Plasmid generation. The selected ancestors were designed with an in-phase N-terminal His-tag and cloned into pET28a using the restriction sites *Nde*I and *Xho*I. The sequences were codon optimised for *E. coli* and the plasmids were synthesised by Twist Biosciences (CA, USA). Site-directed mutagenesis was performed using the primers and conditions in Supplementary Tables 7 and 8. The PCR mixture was treated with *Dpn*I for 1 h at 37 °C, NEB-5α competent *E. coli* cells were transformed with 2 µL of the resulting treated DNA, in accordance with the manufacturer's instructions (New England Biolabs). The constructs and mutations were confirmed *via* DNA sequencing, carried out by Eurofins Genomics (Germany).

Protein overexpression and purification. Chemically competent *E. coli* BL21 (DE3) cells were used for overexpression. Cultures were incubated in 800 mL M9 media in 2 L baffled flasks. M9 media contained M9 salts, 0.4% glucose (w/v), 0.05% $FeCl_3$ (w/v), 1 mM $MgSO_4$, and 1 mM $CaCl_2$, 50 µg/mL kanamycin and 1% starter culture (v/v). The cultures were incubated at 37 °C, 200 rpm until the optical density (OD_{600}) was 0.7−0.9, then induced with β-D-1-thiogalactopyranoside (0.4 mM) and supplemented with 5-aminolaevulinic acid (0.5 mM). Overexpression was carried out for 20 h at 25 °C, and 200 rpm agitation. Cells were harvested by centrifugation (2831 g, 20 min, 4 °C). The cell pellets were stored at −20 °C.

Protein purification was performed *via* immobilized metal ion affinity chromatography. The pellets were resuspended in the lysis buffer (0.1 M potassium phosphate, 20 mM imidazole, 300 mM NaCl, 10% glycerol, pH 8). Cells were lysed by ultrasonication using a multiprobe sonicator (QSONICA) for a 10 mins process time with 10 s on and 10 s off, 25% amplitude. Cell debris was removed *via* centrifugation (48 384 g, 45 min, 4 °C). Protein purification was carried out using gravity flow columns and by binding the protein to Ni-agarose resin (Qiagen), in accordance with the manufacturer's instructions. The supernatant was incubated with the resin at 4 °C for 45 min on a rotating mixer, before loading the column. The column was washed with 10 column volumes of lysis buffer, then eluted with elution buffer (0.1 M potassium phosphate buffer, 300 mM imidazole, 300 mM NaCl, 10 % glycerol, pH 8). The eluted protein was concentrated using a 30 000 MW Vivaspin column (GE Healthcare), before being desalted *via* a PD-10 column (GE Healthcare) and stored in storage buffer (0.1 M potassium phosphate, 300 M NaCl, 30% glycerol, pH 8). SDS-PAGE analysis confirmed the presence of full-length purified protein (Supplementary Figure 8). The concentration of purified protein was determined by Fe(II).CO *vs.* Fe(II)-difference spectroscopy (Supplementary Figure 9) as described by Omura and Sato[78], using the extinction coefficient of 91 mM^{-1} cm^{-1}. CO difference spectroscopy was carried out on an Eppendorf BioSpectrometer. The purified protein was concentrated, aliquoted, flash frozen in liquid nitrogen and stored at −80 °C.

Screening for thermal stability. Extant and ancestral CYP116B proteins were expressed in BL21(DE3) *E. coli* at 25 °C for 48 h and total P450 concentrations were determined in intact cells as described by Johnston *et al*. (2008).[79] *E. coli* cell pellets were resuspended in whole cell spectral assay buffer (0.1 M potassium phosphate buffer pH 7.4, 20 mM D-glucose, 6 mM magnesium acetate) to a final concentration of 0.06 g wet cells per mL before incubation at a range of temperatures (25-80 °C) in a BioRad™ T100 thermal cycler (Bio-Rad Laboratories, Inc., Hercules, California, USA) for 15 or 60 min. The residual P450 content was then measured. The proportion of total P450 remaining was plotted against temperature and the $^{15}T_{50}$ or $^{60}T_{50}$ (*i.e.,* the temperature at which 50 % of the P450 is lost after a 15- or 60-min incubation) was calculated using an IC_{50} curve fit in GraphPad Prism 8.0. Data are means of three replicates.

