Identifying and engineering ancient variants of enzymes using Graphical Representation of Ancestral Sequence Predictions (GRASP)

Gabriel Foley[1], Ariane Mora[1]*, Connie M. Ross[1]*, Scott Bottoms[2]*, Leander Sützl[3]*, Marnie L. Lamprecht[1], Julian Zaugg[1], Alexandra Essebier[1], Brad Balderson[1], Rhys Newell[1], Raine E. S. Thomson[1], Bostjan Kobe[1,4], Ross T. Barnard[1], Luke Guddat[1], Gerhard Schenk[1,5], Jörg Carsten[6], Yosephine Gumulya[1], Burkhard Rost[7], Dietmar Haltrich[3], Volker Sieber[2,6,1], Elizabeth M. J. Gillam[1†] and Mikael Bodén[1†]

[1] School of Chemistry and Molecular Biosciences, The University of Queensland, Australia.

[2] Campus Straubing for Biotechnology and Sustainability, Technische Universität München, Straubing, Germany.

[3] Institut für Lebensmitteltechnologie, Universität für Bodenkultur Wien, Vienna, Austria.

[4] Institute for Molecular Bioscience and Australian Infectious Diseases Research Centre, The University of Queensland, Australia.

[5] Sustainable Minerals Institute, The University of Queensland, Australia.

[6] Zentralinstitut für Katalyseforschung, Technische Universität München, Munich, Germany.

[7] Fakultät für Informatik, Technische Universität München, Munich, Germany.

Abstract: Ancestral sequence reconstruction is a technique which is gaining widespread use in molecular evolution studies and protein engineering. Here we present Graphical Representation of Ancestral Sequence Predictions (GRASP) that can be used to infer and explore ancestral variants of protein families with more than 10,000 members. GRASP uses partial order graphs to represent homology in very large data sets, which are intractable with current inference tools and may, for example, be used to engineer proteins by identifying ancient variants of enzymes. We demonstrate that (1) across three distinct enzyme families, GRASP predicts ancestor sequences, all of which demonstrate enzymatic activity, (2) within-family insertions and deletions can be used as building blocks to support the engineering of biologically active ancestors *via* a new source of ancestral variation, and (3) generous inclusion of sequence data encompassing great diversity leads to *less* variance in ancestor sequence.

*Equal contributors

†Correspondence to: m.boden@uq.edu.au, e.gillam@uq.edu.au

Introduction

Sequencing technology is driving the identification of the *extant* (modern) portion of the universe of biological sequences[1,2,3]. With this increased coverage of natural diversity we are now better placed than ever before to leverage ancestral sequence reconstruction (ASR) to recover the *ancestral* portion and trace the evolutionary events that determine biological function and structure[4]. This is especially useful for protein engineering; the evolutionary record reveals essential cues for the discovery of new enzymes and resurrection of ancestral enzymes often generates enzymes with novel properties that can be exploited in biocatalysis[5,6,7,8].

The ability to perform ASR on large-scale data has been limited by the available methodology and accompanying technology. A recent review highlighted 12 studies from the past decade which each sought to evaluate sources of ambiguity in ancestral inferences[9]. Data set sizes within these studies ranged from 21 to 456 sequences, with an average of 168 sequences. Current methods for performing ASR have reached practical upper limits on data set size, which constrain our ability to adequately represent and accurately analyse enzymes that have been evolving for billions of years. We have developed the tool Graphical Representation of Ancestral Sequence Predictions (GRASP) to take advantage of the rapidly expanding number of known protein sequences and the information from biological diversity that can be mined from large protein families.

Processing large amounts of data is not just a quantitative problem, but a qualitative one as well. The evolutionary models employed to quantify ancestral states depend on an accurate representation of homology and remote homologs are likely to have resulted from numerous evolutionary events that confound current phylogenetic analysis techniques. Lee et al.[10] demonstrated how a partial order graph (POG) can be used to represent and support the alignment of widely different sequences. The risk of aligning sequence fragments with different evolutionary origins motivated us to use the POG data structure to separate distinct sources of sequence diversity at evolutionary branch points. POGs enable us to negotiate sequence variance and to track evolutionary events across time. The premise of our study is that this significant increase of the scope of ASR will provide (a) a rich resource for evolutionary studies, and (b) valuable guidance for protein engineering given the demonstrated usefulness of ancestral enzymes as robust templates for directed evolution[7]. Consequently, our method was designed with a view to identifying substitutions, insertions, and deletions that may be combined to form sequence configurations inspired by, but not necessarily present in, either extant or inferred sequences.

We tested the approach by inferring ancestors from different enzyme families, exemplifying various degrees of sequence number, functional diversity, and sequence similarity. All of the enzyme families studied are attractive from a protein engineering perspective, as ASR offers efficient pathways towards industrially relevant outcomes such as increased thermal stability or altered substrate specificity. Resurrected ancestral proteins from the following families were produced

and evaluated in terms of their structure and function.

1. The glucose-methanol-choline (GMC) oxidoreductases represent a superfamily of enzymes with varying biological functions and industrial applications; we focused on the glucose dehydrogenase (GDH, EC 1.1.5.9) and glucose oxidase (GOx, EC 1.1.2.4) families[11].

2. Members of cytochrome P450 subfamily 2 (CYP2) play a key role in drug and xenobiotic metabolism in metazoans[12]. Here we concentrated on the CYP2U subfamily and two closely-related subfamilies, CYP2R, and CYP2D.

3. The IlvD/ED dehydratase family includes dihydroxy-acid dehydratase (EC 4.2.1.9) and several sugar acid dehydratases all containing iron-sulfur-clusters and has broad taxonomic scope[13]. We refer to this family as DHAD. It is present in bacteria, archaea, fungi, algae, and in some plants.

In addition, we also evaluated a large-scale inference of, but did not resurrect, the following family.

4. The ketol-acid reductoisomerase (KARI) family includes enzymes in the branched-chain amino acid biosynthetic pathway (similar to DHAD) present in bacteria, fungi, and plants. We focused on KARI class I for a large-scale inference, and class II for a comparison between existing tools, having previously successfully resurrected ancestors of class II enzymes[7].

In brief, in this paper we demonstrate the capacity of an ASR approach based on POGs and maximum likelihood inference to:

- perform ASR on proteins in a manner consistent with current tools when restricted to smaller data set sizes;
- perform ASR for very large protein families; specifically, we explore the impact of quantity, diversity, and taxonomic context of input sequences on predicted sequences as well as resurrected structures and functions; and
- assist in the design of biocatalysts; we evaluate the novel prospect of using ASR to track and re-purpose insertion and deletion (indel) events to compose and resurrect hybrid ancestors.

Results

GRASP infers partially ordered ancestor graphs, representing substitutions, insertions, and deletions

Unlike other reconstruction methods, GRASP uses POGs to assign sequence characters from insertions, deletions, and recombination events *over* time and *across*

clades to ensure that homologous positions are processed appropriately by an evolutionary model and to defer decisions when there is ambiguity.

GRASP infers ancestor POGs from an input POG that represents a set of aligned homologous sequences and an input phylogenetic tree describing their evolutionary relationships. It does this in three stages that are designed to deconvolute sources of sequence variation.

1. The most parsimonious history of composite indel events is determined and mapped onto the phylogenetic tree. For each position in the alignment a "character tree" is constructed that only contains phylogenetic branch points with actual sequence content (Fig. 1a).

2. For each character tree, the most probable character (to explain those observed at the leaves) is assigned to each phylogenetic branch point when performing a *joint reconstruction* (Fig. 1b). Alternatively, the probability distribution over all possible characters is inferred for each position at a nominated phylogenetic branch point when performing a *marginal reconstruction* (Fig. 1c).

3. For each phylogenetic branch point or ancestor, character trees are selectively linked to form an individual POG with nodes for characters and edges for *all* inferred combinations of indels, including a preferred path nominating a single sequence.

Inference of ancient character states for the analysis of large protein families is performed using maximum likelihood, leveraging efficient algorithms developed for probabilistic graphical models, which allows unprecedented volumes of non-redundant data to be used[14]. Inference of indel histories is done with a variation of maximum parsimony that we refer to as bi-directional edge parsimony, which tracks and scores the edges of a POG (see Methods for a complete description).

For GDH and GOx, we used GRASP to identify potential substitution variants through analysis of inferred distributions via marginal reconstruction. This analysis is an established approach which is frequently performed in ASR to account for uncertainties in reconstructed sequences, suggest ancestor variants, and explore properties such as thermal stability or substrate preference in inferred variants (Fig. 1c, Supplementary Table 1).

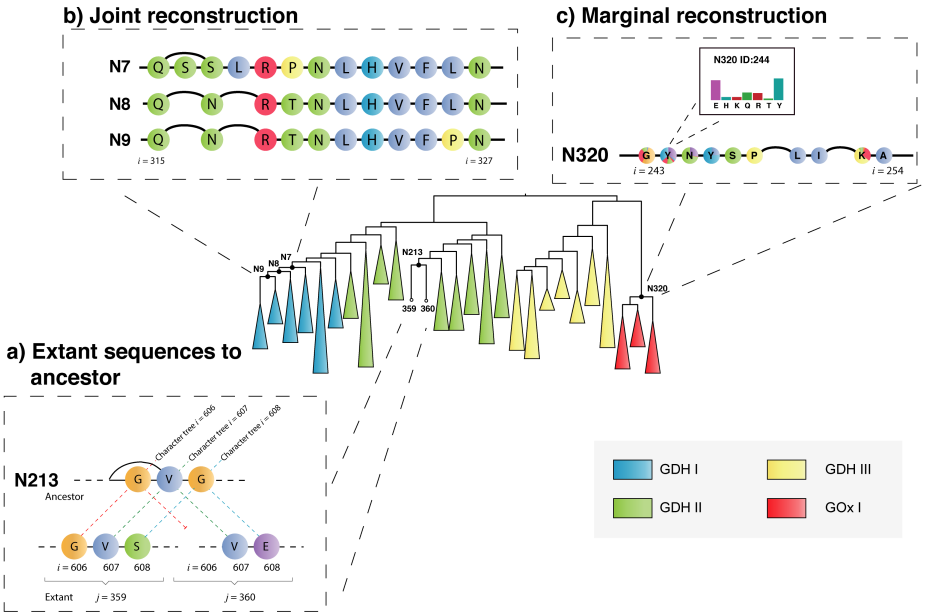

Figure 1: Phylogenetic tree showing a reconstruction of fungal GDH and GOx sequences decorated with illustrations of key concepts used in GRASP. **a**, Two extant POGs (j indicates extant sequence number) mapped to an ancestral POG. Each extant POG has a single path through strictly ordered sequence positions (i indicates position). Ancestral states are influenced by all sequences, which explains why $i = 608$ is inferred as glycine, despite glycine not appearing in either sequence $j = 359$ or 360. **b**, Three ancestor POGs showing most probable assignments from a joint reconstruction at positions $i \in \{315, ..., 327\}$ for nodes N7, N8, and N9. GRASP supports the simultaneous viewing of multiple ancestors from a joint reconstruction, enabling a direct comparison at different time points. **c**, A single ancestor POG showing inferred marginal distributions at positions $i \in \{243, ..., 254\}$ for node N320. For marginal reconstructions, nodes are coloured according to their posterior probabilities and can be queried to view histograms of these underlying distributions, as is done for position $i = 244$. The marginal reconstruction from (c) was used to reconstruct the inferred ancestor (N320) as well as an alternative ancestor in which a single amino acid (N320_Y244E) was altered based on posterior probabilities from the marginal distribution that resulted in increased thermal stability (Supplementary Table 1).