Screening of activity towards decanoic acid. The biotransformations were performed using purified protein with a final enzyme concentration of 10 or 50 µM. Decanoic acid was added from 80 mM stock solution in DMSO, to give a final concentration of 0.5 mM (0.6% v/v final DMSO concentration). The reactions were in 0.1 M potassium phosphate buffer, pH 8, up to a final volume of 200 µL, and utilised a NADPH-regenerating system consisting of glucose dehydrogenase (GDH; 1 mg mL^{-1}, CDX-901, Codexis), $NADP^+$ (1 mg mL^{-1}) and D-glucose (10 mg mL^{-1}). The reactions were conducted in triplicate and incubated for 20 h, at 30 °C, and with agitation at 200 rpm.All biotransformations were worked up by quenching initially with 10 µL 5 M HCl (0.1 M), followed by the addition of an equal volume of dichloromethane containing nonanoic acid (1 mM) as an internal standard. The samples were centrifuged (20 627 g, 10 min); the organic phase was separated and dried with $MgSO_4$, and centrifuged again (20 627 g, 10 min); and then samples were derivatised with *N*-methyl-*N*-(trimethylsilyl)trifluoroacetamide (MTSFA, 20 µL) for 15 min at 50 °C for GC-MS analysis. The GC system (Santa Clara, CA, USA) was coupled to an MS detector (Agilent Technologies) and Agilent HP-1 MS column (30 m x 0.32 mm x 0.25 µm) with helium gas as the carrier. The GC-MS method was 120 °C held for 4 min, increased to 200 °C at a rate of 20 °C min^{-1} then held at 200 °C for 2 min.

Data Availability

The data supporting the findings of this study are available within the article, in its Supplementary Information.

Acknowledgements

This work was supported by the European Research Council (788231-ProgrES-ERC-2017-ADG).

Author contributions

E.M.G., and S.L.F. managed and supervised the project, with input from M.A.H., and Y.G. B.S.J., G.F., N.K., and R.E.S.T., gathered, and curated sequences, then performed the phylogenetic analysis. B.S.J. designed the plasmids, carried out the molecular biology, enzyme purifications, and biotransformations. C. R. undertook the thermal stability and expression assays. B.S.J., C.R., E.M.G., and S.L.F. wrote the manuscript and generated the figures.

Competing Interests Statement

The authors declare no competing interests.

References

1. Kim, K. R. & Oh, D. K. Production of hydroxy fatty acids by microbial fatty acid-hydroxylation enzymes. *Biotechnol Adv* **31**, 1473–1485 (2013).

2. Thesseling, F. A. *et al.* Novel insights into oxidation of fatty acids and fatty alcohols by cytochrome P450 monooxygenase CYP4B1. *Arch Biochem Biophys* **679**, 108216–108225 (2020).

3. Hammerer, L., Winkler, C. K. & Kroutil, W. Regioselective biocatalytic hydroxylation of fatty acids by cytochrome P450s. *Catal Letters* **148**, 787–812 (2018).

4. Yang, Y.-H., Wang, J.-L., Miranda, C. L. & Buhler, D. R. CYP2M1: Cloning, sequencing, and expression of a new cytochrome P450 from rainbow trout liver with fatty acid (omega-6)-hydroxylation activity. *Arch Biochem Biophys* **352**, 271–280 (1998).

5. Park, H. G. *et al.* CYP52A23 from Candida albicans and its substrate preference for fatty acid hydroxylation. *Arch Biochem Biophys* **671**, 27–34 (2019).