On smaller data sets, GRASP's predictions are consistent with the predictions of existing methods

We compared GRASP against two alternative ASR tools, selected due to their dominant use in the literature: FastML[15] and the `aaml` program from the Phylogenetic Analysis by Maximum Likelihood (PAML) package[16]. We were able to produce ancestral proteins from reconstructions produced by GRASP, FastML, and PAML on a CYP2U/CYP2R data set (359 sequences). The ultimate CYP2U ancestors had ~95% sequence identity and regardless of the tool used; ancestral proteins ex-

pressed at similar levels in *Escherichia coli*, displayed characteristic P450 spectra and activities towards the luciferin MultiCYP substrate, and also had similar thermal stabilities (Supplementary Fig. 1).

To make statements about the accuracy of ancestral predictions is problematic as the historically correct and complete evolutionary record is unavailable. To sidestep this issue, we first applied each tool and configuration to generate multiple predictions of the *same* principal ancestor node based on stratified, down-sampled data sets of a given sequence family. Secondly, we performed two tests asking: (a) between tools, how similar is the prediction of one tool to those of others; and (b) how similar is the prediction of one tool from the down-sampled data to a better-sampled ancestor, predicted from the *complete* family? We reasoned that a better method would be one which tended to agree with the majority of others, and one that with *less* data tended to agree with a prediction based on *more* data (assuming that more data help to improve a prediction).

A large alignment with 1,682 sequences (KARI class II, adapted from Gumulya et al.[7]) and the corresponding phylogenetic tree were divided into sub-groups and used to assess the effect of tool, data set size, and reconstruction parameters on ancestral inference (see Methods for details). We sought to corroborate any trends using a second independent data set (CYP2 with 975 sequences).

Test (a) measured similarity between ancestors at a given set of tool parameters and group size (Fig. 2a and Supplementary Fig. 2a); specifically, we observed fractional distances D/L (where D is the number of substitutions, of L non-gapped, homologous positions) between sequences predicted for each condition tested. Test (b) measured similarity in terms of fractional distances between ancestor predictions of an individual tool (with a set of parameters and group sample size) and a better-sampled ancestor using all available data (Fig. 2b and Supplementary Fig. 2b). The better-sampled ancestor for the comparison in (b) was predicted by GRASP, since a data set of this size could not be completed by FastML or PAML. A series of statistical tests were performed; first ANOVA evaluated whether choice of tool, data set size, and rate parameter setting were factors in determining how similar a predicted ancestor sequence (grouped by a specified setting) was, relative to those of alternative tools with the same setting (test a) and relative to those generated from the complete data set (test b). The t-test was then used to identify the pairs of labels (on groups) that best explained observed differences (Supplementary Figs. 3 and 4).

While GRASP performs character inference via standard models and algorithms, it does not support variable evolutionary rates at this stage. When comparing predictions between tools (test a) or between tool and better-sampled ancestor (test b), both the choice of tool and data set size separately and consistently explained the observed differences in distances; however the rate setting did not.

The choice of tool mattered for both types of comparisons across the two data sets; in most cases PAML-predicted sequences have a greater mean fractional distance to those of GRASP and FastML, than any of the alternatives. GRASP and FastML predictions were broadly indifferent, both relative to the better-sampled

ancestor and relative to PAML's ancestors. Indeed, greater sequence numbers generally reduced distances between tools and reduced distances between a tool's predictions and the ancestor based on the complete data set.

We calculated the time taken for all tools to complete the reconstructions with a run time cut-off of 48 hours (Supplementary Fig. 5) and highlighted the time taken for GRASP and FastML to complete the two larger data set sizes (Fig. 2c).

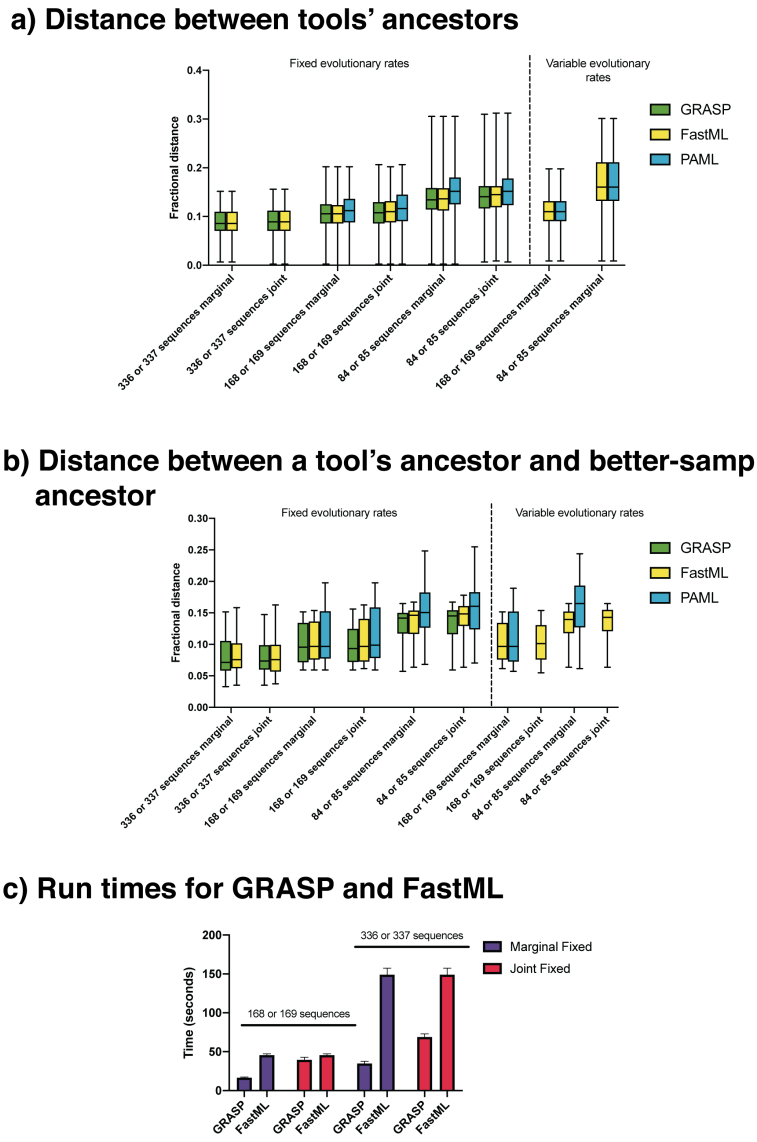

Figure 2: Tool comparison on KARI data. **a**, Average fractional distance between tools, calculated as pairwise fractional distances for each ancestral prediction for a given tool against all other ancestral predictions of other tools at 5 groups of 336 or 337 sequences, 10 groups of 168 or 169 sequences, and 20 groups of 84 or 85 sequences. Parameter combinations are joint and marginal reconstruction; and fixed or variable evolutionary rates (FastML and PAML only). **b**, Average fractional distance between a better-sampled ancestor inferred by GRASP using 1,682 sequences and each tool / parameter combination at 5, 10, and 20 groups. **c**, Run times of tools for GRASP and FastML at 5 and 10 groups; PAML is omitted due to long run times. Run times for all tools at 10 and 20 groups are shown in Supplementary Fig. 5.

Increasing sequences constrains inferred ancestral sequences

GRASP is able to process very large numbers of sequences (i.e., greater than 10,000), which is a requirement to capture the true diversity of the sequence space. Intuitively, more data equates to better coverage (and resolution) of the biological sequence space, which, if we had perfect knowledge of the true homologous relationships between residues, would imply that possible ancestral states are more robustly constrained towards canonical sequences.

In practice, we must account for obscured homology due to substitutions and indel events. Indel handling is critical for ASR, yet routinely problematic, and the accurate management of indel events is essential to decide on which sequence content to include for any particular ancestor. As data set sizes grow, the number of columns in a sequence alignment, or positions in a POG, increases substantially and the indel histories become more complicated. Therefore, increasing the number of sequences does not necessarily lead to data saturation and ancestral inferences that approach a stable, canonical sequence.

To test the effect of increasing data set size on ancestral inference, we assembled sequence data sets for the DHAD and CYP2U protein families via increments of sequence data (see Methods) and compared the ancestral inferences for each data set size (Fig. 3a-d), ranging from between 1,612 to 9,112 sequences for DHAD and between 165 and 595 sequences for CYP2U. The DHAD data sets were increased by adding sequences from across the DHAD taxonomic space, while the CYP2U data sets were increased by adding sequences from sister groups CYP2R and CYP2D while retaining the same number of CYP2U sequences at each point. For the DHAD data set we also performed a sparse reconstruction of 585 sequences, containing primarily reviewed Swiss-Prot sequences.

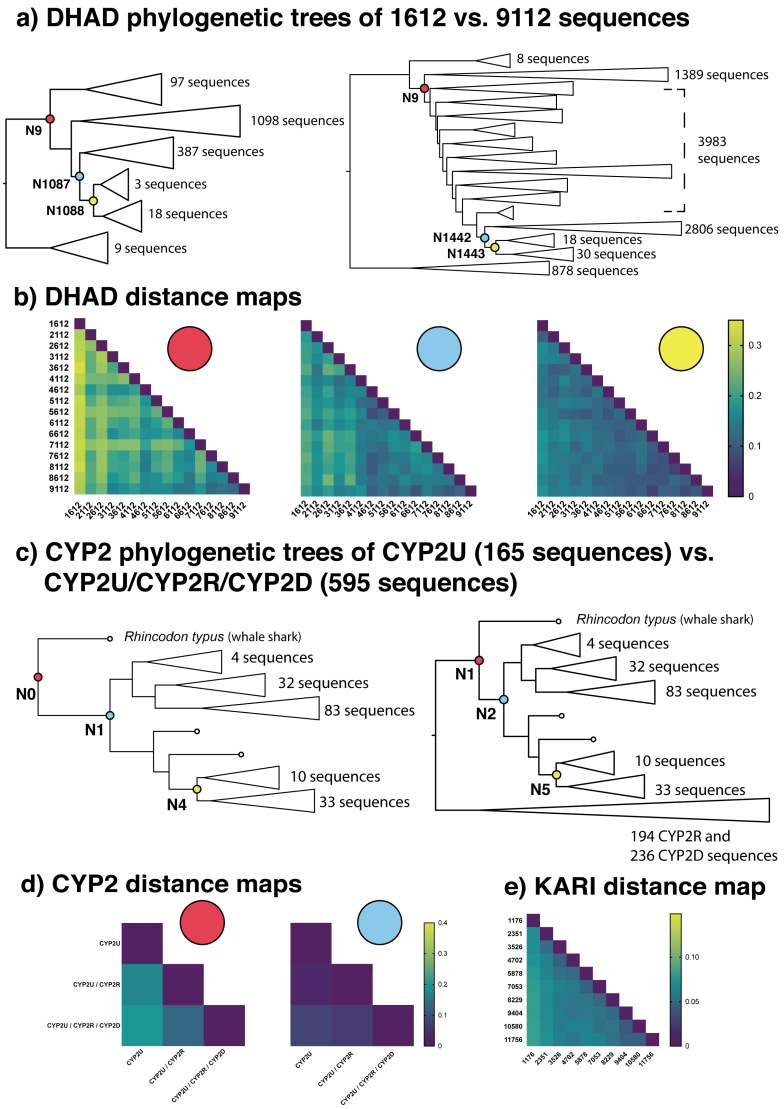

Figure 3: **a**, Phylogenetic trees of the smallest and largest DHAD data sets after producing 14 randomly sampled data sets in 500 sequence increments, added to our base data set of 1,612 sequences and reaching a maximum size of 9,112 sequences. **b**, Distance maps of the fractional distance between three nominated nodes from DHAD at incremental data set sizes. **c**, Phylogenetic trees of the smallest and largest data sets after increasing CYP2U sequences via addition of homologous subfamilies, starting with 165 CYP2U sequences then growing to 359 sequences and reaching a maximum of 595 sequences via addition of sequences from CYP2R and CYP2D, respectively. **d**, Distance maps of the fractional distance between two nominated nodes from CYP2U at incremental data set sizes. Ancestors from the N4/N5 equivalent nodes across the three data set sizes are not shown but had 98% identity. **e**, Distance map of the average fractional distance between 50 randomly selected ancestors in the KARI I data set, ranging from 1,176 to 11,756 sequences.