6. Porter, J. L. *et al.* Characterisation of CYP102A25 from Bacillus marmarensis and CYP102A26 from Pontibacillus halophilus: P450 homologues of BM3 with preference towards hydroxylation of medium-chain fatty acids. *ChemBioChem* **19**, 513–520 (2018).

7. Munday, S. D., Maddigan, N. K., Young, R. J. & Bell, S. G. Characterisation of two self-sufficient CYP102 family monooxygenases from Ktedonobacter racemifer DSM44963 which have new fatty acid alcohol product profiles. *Biochim Biophys Acta* **1860**, 1149–1162 (2016).

8. Whitehouse, C. J. C. C., Bell, S. G. & Wong, L.-L. L. P450 BM3 (CYP102A1): Connecting the dots. *Chem Soc Rev* **41**, 1218–1260 (2012).

9. Porter, J. L. *et al.* Cloning, expression and characterisation of P450-Hal1 (CYP116B62) from Halomonas sp. NCIMB 172: A self-sufficient P450 with high expression and diverse substrate scope. *Enzyme Microb Technol* **113**, 1–8 (2018).

10. Child, S. A., Rossi, V. P. & Bell, S. G. Selective ϖ-1 oxidation of fatty acids by CYP147G1 from Mycobacterium marinum. *Biochim Biophys Acta Gen Subj* **1863**, 408–417 (2019).

11. Girhard, M., Schuster, S., Dietrich, M., Dürre, P. & Urlacher, V. B. Cytochrome P450 monooxygenase from Clostridium acetobutylicum: A new α-fatty acid hydroxylase. *Biochem Biophys Res Commun* **362**, 114–119 (2007).

12. Honda Malca, S. *et al.* Bacterial CYP153A monooxygenases for the synthesis of omega-hydroxylated fatty acids. *Chem. Commun.* **48**, 5115–5117 (2012).

13. Manning, J. *et al.* Regio- and enantio-selective chemo-enzymatic C−H-lactonization of decanoic acid to (S)-δ-decalactone. *Angew. Chem., Int. Ed.* **58**, 5668–5671 (2019).

14. Cirino, P. C. & Arnold, F. H. Regioselectivity and activity of cytochrome P450 BM-3 and mutant F87A in reactions driven by hydrogen peroxide. *Adv Synth Catal* **344**, 932–937 (2002).

15. Maseme, M. J., Pennec, A., van Marwijk, J., Opperman, D. J. & Smit, M. S. CYP505E3: A novel self-sufficient ω-7 in-chain hydroxylase. *Angew. Chem., Int. Ed.* **59**, 10359–10362 (2020).

16. Brühlmann, F. *et al.* Engineering cytochrome P450 BM3 of Bacillus megaterium for terminal oxidation of palmitic acid. *J Biotechnol* **184**, 17–26 (2014).

17. Ost, T. W. B. *et al.* Rational re-design of the substrate binding site of flavocytochrome P450 BM3. *FEBS Lett* **486**, 173–177 (2000).

18. Hammerer, L., Friess, M., Cerne, J., Fuchs, M. & Steinkellner, G. Controlling the regioselectivity of fatty acid hydroxylation (C 10) at α - and β -position by CYP152A1 (P450Bs β) variants. *ChemCatChem* **11**, 5642–5649 (2019).

19. Zhang, K. *et al.* Biocatalytic enantioselective β-hydroxylation of unactivated C−H bonds in aliphatic carboxylic acids. *Angew. Chem., Int. Ed.* **61**, e20220429 (2022).

20. Foley, G. *et al.* Engineering indel and substitution variants of diverse and ancient enzymes using Graphical Representation of Ancestral Sequence Predictions (GRASP). *PLoS Comput Biol* **18**, e1010633 (2022).