With GRASP, we observed that as data set size increased, the predicted ancestor sequences approached canonical forms in terms of amino acid composition at equivalent phylogenetic nodes between different tree sizes. To further illustrate these trends, we inferred KARI ancestors in regular increments ranging from 1,176 sequences to 11,756 sequences. These ancestors also converged towards canonical forms with the addition of sequences (Fig. 3e). While the number of positions in the input sequence alignment generally increases with coverage, the length of the ancestor sequences is not correlated with the number of input sequences (Supplementary Fig. 6).

GRASP is able to complete the reconstruction of the largest data sets in this study within 7 hours for DHAD (9,112 sequences, 1,381 positions in alignment) and within 6 hours for KARI (11,756 sequences, 667 positions) (Supplementary Fig. 7).

Ancestral proteins inferred from the smallest and largest data sets for both DHAD and CYP2U are active towards expected substrates, despite differences in ancestral sequence identity between the two extremes of data set size (DHAD 75%, CYP2U 80%). All DHAD ancestors displayed enzymatic activity to D-gluconate and included products of a control DHAD and traces of additional products (Supplementary Fig. 8a). We observed that three DHAD ancestral proteins from the smallest data set have thermal shift profiles comparable to those of the three ancestors that are located in equivalent tree positions in the largest data set (Supplementary Fig. 8b). For two of the three DHAD reference ancestors the melting points in the proteins from the larger reconstruction are increased by approximately 5 °C relative to their counterparts from the smaller data set (Supplementary Fig. 8b). Likewise, the inclusion of the sister clades for the CYP2U reconstruction increased the thermal stability of the ultimate CYP2U ancestors and the ancestors at each point (165, 359, and 595 sequences) were all shown to be active towards the substrate luciferin MultiCYP (Supplementary Fig. 9).

Indel variation can be used to create hybrid ancestors

A common technique to explore plausible alternative amino acids at particular sites is to select residues that show a relatively high posterior probability in a marginal reconstruction[7]. Mutations can be introduced at these positions to test the robustness of prediction and to create alternative ancestors. GRASP is able to prioritise mutations that best capture inferred probability distributions by minimising the expected relative entropy.

We hypothesised that indel events suggest plausible blocks of sequence content that could be included or excluded in identified ancestors as a novel approach to creating ancestral variants, orthogonal to substitution. GRASP utilises the history of indel events to predict modular blocks of content capable of being removed from ancestors in which they occur or inserted into ancestors that never contained these modules. In doing this, GRASP fundamentally extends the nature and practical application of modulating variation within ancestors and is capable of identifying

modular insertions that, in the case of the CYP2U ancestors, alter the protein thermal stability and substrate selectivity towards two different probe substrates. The ability to manipulate both of these properties is desirable for protein engineering.

We used GRASP to identify two distinct lineage-specific insertions within the CYP2U/CYP2R/CYP2D data set, occurring at the nodes N5 and N51 (Fig. 4). We synthesised the inferred ancestors N5 and N51, as well as a more ancient ancestor, N2, that did not contain either insertion. We removed the insertion LSEE from N5 at sequence position 153 (N5_153dLSEE) and removed the insertion LLSPP from N51 at sequence position 27 (N51_27dLLSPP). We preempted their predicted occurrence by separately inserting them into N2 at the equivalent sequence positions 152 (N2_152iLSEE) and 27 (N2_27iLLSPP). We also tested two variants of the N1 CYP2U ancestor. One form contained a CYP2U-specific insertion of 19 amino acids (N1), and the other removed this insertion to resemble the CYP2R and CYP2D sequences (N1_19dIPP...RR).

All ancestral proteins inferred via this process were heterologously expressed in *E. coli* and characterised. They were shown to express at similar levels, have the same fold, and form intact haem-thiolate linkages as indicated by the characteristic spectral peak at 450 nm in the Fe(II).CO vs. Fe(II) difference spectrum (Supplementary Fig. 10). All were catalytically active towards at least one substrate, when tested with three different P450-Glo^{TM} pro-luciferin probe substrates, luciferins MultiCYP, ME-EGE, and CEE. Therefore, the presence or absence of these lineage-specific insertions was not essential for the protein folding, co-factor binding, or interaction with the cytochrome P450 reductase. However, it was observed that the lineage-specific insertions did alter the substrate selectivity of the otherwise identical ancestors.

Both N5 and N51 are active towards both luciferins CEE and ME-EGE, while N2 is only active toward luciferin ME-EGE. Loss of the insertion LLSPP from N51 reduces its activity towards luciferin CEE, and the corresponding gain of the LLSPP insertion in the N2 ancestor increases its activity towards luciferin CEE. Neither loss of the insertion LSEE from N5 or gain of the insertion LSEE in N2 has an effect on luciferin CEE activity. The presence of the LSEE insertion in the N2 and N5 ancestors increased both ancestors' activity towards luciferin ME-EGE. Inclusion of the LLSPP insertion did not have a consistent effect in activity towards luciferin ME-EGE, whereas inclusion increased activity towards ME-EGE in N2, but not in N51. The N1 ancestor was only active towards luciferin MultiCYP, but N1_19dIPP...RR was slightly active towards all three pro-luciferin substrates, suggesting this insertion may also alter the selectivity of the ancestor.

The LLSPP insertion also modulated the thermal stability of the ancestors; the insertion produced a small but statistically significant increase in the thermal stability in both the N2 and N51 ancestors, compared to their variants lacking this insertion (Fig. 4c). This effect was not seen for the LSEE insertion (Fig. 4c), suggesting that these effects are protein and sequence specific.

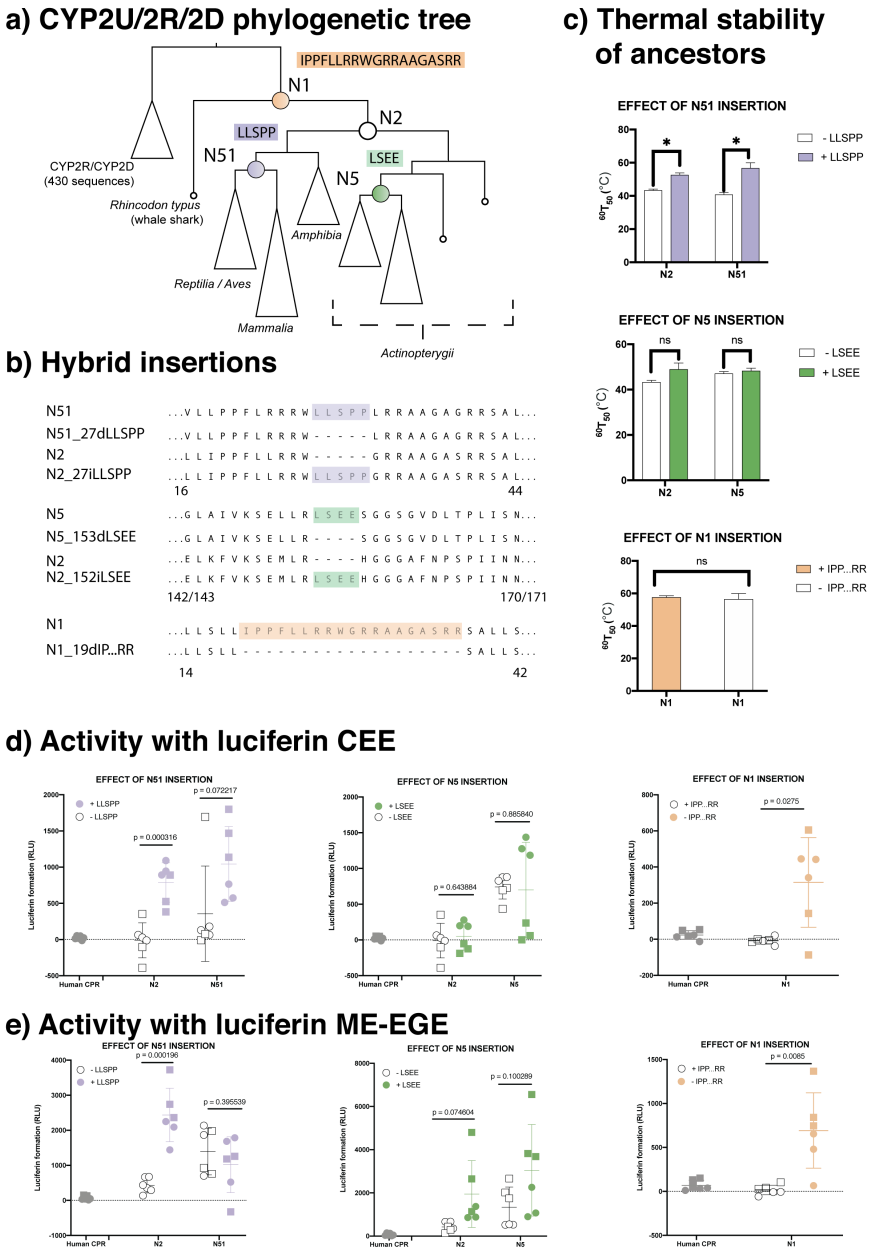

Figure 4: **a**, Phylogenetic tree showing positions of ancestors chosen for synthesis and evaluation. Coloured boxes indicate the content that was removed from correspondingly coloured ancestral nodes (N51, N5, and N1) and, in the case of the N51 and N5 insertions, preemptively inserted into N2. **b**, Amino acid sequences surrounding and including the insertion or deletion of the content at each ancestor. Numbers under sequences indicate the position numbers of the start and end columns represented in the alignment. **c**, Thermal stability assays for each ancestor with and without inserted content. **d,e,** Activity assays for the substrates luciferin CEE and luciferin ME-EGE for each ancestor with and without inserted content. Different symbols indicate different experimental repeats, lines indicate mean and standard deviation, and p-values were determined by a two-tailed Student's t-test. No data points were excluded.

Discussion

Increasing the scale at which ASR can be performed means that greater sequence and functional diversity can be explored and more complicated phylogenetic relationships can be assessed[9]. Incorporating more data from sister clades and remote homologs can allow ancestors to be inferred for more ancient evolutionary times. We substantiated these points computationally and experimentally by resurrecting twenty ancestral variants from different time points across three enzyme families (GDH/GOx, DHAD, and CYP2U), all of which were shown to be catalytically active.

ASR has been used extensively in recent years[4,9,8] and it is important to understand the relative performance of different tools, and to recognise proven principles that underpin successful methods. We note FastML is in broad agreement with GRASP, with it having closer evolutionary distances to GRASP's predictions than to PAML. This trend holds true even when FastML and PAML incorporate variable evolutionary rates. We also demonstrated that incorporating more sequence data resulted in smaller fractional distances between inferred ancestors, regardless of tool. GRASP uniquely scales to input data with more than 10,000 sequences.

Based on the analyses of KARI and DHAD we demonstrated that despite an increase in diversity, ancestral sequences converge toward canonical forms when using large data sets. Ancestral sequences generated from smaller data sets exhibit greater variation in ancestral sequence identity relative to the ancestral sequences from the larger data sets. This supports the notion that greater representation of a family provides a constraint for the ancestor, i.e. that robust reconstructions are best achieved when available sequence data is exploited to the most full extent possible.

At the core of GRASP's approach is the POG data structure, originally proposed to facilitate multiple sequence alignment[17,18]. We developed bi-directional edge parsimony to directly use the POG data structure during inference and pinpoint likely phylogenetic positions for indel events once homologs are placed in an alignment; it effectively delineates sequence content at all internal nodes of a given phylogenetic tree, collectively tracing the evolutionary relationships between all sequences. As a consequence, evolutionary events are isolated to specific clades, and alignment ambiguities that are difficult to resolve at a single branch point can be disentangled across evolutionary time, here applied to enzyme families with members across different phyla.

ASR often requires judgements to be made as to whether ancestors contained an insertion present in only some of the multiple descendent branches of a phylogenetic tree. The presence of these events may pinpoint loop remodelling events[19] or other conformational diversity in the family. In turn, the evolution of conformational diversity may promote new functions[8].

In contrast to current approaches based on gapped sequence representations, POGs enable the identification of all supported indel histories across a reconstructed family. Hybrid ancestors represent a novel class of variant that can readily

be identified and resurrected through the partitioning of indel events onto individual edges within a POG. In GRASP, edges that are parsimonious but are not chosen to form the preferred ancestral sequence are visualised as alternative paths through the POG; multiple branch points can be inspected and contrasted at once. GRASP provides a framework to delete, reintroduce, or preemptively include indel variation that supports both exploring and creating new function.

The modular identification of alternative indels, each compatible with a given ancestor (or those near it) can be used to test alternative hypotheses about the true progression of events. In addition, combinations of compatible indels can be sampled in order to engineer novel sequences by including or removing blocks of sequence content. Using the analysed CYP2U data set, we showed that inclusion of these modular blocks allowed for increased thermal stability and altered substrate preference. We stress that we are not attributing the increased stability or interaction with a specific substrate solely to the identified insertion, but rather that we have identified blocks of content that are likely to be tolerated and which in turn affect the folding and function of these ancestral proteins. Due to the complex nature of protein folding, these blocks will not always behave in predictable ways and effects will depend on the ancestral sequence and sequence context into which they are being inserted. Strikingly, given the substantial impact that indel events are likely to have on any protein sequence, coupled with the divergence between CYP2U ancestral sequences chosen, the hybrid ancestors folded to form holoenzymes that are catalytically active when tested *in vitro* and are capable of interacting with the native human reductase.

The identification of modular insertions altered the substrate selectivity, not unlike how substitution variants identified by marginal reconstruction have in previous studies[7]. This study provides a proof-of-concept that indel histories can suggest a form of variation that protein engineers can use that is orthogonal to varying specific amino acids. We foresee this as being of practical use for (1) altering function through the addition and removal of discrete, evolutionarily-defined building blocks to engineer variants with altered catalytic and physical properties (e.g. thermal stability) and (2) exploring alternative ancestors where there is ambiguity in the true phylogenetic position of an indel.

Methods

The three main stages of GRASP are (1) to construct an indel history for every position in the alignment, (2) to infer character states for all positions not removed in each ancestor, and (3) to form a POG by linking positions inferred for each ancestor.

GRASP infers ancestor character states from a set of M input sequences $\mathbf{S} = \{S_j : j \in \mathbb{J}\}$ where $\mathbb{J} = \{1, 2, ..., M\}$; $\mathbf{S}$ has N aligned positions, indexed with $i \in \mathbb{I}$, where $\mathbb{I} = \{1, 2, ..., N\}$. In classical sequence alignments, positions without sequence content are padded, often shown as '$-$'; we use $\mathbb{I}^{(j)} \subseteq \mathbb{I}$ to index

positions in sequence j with actual sequence content, $S_{ji} = x$ where $x \in \mathcal{A}$ when $i \in \mathbb{I}^{(j)}$ and $\mathcal{A}$ is the set of amino acids. Later, it will be convenient to refer to the transpose of $\mathbb{I}^{(j)}$, namely $\mathbb{J}^{(i)}$ which indexes all sequences with content at position i.

Inference is based on a given phylogenetic tree T with a nominated root, that has $M-1$ branch points (if bifurcating, fewer when multifurcating) indexed by $\mathbb{K} = \{M+1, M+2, ..., 2M-1\}$; we designate the index $k = M+1$ for the root of the tree. The superset of extant and ancestor sequences (matched to POGs) is indexed by $\mathbb{Z} = \mathbb{J} \cup \mathbb{K}$. The topology of T defines parent-child relationships, $\mathbb{Z}^{(k)} \subseteq \mathbb{Z}$ indexes the ancestral descendants of an ancestor k; conversely, we define a function $\kappa(k') = k''$ to indicate that k'' is the direct ancestor of k', where $k' \in \mathbb{Z}^{(k'')}$.

Character states are inferred with an evolutionary model (in the form of an instantaneous rate matrix, indexed by $\mathcal{A}$); and maximum likelihood[20], by using a Bayesian network that shares the topology of the position-specific character tree, which is determined by parsimony.

Below, we first define key data structures, then we distinguish between (a) the handling of *where* homologous positions are placed relative to one another in the trace of ancestral sequences via POGs; and (b) the principles with which homologous positions in extant sequences are used to determine ancestral character states at branch points in the phylogenetic tree. The principles under (b) are unremarkable in themselves, but key benefits are achieved by using them in the ancestor POG from (a). For succinctness, we describe this procedure as it applies to a bifurcating tree, however the same principles seamlessly extend to multifurcating trees.

Representing sequence content as a partial order graph (POG)

A POG is a directed acyclic graph whose elements are ordered relative to other elements; a strict ordering is enforced *within* a subset of elements, but not always *between* subsets. When an order is imposed amongst elements the relationship must be reflexive, anti-symmetric, and transitive[10]. A growing body of work in sequence alignment has demonstrated the flexibility that POGs offer for detecting and representing homologous sequence elements during alignment[17,18,21]. We take advantage of the flexibility of POGs when projecting homologous elements back in time; they represent deletions and insertions by edges that exclude and include alternative character subsets, respectively, allowing for optional histories by offering multiple paths at ancestral branch points.

Formally, a POG is defined by a set of (up to) N nodes that are indexed by $i \in \mathbb{I}$. The indices are determined by performing a topological sort on the input POG (see below); this gives at least one linear and complete ordering (out of several possible). Nodes are connected by a set of directed edges, which is conveniently represented by a matrix $\mathbf{E}$, where $\mathbf{E}(a, b)$ is set to 1 if there is an edge from a to b, else 0. We introduce an extended index-set $\mathbb{I}*$ for rows and columns in $\mathbf{E}$, with 0 and $N+1$ to start and terminate the POG, so $a \in \mathbb{I}*$ and $b \in \mathbb{I}*$. We define

$next(\mathbf{E}, a) = \{b : \mathbf{E}(a, b) > 0\}$ and $prev(\mathbf{E}, b) = \{a : \mathbf{E}(a, b) > 0\}$ to refer to sets of nodes that occur after and before a node with a given index, respectively. $next(\mathbf{E}, 0)$ would thus give all possible start indices, and $prev(\mathbf{E}, N + 1)$ all terminating indices. Moreover, we define $path(\mathbf{E})$ to return all indices in $\mathbb{I}$ that can be accessed from 0, and $N + 1$, via recursive application of $next$ and $prev$.

We distinguish between three types of POGs, the first two are determined directly from $\mathbf{S}$, and the third by inference. All POGs share the node index $\mathbb{I}$, which allows character states to be mapped across extant sequences and ancestors (illustrated in Fig. 1).

- an "extant POG", is defined by a set of edges $\mathbf{E}^{(j)}$ specific to an extant sequence S_j, where $j \in \mathbb{J}$. $path(\mathbf{E}^{(j)})$ recovers the indices in $\mathbb{I}^{(j)}$; it forms a single path of "character" nodes $X_{ji} = S_{ji}$ where $i \in \mathbb{I}^{(j)}$.
- an "input POG", denoted $\mathbf{E}* = \sum_{j \in \mathbb{J}} \mathbf{E}^{(j)}$ represents the joint set of edges collected from extant sequences. The presence of an edge between a and b is indicated by $\mathbf{E}*(a, b) > 0$.
- an "ancestor POG", is inferred to have a set of edges $\mathbf{E}^{(k)}$ where $k \in \mathbb{K}$. It links a series of nodes Y_{ki} where $i \in \mathbb{I}^{(k)}$; each node either identifies a character state, or defines a probability distribution over character states; the latter is referred to as a "distribution" node. Once the POG for ancestor k is inferred, $path(\mathbf{E}^{(k)})$ recovers its valid indices $\mathbb{I}^{(k)}$.

Inference of ancestral states, insertions, and deletions

The phylogenetic tree with a nominated root and the collection of extant POGs serve as input to inference. GRASP supports two types of inference:

- Marginal reconstruction at a specified ancestral branch point in the phylogenetic tree; as a result of inference, the nominated ancestor POG will contain distribution nodes that represent the marginal distributions of character states.
- Joint reconstruction of all ancestral branch points; all ancestor POGs will contain character nodes that represent the most probable character state.

Inferring insertions and deletions at ancestor branch points

POG $\mathbf{E}^{(k)}$ at an ancestor k defines all possible paths that can form a valid sequence and therefore determines if a character state needs to be inferred at any given position. This subsection describes how $\mathbf{E}^{(k)}$ is determined, and $\mathbb{I}^{(k)}$ by implication.

GRASP considers all edges in $\mathbf{E}*$ and seeks to jointly identify the most parsimonious set of edges across all branches in the tree. To decompose this problem, GRASP scores edges *leaving* ($\delta = \texttt{OUT}$) and edges *entering* ($\delta = \texttt{IN}$) for a single position at a time. This process starts at the top-most branch point in the tree, and

by dynamic programming finds the edges at all descendant ancestors that imply the smallest cost across the tree.