21. Watanabe, K., Ohkuri, T., Yokobori, S. I. & Yamagishi, A. Designing thermostable proteins: Ancestral mutants of 3-isopropylmalate dehydrogenase designed by using a phylogenetic tree. *J Mol Biol* **355**, 664–674 (2006).

22. Gonzalez, D. *et al.* Ancestral mutations as a tool for solubilizing proteins: The case of a hydrophobic phosphate-binding protein. *FEBS Open Bio* **4**, 121–127 (2014).

23. Whitfield, J. H. *et al.* Construction of a robust and sensitive arginine biosensor through ancestral protein reconstruction. *Protein Sci.* **24**, 1412–1422 (2015).

24. Wilson, C. *et al.* Using ancient protein kinases to unravel a modern cancer drug's mechanism. *Science* **347**, 882–886 (2015).

25. Risso, V. A. *et al.* Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history. *Mol Biol Evol* **32**, 440–455 (2015).

26. Thomas, A., Cutlan, R., Finnigan, W., van der Giezen, M. & Harmer, N. Highly thermostable carboxylic acid reductases generated by ancestral sequence reconstruction. *Commun Biol* **2**, 429–440 (2019).

27. Hueting, D. A., Vanga, S. R. & Syrén, P.-O. Thermoadaptation in an ancestral diterpene cyclase by altered loop stability. *J Phys Chem B* **126**, 3809–3821 (2022).

28. Barandiaran, L. *et al.* Enzymatic upgrading of nanochitin using an ancient lytic polysaccharide monooxygenase. *Commun Mater* **3**, 55–64 (2022).

29. Thomson, R. E. S., Carrera-Pacheco, S. E. & Gillam, E. M. J. Engineering functional thermostable proteins using ancestral sequence reconstruction. *J. Biol. Chem.* **298**, 102435 (2022).

30. Livada, J., Vargas, A., Martinez, C. A. & Lewis, R. Ancestral sequence reconstruction enhances gene mining efforts for industrial ene reductases by expanding enzyme panels with thermostable catalysts. *ACS Catal* **13**, 2576–2585 (2023).

31. Babkova, P., Sebestova, E., Brezovsky, J., Chaloupkova, R. & Damborsky, J. Ancestral haloalkane dehalogenases show robustness and unique substrate specificity. *ChemBioChem* **18**, 1448–1456 (2017).

32. Nicoll, C. R. *et al.* Ancestral-sequence reconstruction unveils the structural basis of function in mammalian FMOs. *Nat Struct Mol Biol* **27**, 14–24 (2020).

33. Sun, Y., Calderini, E. & Kourist, R. A reconstructed common ancestor of the fatty acid photo-decarboxylase clade shows photo-decarboxylation activity and increased thermostability. *ChemBioChem* **22**, 1833–1840 (2021).

34. Gumulya, Y. & Gillam, E. M. J. Exploring the past and the future of protein evolution with ancestral sequence reconstruction: The 'retro' approach to protein engineering. *Biochem.* **474**, 1–19 (2017).

35. Spence, M. A., Kaczmarski, J. A., Saunders, J. W. & Jackson, C. J. Ancestral sequence reconstruction for protein engineers. *Curr Opin Struct Biol* **69**, 131–141 (2021).

36. Gumulya, Y. *et al.* Engineering highly functional thermostable proteins using ancestral sequence reconstruction. *Nat Catal* **1**, 878–888 (2018).

37. Gumulya, Y. *et al.* Engineering thermostable CYP2D enzymes for biocatalysis using Combinatorial Libraries of Ancestors for Directed Evolution (CLADE). *ChemCatChem* **11**, 841–850 (2019).

38. Hartz, P. *et al.* Resurrection and characterization of ancestral CYP11A1 enzymes. *FEBS Journal* **288**, 6510–6527 (2021).