Eq. 1 defines a score for each edge (between a and b, with direction δ) at a given ancestor k. The parsimony score of that edge depends on which edges are selected at its descendants: staying with the same edge is costless, changing to any other edge will cost 1. Except for the base case when the descendant is an extant sequence, the cost from the descendants are propagated recursively, and (all) edge choices that ultimately lead to the best parsimony score at the top-most branch point are recorded. The edge with the best score (i' in Eq. 1, relative to either a or b depending on δ) is assigned a score of 1 to indicate it was most parsimonious in that direction: $\mathbf{E}^{(k)}(i, i') = 1$ for both $\delta = \texttt{OUT}$ and $\delta = \texttt{IN}$, at any position $i \in \mathbb{I}$. (Note that $\delta = \texttt{OUT}$ references only one half of the matrix ($i < i'$) and $\delta = \texttt{IN}$ references the other ($i > i'$).)

$$\sigma(k, \delta, a, b) = \sum_{c}^{\mathbb{Z}^{(k)}} \begin{cases} \min_{i'}^{next(\mathbf{E}*,a)} \begin{cases} 0 & \text{if } b = i' \\ 1 & \text{otherwise} \end{cases} + \begin{cases} 0 & \text{if } c \in \mathbb{J} \\ \sigma(c, \delta, a, i') & \text{if } c \in \mathbb{K} \end{cases} & \text{if } \delta = \texttt{OUT} \\ \min_{i'}^{prev(\mathbf{E}*,b)} \begin{cases} 0 & \text{if } a = i' \\ 1 & \text{otherwise} \end{cases} + \begin{cases} 0 & \text{if } c \in \mathbb{J} \\ \sigma(c, \delta, i', b) & \text{if } c \in \mathbb{K} \end{cases} & \text{if } \delta = \texttt{IN} \end{cases} \tag{1}$$

For an ancestor k, an edge (a, b) is included if $\mathbf{E}^{(k)}(a, b) + \mathbf{E}^{(k)}(b, a) > 0$; if the sum is 2, it is bi-directionally parsimonious, which implies it is preferred when identifying ancestor sequences. There is no guarantee that there is a complete path through the POG where all edges are bi-directionally parsimonious, but in practice this turns out to be mostly the case.

Inferring the character state of ancestor nodes

For GRASP to infer character states and operate efficiently, we make several standard assumptions. First, each sequence position, i, can be modelled independently[20]. Second, we assume that character substitutions depend only on the state of the immediate ancestor[20]. Third, we assume that each position mutates at the same rate. Modelling variable rates across positions[22,23] compromises our ability to efficiently process large data sets, we presently opt not to do this. Instead, we leverage efficient procedures of graphical models for inference[14].

The topology of the phylogenetic tree maps to a character tree for each position, subject to the position $i \in \mathbb{I}^{(k)}$ in an ancestor k forming part of a valid sequence; for later, we define the transpose of that mapping as $k \in \mathbb{K}^{(i)}$, i.e. the subset of ancestors that have character content for a position i.

Each position-specific character tree maps to a directed Bayesian network, which is parameterised to reflect evolutionary distances (additively) at each branch, from the provided phylogenetic tree. The network is created with "observable" variables, instantiated to the characters in extant sequences $X_{ji} = x$. "Non-observable" variables in the Bayesian network correspond to the ancestors Y_{ki}

where $i \in \mathbb{I}^{(k)}$; how the character state or distribution for Y_{ki} is inferred is described below.

A Bayesian network node is a conditional probability $P(X_{ji}|Y_{\kappa(j)i}, d_j)$ or $P(Y_{ki}|Y_{\kappa(k)i}, d_k)$, for $j \in \mathbb{J}^{(i)}$ and $k \in \mathbb{K}^{(i)}$ and is parameterised by their respective distances (d_j or d_k; which refer to their closest ancestor branch point, $\kappa(j)$ or $\kappa(k)$, respectively).

The matrix of conditional probabilities is $e^{Q(d)}$ where Q is the instantaneous rate matrix given by the evolutionary model. GRASP supports all popular models[24,25,26,27]. Inference of the joint ancestral character state at a position i is then defined by:

$$P(\{Y_{ki} : k \in \mathbb{K}^{(i)}\}|\{X_{ji} : j \in \mathbb{J}^{(i)}\}, T) \propto \prod_{j \in \mathbb{J}^{(i)}} P(X_{ji}|Y_{\kappa(j)i}, d_j) \prod_{k \in \mathbb{K}^{(i)}} P(Y_{ki}|Y_{\kappa(k)i}, d_k) \tag{2}$$

where T is the tree with distances for all branches. The implementation uses an adaptation of variable elimination[14,28], which decomposes the inference into an efficient series of products given the hierarchical topology of the tree. Ancestral states are determined by the highest *joint probability* across all non-observed variables (all ancestors, all positions). From the above, GRASP is also capable of inferring the *marginal probability distribution* for each position in a given ancestor, by summing out all other non-observed variables. All inferences are exact (not approximated).

Identifying a single, preferred ancestor sequence

Not uncommonly, multiple indel histories are equally parsimonious, implying that several ancestor candidate sequences can be identified by traversing an ancestor POG; however, in some applications it is necessary to nominate a single sequence.

To determine a "preferred" path through an ancestor POG, we first define a subset of extant sequences $\mathbb{J}^{(k)}$ that are in the subtree under a given ancestor, k. To express preference between multiple edges, we calculate the proportion of extant sequences that contain a particular edge (see Eq. 3).

$$w_k(a, b) = \frac{\sum_{j \in \mathbb{J}^{(k)}} \begin{cases} 1 & \text{if } \mathbb{E}^{(j)}(a, b) = 1 \\ 0 & \text{otherwise} \end{cases}}{|\mathbb{J}^{(k)}|} \tag{3}$$

Identifying the preferred path

GRASP uses the A* algorithm[29] to determine the selection of edges in a POG that jointly minimise the cost, travelling from the N- to the C-terminus.

The cost assigned to an edge is given by Eq. 4:

$$\gamma_k(a, b) = (1 + (\eta_k(a, b) \cdot (1 - w_k(a, b)))) \cdot (b - a) \tag{4}$$

η is defined in Eq. 5 and imposes an absolute preference for bi-directionally parsimonious edges; a uni-directional edge is only chosen in the absence of bi-directional edges to complete the traversal. The exception is the edge to the first node, and the edge from the last node, where bi-directionality is disregarded. The impact of the weight is normalised by the number of positions skipped by a given edge, $b - a$. This ensures that each complete ancestral sequence is scored evenly, regardless of the number of edges it takes to form.

$$\eta_k(a,b) = \begin{cases} N & \text{if } \mathbf{E}^{(k)}(a,b) + \mathbf{E}^{(k)}(b,a) < 2,\ a \neq 0 \text{ and } b \neq N+1 \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

Access to tools and data

GRASP is freely accessible via a web server at `http://grasp.scmb.uq.edu.au`. The online service allows users to upload their own data sets and predict ancestors. The results are presented to allow exploration of ancestral POGs and their states via an interactive phylogenetic tree. Numerous other functions are available including annotation of trees with taxonomy and user specified terms, inspection of probability distributions for the identification of mutations for alternative ancestors, and sharing of entire reconstructions. A tutorial, user guide, and several example reconstructions are also available from the web site.

Data sets reported in the manuscript and a suite of tools to assist in the application of GRASP are available at `https://bodenlab.github.io/GRASP-suite`. In particular, a command-line version of the prediction method without visualisation features is available.

The implementation in Java and a web application are available from the same site. The software is available under the GNU General Public License v3.0.

GDH-GOx experimental methods

GDH-GOx ancestral inference

Starting from an aligned data set and phylogenetic tree previously established by Sützl et al.[11] for the GDH-GOx cluster, only the four major clades (GOx, GDH I, GDH II, and GDH III) were selected together with the second small GDH III clade. All sequences with >800 amino acids as well as manually selected sequences showing large insertions were removed from the selection, resulting in 399 sequences. This sequence selection was aligned by MAFFT v7.271 G-INS-i[30], the alignment trimmed for positions with >99% gaps by trimAl v1.2[31], and pruned using Gblocks 0.91b[32] with a less stringent block selection. The phylogenetic tree was inferred by PhyML[33] with default settings with SPR moves to optimise tree topology, Smart Model Selection (SMS), and aLRT SH-like branch support. The tree was rooted on the midpoint. Marginal reconstruction of ancestral nodes was

performed with the LG evolutionary rate model[27] after the N- and C-termini of the alignment had been trimmed.

GDH-GOx synthesis and cloning

The N- and C-terminal sequences not present in the ancestral sequences were replaced by the equivalent amino acid sequences of GOx from *Aspergillus niger*, 'MQTLLVSSLVVSLAAALPHYIRSNGIEASLLTDPKDVSGRT' and 'ASMQ', respectively. Resulting ancestral genes were ordered at BioCat GmbH, cloned into the expression vector pPICZ A together with an added polyhistidine tag (6 x His), and codon-optimised for *Komagataella phaffi* (formerly *Pichia pastoris*) expression. Constructs were linearized with *Pvu*II and transformed into *K. phaffii* via electroporation.

GDH-GOx expression

Ancestral and extant GOx and GDH genes were expressed in *K. phaffii* under the AOX1 promoter with methanol induction. Routine cultivations and selection of the transformed cells were done in liquid YPD medium supplemented with zeocin (100 mg/L) at 30 °C and 130 rpm. Expression was done in shake flasks at 30 °C and 130 rpm on modified BMMY medium (20 g/L peptone from casein, 10 g/L yeast extract, 100 mM potassium phosphate buffer pH 6.0, 10 g/L $(NH_4)_2SO_4$, 3.4 g/L yeast nitrogen base (without amino acids and $(NH_4)_2SO_4$), and 0.4 mg/L biotin) together with 12 g/L sorbitol and 2% methanol. After centrifugation at 6,000x*g* and 4 °C for 30 minutes, supernatants were loaded onto an equilibrated 5 mL HisTrap column (GE Healthcare) and washed with binding buffer (50 mM potassium phosphate buffer pH 6.5, 500 mM NaCl, and 20 mM imidazole). Proteins were eluted using a linear gradient of 50 mM potassium phosphate buffer pH 6.5 containing 500 mM NaCl and 500 mM imidazole. Manually collected fractions were concentrated and desalted (50 mM phosphate buffer pH 6.5) in Vivaspin 20 tubes (Sartorius) with 30,000 Da molecular mass cut-off.

GDH-GOx activity assays

Both GDH and GOx activity were measured spectrophotometrically at 30 °C on a UV/Vis spectrophotometer (Lambda 35, Perkin Elmer), using appropriately diluted enzyme solution, 20 mM D-glucose, and the respective electron acceptor in 50 mM potassium phosphate buffer pH 6.5. The electron acceptors 1,4-benzoquinone (BQ) and ferrocenium-hexafluorophosphate (FcPF6) were used at 0.5 and 0.2 mM, and their reduction was followed at 290 and 300 nm, respectively. Reduction of the electron acceptor oxygen was measured using the peroxidase-coupled ABTS assay[34], following the reduction of 0.1 mM 2,2'-azino-bis(3-ethylbenzothiazoline-6-sulphonic acid) (ABTS) at 420 nm.

GDH-GOx thermal stability assays

Thermal stability of GDH-GOx enzymes was assessed by differential scanning calorimetry conducted on a PEAQ-DSC automated instrument (Malvern Panalytical). All enzyme samples were diluted to 5 µM (~0.33 mg/ml) in 50 mM potassium phosphate buffer pH 6.5, and scanned from 20–90 °C with a scan rate of 60 °C/h and feedback set to high. Instrument blanks were recorded using buffer only and rescans were measured for all samples. Data analysis was performed by using the MicroCal PEAQ-DSC software V.1.22. The background signal was subtracted using rescans whenever applicable or buffer blanks otherwise, the baseline was fitted using the spline method, and peaks were fitted with a non-two-state model.