39. Harris, K. L. *et al.* Ancestral sequence reconstruction of a cytochrome P450 family involved in chemical defense reveals the functional evolution of a promiscuous, xenobiotic-metabolizing enzyme in vertebrates. *Mol Biol Evol* **39**, msac116 (2022).

40. O'Reilly, E., Köhler, V., Flitsch, S. L. & Turner, N. J. Cytochromes P450 as useful biocatalysts: Addressing the limitations. *Chem. Commun.* **47**, 2490–2501 (2011).

41. Tavanti, M., Porter, J. L., Sabatini, S., Turner, N. J. & Flitsch, S. L. Panel of new thermostable CYP116B self-sufficient cytochrome P450 monooxygenases that catalyze C−H activation with a diverse substrate scope. *ChemCatChem* **10**, 1042–1051 (2018).

42. Devamani, T. *et al.* Catalytic promiscuity of ancestral esterases and hydroxynitrile lyases. *J Am Chem Soc* **138**, 1046–1056 (2016).

43. Wilding, M. *et al.* Reverse engineering: Transaminase biocatalyst development using ancestral sequence reconstruction. *Green Chemistry* **19**, 5375–5380 (2017).

44. Alcolombri, U., Elias, M. & Tawfik, D. S. Directed evolution of sulfotransferases and paraoxonases by ancestral libraries. *J Mol Biol* **41**, 837–853 (2011).

45. Conti, G., Pollegioni, L., Molla, G. & Rosini, E. Strategic manipulation of an industrial biocatalyst - Evolution of a cephalosporin C acylase. *FEBS Journal* **281**, 2443–2455 (2014).

46. Shih, P. M. *et al.* Biochemical characterization of predicted Precambrian RuBisCO. *Nat Commun* **7**, 10382–10392 (2016).

47. Schriever, K. *et al.* Engineering of ancestors as a tool to elucidate structure, mechanism, and specificity of extant terpene cyclase. *J Am Chem Soc* **143**, 3794–3807 (2021).

48. Zeng, B. *et al.* Engineering and screening of novel β-1,3-xylanases with desired hydrolysate type by optimized ancestor sequence reconstruction and data mining. *Comput Struct Biotechnol J* **20**, 3313–3321 (2022).

49. Chen, S. *et al.* Big data mining, rational modification, and ancestral sequence reconstruction inferred multiple xylose isomerases for biorefinery. *Sci Adv* **9**, eadd8835 (2023).

50. Bailleul, G. *et al.* Evolution of enzyme functionality in the flavin-containing monooxygenases. *Nat Commun* **14**, 1042–1031 (2023).

51. Chiang, C.-H. *et al.* Deciphering the evolution of flavin-dependent monooxygenase stereoselectivity using ancestral sequence reconstruction. *PNAS* **120**, e2218248120 (2023).

52. Tavanti, M. *et al.* The crystal structure of P450-TT heme-domain provides the first structural insights into the versatile class VII P450s. *Biochem Biophys Res Commun* **501**, 846–850 (2018).

53. Ciaramella, A., Catucci, G., Gilardi, G. & Di Nardo, G. Crystal structure of bacterial CYP116B5 heme domain: New insights on class VII P450s structural flexibility and peroxygenase activity. *Int J Biol Macromol* **140**, 577–587 (2019).

54. Zhang, L. *et al.* Structural insight into the electron transfer pathway of a self-sufficient P450 monooxygenase. *Nat Commun* **11**, 2676–2681 (2020).

55. Gong, R. *et al.* Crystal structure of an intact type IV self-sufficient cytochrome P450 monooxygenase. Preprint at https://doi.org/10.2210/pdb6KBH/pdb (2020).

56. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* **36**, 996–1004 (2018).

57. Coleman, G. A. *et al.* A rooted phylogeny resolves early bacterial evolution. *Science* **372**, eabe0511 (2021).

58. Chen, W. M. *et al.* Caldimonas taiwanensis sp. nov., a amylase producing bacterium isolated from a hot spring. *Syst Appl Microbiol* **28**, 415–420 (2005).

59. Elbanna, K. *et al.* Schlegelella thermodepolymerans gen. nov., sp. nov., a novel thermophilic bacterium that degrades poly(3-hydroxybutyrate-co-3-mercaptopropionate). *Int J Syst Evol Microbiol* **53**, 1165–1168 (2003).

60. Harris, K. L., Thomson, R. E. S., Strohmaier, S. J., Gumulya, Y. & Gillam, E. M. J. Determinants of thermostability in the cytochrome P450 fold. *Biochim Biophys Acta Proteins Proteom* **1866**, 97–115 (2018).

61. Renata, H. Engineering catalytically self-sufficient P450s. *Biochemistry* **62**, 253–261 (2023).

62. Nodate, M., Kubota, M. & Misawa, N. Functional expression system for cytochrome P450 genes using the reductase domain of self-sufficient P450RhF from Rhodococcus sp. NCIMB 9784. *Appl Microbiol Biotechnol* **71**, 455–462 (2006).

63. Li, S., Podust, L. M. & Sherman, D. H. Engineering and analysis of a self-sufficient biosynthetic cytochrome P450 PikC fused to the RhFRED reductase domain. *J Am Chem Soc* **129**, 12940–12941 (2007).

64. Robin, A. *et al.* Engineering and improvement of the efficiency of a chimeric [P450cam-RhFRed reductase domain] enzyme. *Chem. Commun.* **18**, 2478–2480 (2009).

65. Robin, A. *et al.* Chimeric self-sufficient P450cam-RhFRed biocatalysts with broad substrate scope. *Beilstein J. Org. Chem.* **7**, 1494–1498 (2011).

66. Chen, C. C. *et al.* Molecular basis for a toluene monooxygenase to govern substrate selectivity. *ACS Catal* **12**, 2831–2839 (2022).

67. Zetzsche, L. E. *et al.* Biocatalytic oxidative cross-coupling reactions for biaryl bond formation. *Nature* **603**, 79–85 (2022).

68. Butzin, N. C. *et al.* Reconstructed ancestral myo-inositol-3-phosphate synthases indicate that ancestors of the Thermococcales and Thermotoga species were more thermophilic than their descendants. *PLoS One* **8**, (2013).

69. Hart, K. M. *et al.* Thermodynamic system drift in protein evolution. *PLoS Biol* **12**, e1001994 (2014).

70. Williams, P. D., Pollock, D. D., Blackburne, B. P. & Goldstein, R. A. Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comput Biol* **2**, 0598–0605 (2006).

71. Behrendorff, J. B. Y. H. *et al.* Directed evolution reveals requisite sequence elements in the functional expression of P450 2F1 in Escherichia coli. *Chem Res Toxicol* **25**, 1964–1974 (2012).

72. Foley, G., Sützl, L., D'Cunha, S. A., Gillam, E. M. J. & Bodén, M. SeqScrub: A web tool for automatic cleaning and annotation of FASTA file headers for bioinformatic applications. *Biotechniques* **67**, 50–54 (2019).

73. Katoh, K. & Toh, H. Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics* **26**, 1899–1900 (2010).

74. Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* **37**, 1530–1534 (2020).

75. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**, 587–589 (2017).

76. Thi Hoang, D. *et al.* UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol* **35**, 518–522 (2017).

77. Pupko, T., Pe, I., Shamir, R. & Graur, D. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol Biol Evol* **17**, 890–896 (2000).

78. Omura, T. & Sato, R. The carbon moxoide-binding pigment of liver microsomes. *J Biol Chem* **239**, 2370–2378 (1964).

79. Johnston, W. A., Huang, W., De Voss, J. J., Hayes, M. A. & Gillam, E. M. J. Quantitative whole-cell cytochrome P450 measurement suitable for high-throughput application. *J Biomol Screen* **13**, 135–141 (2008).