DHAD experimental methods

DHAD ancestral inference

A *minimal* 585-sequence set was created, consisting of members annotated with family "Ilvd/Edd" in UniProt, and excluding sequence fragments (as defined by UniProt). Most sequences were from Swiss-Prot, with several TrEMBL entries added due to function and structural data being available. A *baseline* data set of 1,612 sequences was created from the minimal data set, ensuring that 19 nominated enzymes with experimental data (functional and/or structural) were included, as well as members of their UniRef90 clusters. A *background* data set of 8,221 sequences included annotated members from the broadest assortment of species, only filtered to be non-redundant at 90% identity (using UniRef90). All three data sets were checked for the aligned location of two motifs (CDK and PCN/PGH/SAH with provision for a substitution) that are associated with the active site[35]; sequences that did not exhibit these motifs were removed.

The background data set was then repeatedly and independently sampled to extend the baseline data set to up to 9,112 sequences. At each size increment of 500 sequences an alignment was created using Clustal Omega[36], and a phylogenetic tree was inferred using FastTree[37] and rooted using phosphogluconate dehydratase as an outgroup. Despite differences in alignments and phylogenetic trees at each data size increment, we were able to map any ancestor in a smaller tree to an ancestor in a larger tree by maximising shared inclusions and exclusions of member proteins of the ancestral subtrees. Joint reconstruction was performed with the JTT evolutionary model[25].

DHAD synthesis and cloning

The inferred DHAD ancestral genes N1, N423, and N560 from the 585 data set, and the equivalent nodes N9, N1442, and N1443 from the 9,112 data set were optimised for *E. coli* expression and synthesised by Twist Bioscience and ATG:biosynthetics

GmbH, respectively. After amplification, the purified DNA fragments were digested with *Sap*I, followed by ligation into a modified pET26 vector (p7XNH3)[38].

DHAD expression

Expression of the inferred DHAD and the *Pi*DHAD genes was performed in shaking flasks. *E. coli* BL21 (DE3) cells transformed with the p7XNH3 plasmid and the appropriate inserted gene fragment were grown as an overnight pre-culture in Lysogeny Broth supplemented with kanamycin (100 µg/ml), and then 1:50 inoculated into auto-induction ZP-5052 medium[39] supplemented with kanamycin (100 µg/ml). These cultures were incubated at 90 rpm and 37 °C for 3 hours and then overnight at 18 °C in a horizontal orbital shaking incubator. Cells were disrupted by sonication in binding buffer (50 mM potassium phosphate buffer, 500 mM NaCl, 10% glycerol, and 20 mM imidazole) at pH 8.0. Cell debris was pelleted by centrifugation. Proteins were purified using an ÄKTA Purifier FPLC system and a HisTrap HP Nickel column (GE Healthcare). Filtered samples were loaded onto the column and washed with binding buffer. The His-tagged proteins were then eluted with elution buffer (50 mM potassium phosphate buffer, 500 mM NaCl, 10% glycerol, and 500 mM imidazole) at pH 8.0. Desalting of the enzymes was carried out using HEPES buffer pH 7.0.

DHAD activity assays

DHAD activity was analysed by HPLC of an assay mixture containing the respective DHAD, 25 mM HEPES buffer pH 7.0, 5 mM $MgSO_4$, and 25 mM of sodium D-gluconate, and incubated at 30 °C. Samples were taken every few hours for 3 days. The enzyme was removed by ultrafiltration (PES 10 kDa MWCO, VWR) and the samples were stored at -20 °C until analysed by HPLC. HPLC measurements were performed on an Ultimate-3000 HPLC system (Dionex), equipped with an auto-sampler and diode-array detector. D-gluconate and products were separated by using a Metrosep A supp10-250/40 column (250 mm, particle size 4.6 µm, Metrohm) at 65 °C by isocratic elution with 12 mM ammonium bicarbonate at pH 10.02, followed by a washing step with 30 mM sodium carbonate at pH 10.4 and a flow rate of 0.2 ml/min. Each sample injection volume was 10 µl. System peak calibration was performed using external standards of the known compounds.

CYP2U experimental methods

CYP2U ancestral inference

Five candidate CYP2U proteins were chosen, one each from *Andrias davidianus, Python bivittatus, Marmota marmota, Poecilia reticulata*, and *Amazona aestiva*. A pBLAST search of each of the candidates excluding hits from plants (taxonomic

id:3193) or fungi (taxonomic id:4751) was conducted (E-Value = 0.00001). Sequences from the pBLAST search were retained if they had at least 55% sequence identity to the original candidate sequence. This procedure was also repeated retaining sequences with at least 50% sequence identity, however, the additional sequences from this lower bound were all removed at later stages of curation, indicating that 55% was an appropriate level of identity. Sequences from the pBLAST searches were collated and identical sequences were removed. Sequences were excluded if they were below 400 amino acids in length or if they contained unidentified amino acids in their sequences. Sequences were aligned using MAFFT (L-INS-i) with default parameters[30]. Removal of sequences with indel events over 20 amino acids (suggestive of incorrect annotation of splice sites) was completed in an iterative manner by first identifying which sequence had the longest indel over 20 amino acids, removing it and realigning the remaining sequences, and then continuing until no sequence had an indel over 20 amino acids. Sequences were manually inspected and any sequences with apparent frameshift mutations were removed. Sequences were mapped back to their exon structure and removed if they had more than two exons difference to the accepted number of five exons for CYP2U sequences. Sequences missing the conserved cysteine residue characteristic of cytochrome P450 enzymes were removed. Similar procedures were used to generate the CYP2R and CYP2D families. Phylogenetic trees were inferred using RAxML[40]. A CYP2R *Latimeria chalumnae* sequence (XP_005989762.1) was manually shifted on the phylogenetic tree to better represent the known phylogeny[41], while retaining each sequence's overall evolutionary distance to the root. Joint reconstruction was performed with the JTT evolutionary rate model[25].

CYP2U synthesis and cloning

The amino acid sequences of CYP2U ancestors were inferred starting from the conserved PPGP motif, which signifies the end of the transmembrane domain. For expression of the resurrected ancestors in bacteria, this region was replaced with an N-terminal sequence (MAKKTSSKGKL) that is known to improve expression yields of microsomal P450s in bacteria[42] and had been used to express human CYP2U1 in *E. coli*[43]. To enable purification, a flexible ST linker followed by a polyhistidine tag (6 x His) was added to the C-terminus of the sequences. All ancestor sequences were codon-optimised for *E. coli* expression, and the N-termini were optimised initially using mRNA optimiser[44] and subsequently manually until the free energy was less than -15 kJ/mol. The genes were synthesised as GeneStrings (GeneArt, Invitrogen) designed with 60 bp 5' and 3' end extensions complementary to the WW vector, cloned by Gibson assembly, and then sequence-verified by dideoxy sequencing (Australian Genome Research Facility). Correct inserts were subcloned into a bicistronic pCW vector upstream of the open reading frame for the human cytochrome P450 reductase (hCPR) using the *Nde*I and *Xba*I sites.

CYP2U expression

DH5a F' IQ *E. coli* cells carrying the pGro7 plasmid were transformed with pCW vectors containing the relevant P450 and CPR genes or the empty vector ("pCW controls"), and selected using chloramphenicol (20 µg/ml) and ampicillin (100 µg/ml). Single colonies were used to inoculate overnight cultures in Lysogeny Broth with antibiotics. Batch cultures were grown at 25 °C, 180 rpm in 500 ml flasks containing 50 ml Terrific Broth supplemented with trace elements, 1 mM thiamine, and antibiotics. Cultures were induced after 5 hours with 1 mM IPTG and 4 mg/ml L-arabinose, and supplemented with 500 mM delta-aminolaevulinic acid. Cultures were grown for a further 43 hours before harvesting by centrifugation at 6,000x*g* for 10 minutes. *E. coli* pellets were weighed and resuspended in 2 ml/g (wet weight) sonication buffer (100 mM potassium phosphate buffer pH 7.4, 20% (w/v) glycerol, 6 mM magnesium acetate, 1 mM PMSF, and protease inhibitor cocktail (Sigma-Aldrich)). Cells were lysed using a Constant Systems OneShot cell disruptor followed by centrifugation at 10,000x*g* for 20 minutes. The supernatant was centrifuged at 180,000x*g* for 1 hour and the pellet was resuspended in TES (100 mM Tris acetate, 500 mM sucrose, and 0.5 mM EDTA pH 7.6) or the relevant solubilisation buffer using a Potter-Elvehjem homogeniser. The P450 concentration was determined in intact cells and membranes using Fe(II).CO vs. Fe(II) difference spectroscopy[45].

CYP2U activity assays

P450 (0.02 μM), added in membranes prepared from bacteria coexpressing hCPR, was premixed with 50 μM luciferin CEE or luciferin ME-EGE (Promega) in 100 mM potassium phosphate pH 7.4, and incubated at 37 °C for 10 minutes. Reactions were initiated by addition of the NADPH-regenerating system (NGS; 0.25 mM $NADP^+$, 10 mM glucose-6-phosphate, and 0.5 U/ml glucose-6-phosphate dehydrogenase), and incubated with shaking at 37 °C for 30 minutes. An equal volume of the luciferin detection reagent was added, and reactions were incubated for a further 20 minutes at room temperature. Luminescence was measured using a CLARIOstar multimodal plate reader (BMG Labtech).

CYP2U thermal stability assays

Ancestors were expressed in *E. coli* as described above and cell pellets were resuspended in whole cell spectral assay buffer (WCAB; 100 mM potassium phosphate, 20 mM D-glucose, and 6 mM magnesium acetate pH 7.4) to one eighth of the original culture volume. The resuspended cultures, distributed into tubes in 200 μL volumes, were incubated at a range of temperatures (25-80 °C, in 5 °C increments) for 60 minutes, followed by a 5 minute recovery at 4 °C and equilibration at 25 °C. The remaining P450 content was measured in intact cells using the method of Johnston et al.[45]. The proportion of total P450 content compared to the unheated

control (25 °C) was plotted against temperature and the $^{60}T_{50}$ value was calculated by fitting the data to a variable slope (4-parameter) dose response curve in GraphPad Prism 8.0.

KARI experimental methods

KARI ancestral inference

We created two separate data sets representing KARI class I and class II, respectively. The sequence alignment for class II was taken directly from Gumulya et al.[7] and used to compare tools. Class I sequences were compiled by searching for both reviewed and unreviewed proteins in UniProt, designated as bacterial and belonging to the family (26,485 sequences). We removed all fragments and sequences above the length of 400 to exclude obvious cases of class II enzymes. The sequence set was redundancy-reduced with CD-HIT at 99%[46], resulting in 11,920 sequences, from which 57 sequences were manually removed by observing a C-terminal knotted domain, indicative of class II. After aligning all sequences with Clustal Omega[36], the Shannon entropy of gap vs. character content was determined for all columns and sequences with high entropy over consecutive columns were removed, resulting in a final set of 11,756 sequences. Phylogenetic tree inference was carried out using FastTree[37].

In contrast to the DHAD data sets, the KARI class I data sets were created by decreasing their size from 11,756 via 10 regular decrements reaching a minimum representation of 1,176 sequences. For each subset, the alignment was recalculated independently. For each alignment, a new tree was calculated, and rooted by using KARI sequences in Aquificae and Thermotogae as an outgroup. For each subset, we computed reconstructions for 50 randomly chosen ancestor nodes (mapped between each subset as described for the DHAD data sets). Joint reconstruction was performed with the JTT evolutionary model[25].

GRASP, FastML, and PAML comparison method

The following procedure was used to evaluate each of the tools: (1) the input multiple sequence alignment was randomly divided into G groups of alignments with approximately equal numbers of sequences, where $G \in \{5, 10, 20\}$; (2) for each sub-alignment, the input phylogenetic tree constructed from the full alignment was pruned to remove sequences not in the sub-alignment. To represent the same principal ancestor across all groups, as well as to maintain a valid tree, branches in the original phylogenetic tree with removed sequences were collapsed and branch distances added together; (3) sub-alignments and corresponding pruned trees were therefore pared-down representatives of the same family and used as input to each of the ASR tools; (4) the process was repeated until 20 ancestral sequences had been generated for each configuration, e.g., when $G = 5$ the process is repeated

four times. This procedure therefore results in sequences that belong to multiple groups across replicates for $G = 5$ and $G = 10$.

The JTT evolutionary rate model[25] was used for all inferences and variable rates were calculated from a discrete gamma distribution with eight categories. To remove confounding effects of different strategies for dealing with gaps, we removed any column that contained a deletion, leaving 455 and 242 columns in the KARI and CYP2 multiple sequence alignments, respectively.

Acknowledgements

We thank Broder Rühmann who helped prepare data for the manuscript. This work has been supported by the Australian Research Council (ARC) Discovery Project grants 160100865 and 120101772. BK is an ARC Laureate Fellow (FL180100109).

Author Contributions

Study design: MB, EMJG, GF; Experimental design and work: EMJG, VS, DH, CMR, LS, SB, JC; Software and algorithm development: MB, AM, MLL, GF, JZ, AE, BB, RN; Supporting scientific advice: BR, BK, REST, RTB, LG, GS, JC, YG; Manuscript: GF, MB, AM, CMR, EMJG, LS, SB, BK, RTB, LG, GS, BR, DH, VS.

Competing Interests Statement

The authors declare that they have no competing financial interests.

References

[1] Land, M. *et al.* Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics* **15**, 141–161 (2015).

[2] Lewin, H. A. *et al.* Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences* **115**, 4325–4333 (2018).

[3] Gregory, A. C. *et al.* Marine DNA viral macro - and microdiversity from pole to pole. *Cell* **177**, 1109–1123.e14 (2019).

[4] Hochberg, G. K. A. & Thornton, J. W. Reconstructing ancient proteins to understand the causes of structure and function. *Annual Review of Biophysics* **46**, 247–269 (2017).

[5] Bar-Rogovsky, H. *et al.* Assessing the prediction fidelity of ancestral reconstruction by a library approach. *Protein Engineering, Design & Selection* **28**, 507–518 (2015).

[6] Sugrue, E., Scott, C. & Jackson, C. J. Constrained evolution of a bispecific enzyme: Lessons for biocatalyst design. *Organic & Biomolecular Chemistry* **15**, 937–946 (2017).

[7] Gumulya, Y. *et al.* Engineering highly functional thermostable proteins using ancestral sequence reconstruction. *Nature Catalysis* **1**, 878–888 (2018).

[8] Trudeau, D. L. & Tawfik, D. S. Protein engineers turned evolutionists–the quest for the optimal starting point. *Current Opinion in Biotechnology* **60**, 46–52 (2019).

[9] Garcia, A. K. & Kaçar, B. How to resurrect ancestral proteins as proxies for ancient biogeochemistry. *Free Radical Biology and Medicine* **140**, 260–269 (2019).

[10] Lee, C., Grasso, C. & Sharlow, M. F. Multiple sequence alignment using partial order graphs. *Bioinformatics* **18**, 452–464 (2002).

[11] Sützl, L., Foley, G., Gillam, E. M. J., Bodén, M. & Haltrich, D. The GMC superfamily of oxidoreductases revisited: Analysis and evolution of fungal GMC oxidoreductases. *Biotechnology for Biofuels* **12**, 118 (2019).

[12] Kirischian, N., McArthur, A. G., Jesuthasan, C., Krattenmacher, B. & Wilson, J. Y. Phylogenetic and functional analysis of the vertebrate cytochrome p450 2 family. *Journal of Molecular Evolution* **72**, 56–71 (2011).

[13] Gao, H. *et al.* Function and maturation of the Fe-S center in dihydroxyacid dehydratase from Arabidopsis. *The Journal of Biological Chemistry* **293**, 4422–4433 (2018).

[14] Koller, D. & Friedman, N. *Probabilistic Graphical Models* (The MIT Press, 2009).

[15] Ashkenazy, H. *et al.* FastML: A web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Research* **40**, W580–4 (2012).

[16] Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**, 1586–1591 (2007).

[17] Grasso, C. & Lee, C. Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. *Bioinformatics* **20**, 1546–1556 (2004).

[18] Löytynoja, A. & Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences* **102**, 10557–10562 (2005).

[19] Afriat-Jurnou, L., Jackson, C. J. & Tawfik, D. S. Reconstructing a missing link in the evolution of a recently diverged phosphotriesterase by active-site loop remodeling. *Biochemistry* **51**, 6047–6055 (2012).

[20] Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376 (1981).

[21] Löytynoja, A., Vilella, A. J. & Goldman, N. Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. *Bioinformatics* **28**, 1684–1691 (2012).

[22] Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution* **39**, 306–314 (1994).

[23] Felsenstein, J. Taking variation of evolutionary rates between sites into account in inferring phylogenies. *Journal of Molecular Evolution* **53**, 447–455 (2001).

[24] Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure* 345–352 (1978).

[25] Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* **8**, 275–282 (1992).

[26] Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution* **18**, 691–699 (2001).

[27] Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Molecular Biology and Evolution* **25**, 1307–1320 (2008).

[28] Dechter, R. Bucket elimination: A unifying framework for reasoning. *Artificial Intelligence* **113**, 41–85 (1999).

[29] Hart, P. E., Nilsson, N. J. & Raphael, B. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics* **4**, 100–107 (1968).

[30] Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* **30**, 772–780 (2013).

[31] Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

[32] Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* **56**, 564–577 (2007).

[33] Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology* **59**, 307–321 (2010).

[34] Spadiut, O. *et al.* Improving thermostability and catalytic activity of pyranose 2-oxidase from Trametes multicolor by rational and semi-rational design. *The FEBS Journal* **276**, 776–792 (2009).

[35] Rahman, M. M. *et al.* The crystal structure of a bacterial L-arabinonate dehydratase contains a [2Fe-2S] cluster. *ACS Chemical Biology* **12**, 1919–1927 (2017).

[36] Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Science* **27**, 135–145 (2018).

[37] Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution* **26**, 1641–1650 (2009).

[38] Geertsma, E. R. & Dutzler, R. A versatile and efficient high-throughput cloning tool for structural biology. *Biochemistry* **50**, 3272–3278 (2011).

[39] Studier, F. W. Protein production by auto-induction in high density shaking cultures. *Protein Expression and Purification* **41**, 207–234 (2005).

[40] Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

[41] Amemiya, C. T. *et al.* Analysis of the African coelacanth genome sheds light on tetrapod evolution. *Nature* **496**, 311–316 (2013).

[42] von Wachenfeldt, C., Richardson, T. H., Cosme, J. & Johnson, E. F. Microsomal P450 2C3 is expressed as a soluble dimer in Escherichia coli following modifications of its N-terminus. *Archives of Biochemistry and Biophysics* **339**, 107–114 (1997).

[43] Siller, M. *et al.* Oxidation of endogenous N-arachidonoylserotonin by human cytochrome P450 2U1. *The Journal of Biological Chemistry* **289**, 10476–10487 (2014).

[44] Gaspar, P., Moura, G., Santos, M. A. S. & Oliveira, J. L. mRNA secondary structure optimization using a correlated stem–loop prediction. *Nucleic Acids Research* **41**, e73–e73 (2013).

[45] 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. *Journal of Biomolecular Screening* **13**, 135–141 (2008).

[46] Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682 (2010).

Supplementary Material

Thermal transitions [°C]

	Tm1	Tm2
*An*GOx	58.2	63.1
N320	67.4	70.0
N320 Y244E	71.0	73.9

Table 1: Comparison of thermal transitions from differential scanning calorimetry of an extant glucose oxidase from *Aspergillus niger*, the ancestor inferred at node N320, and the ancestor inferred at node N320 with a single amino acid change based on marginal distributions.

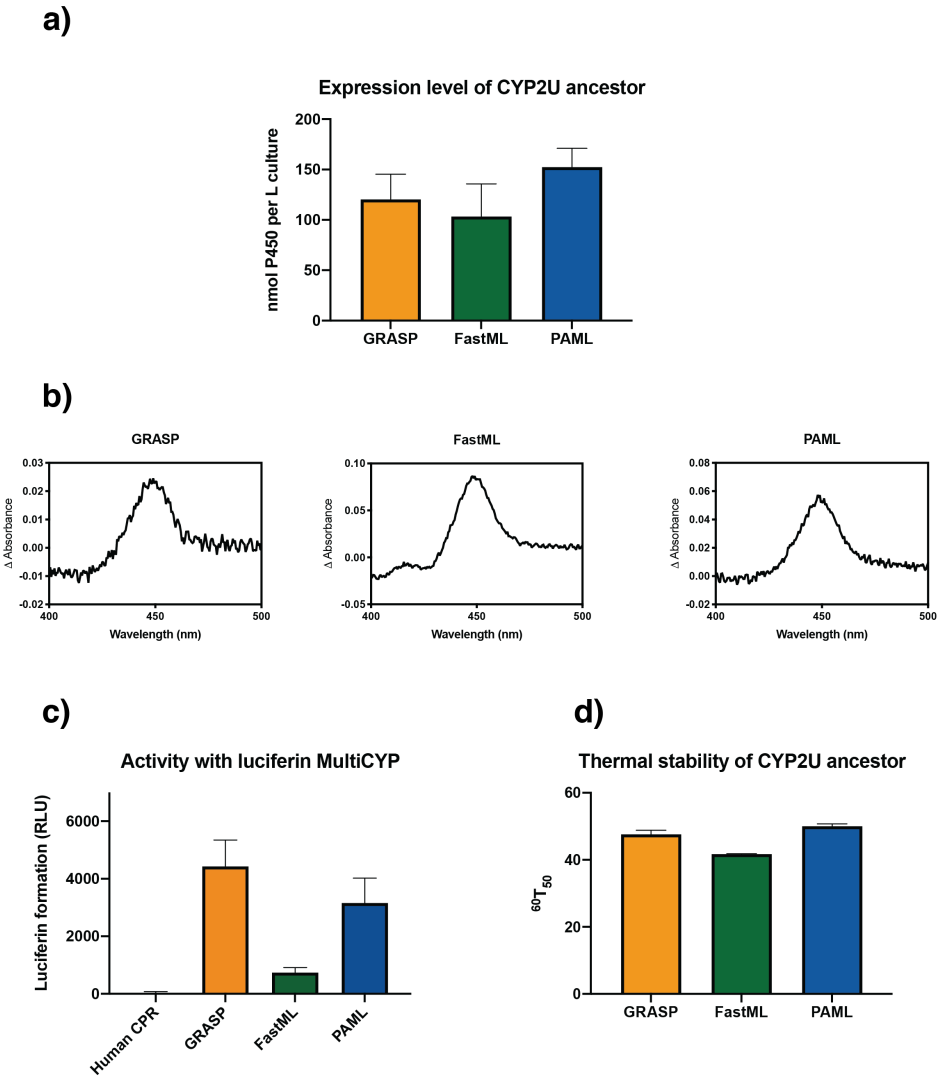

Figure 1: Comparison between ancestors generated using GRASP, FastML, and PAML. **a**, Expression level of CYP2U ancestors in *E. coli* cultures quantified using Fe(II) vs. Fe(II).CO difference spectroscopy. Data are means +/- SEM, $N = 3$. **b**, Fe(II) vs. Fe(II).CO difference spectra for ancestors generated using GRASP, FastML, and PAML in *E. coli* membranes. **c**, Turnover of luciferin MultiCYP by CYP2U ancestors in *E. coli* membranes, also containing human CPR, after 30 minutes at 37 °C. Data are means +/- SEM, $N = 3$. **d**, Comparison of T_{50} values after a 60 minute incubation at a range of temperatures (25-80 °C) for ancestors generated using GRASP, FastML, and PAML. Data are means +/- SEM, $N = 2$.

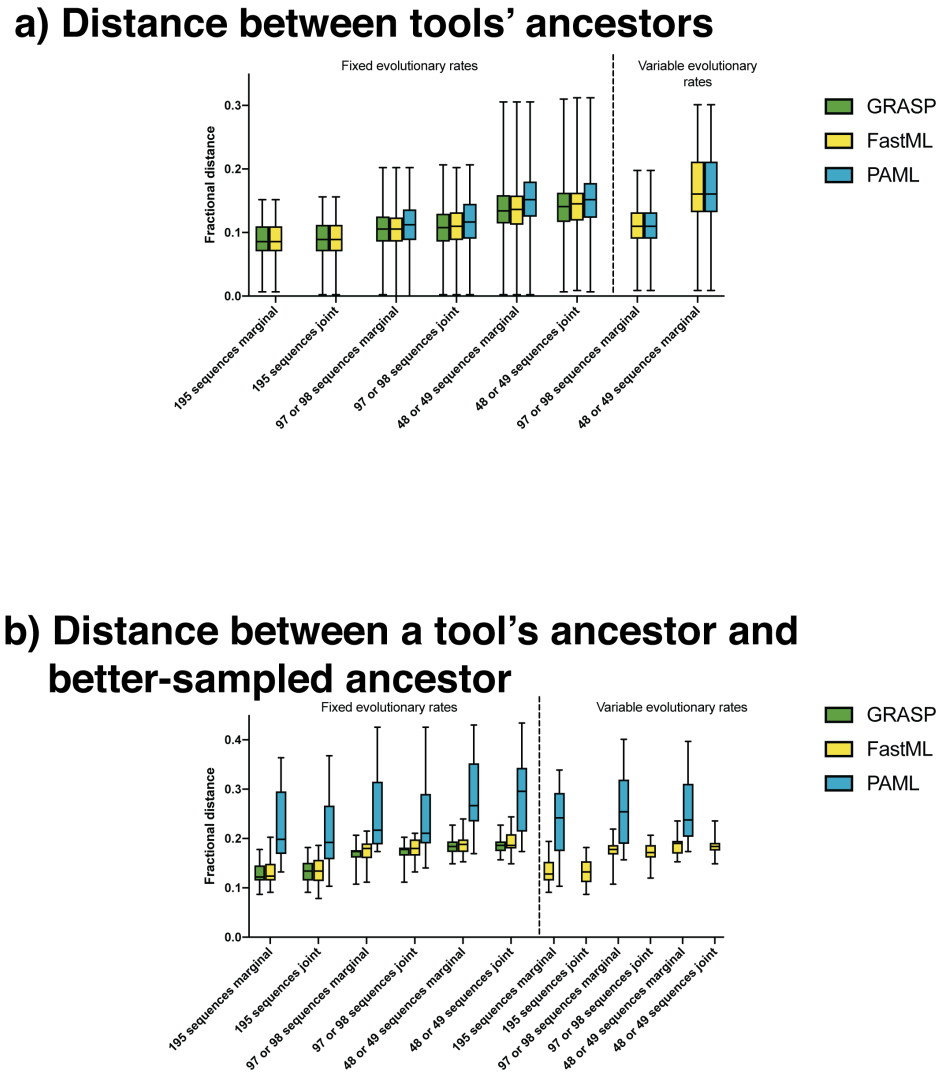

Figure 2: Tool comparison on CYP2 data. **a**, Average fractional distance between tools, calculated as pairwise fractional distances for each ancestral prediction for a given tool against all other ancestral predictions of other tools at 5 groups of 195 sequences, 10 groups of 97 or 98 sequences, and 20 groups of 48 or 49 sequences. Parameter combinations are joint and marginal reconstruction; and fixed or variable evolutionary rates (FastML and PAML only). **b**, Average fractional distance between a better-sampled ancestor inferred by GRASP using 975 sequences and each tool / parameter combination at 5, 10, and 20 groups.

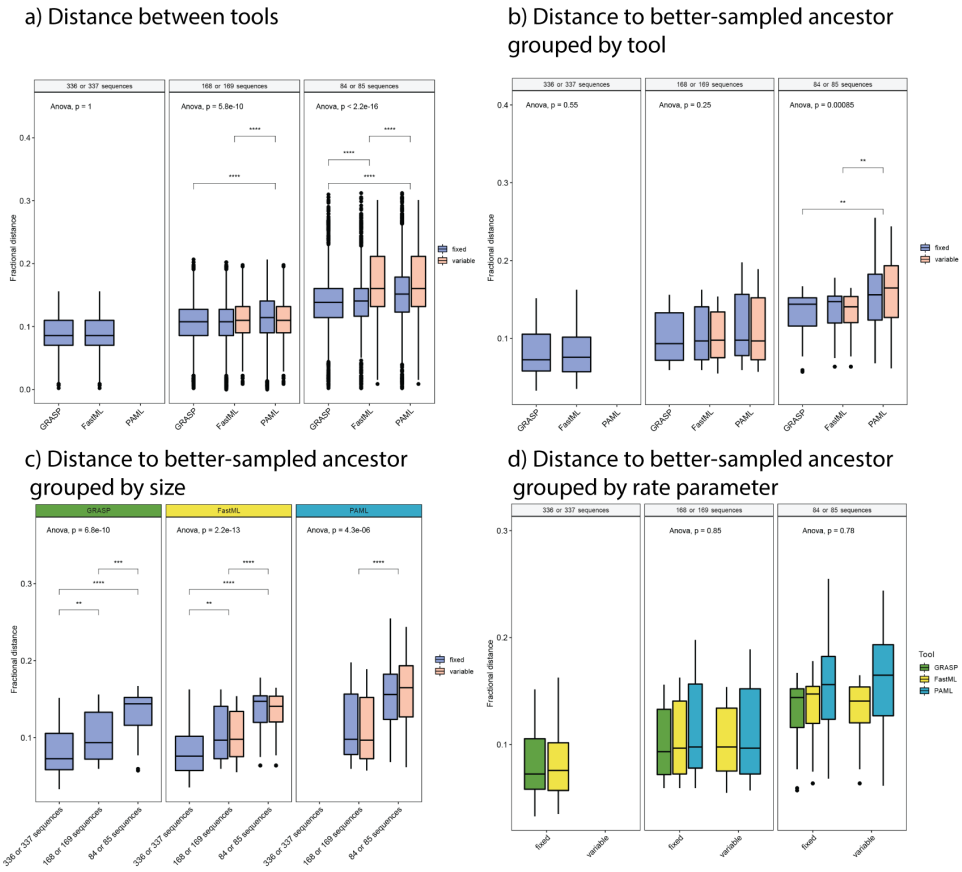

Figure 3: Statistical evaluation of determinants of ancestor prediction performance using 1,682 KARI sequences. **a**, Between tool distances grouped by tool within data set size. **b**, Distance to better-sampled ancestor grouped by tool within data set size. **c**, Distance to better-sampled ancestor grouped by size within tool. **d**, Distance to better-sampled ancestor grouped by rate parameter within data set size. PAML was excluded for the largest data set size; variable rates were not used for the largest data set size. All p-values were determined by a two-tailed Student's t-test. Only significant comparisons are shown (* means $p < 0.05$, ** means $p < 0.01$, *** means $p < 0.001$, **** means at limits of precision of test). All parameter settings are from Fig. 2.

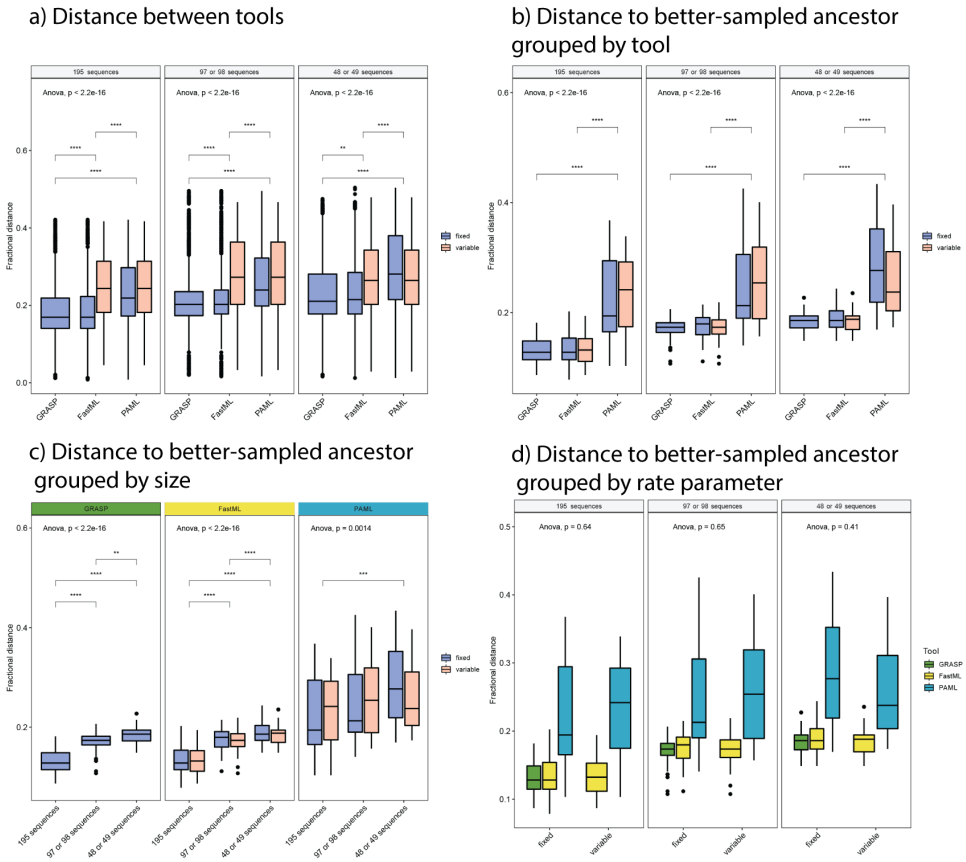

Figure 4: Statistical evaluation of determinants of ancestor prediction performance using 975 CYP2 sequences. **a**, Between tool distances grouped by tool within data set size. **b**, Distance to better-sampled ancestor grouped by tool within data set size. **c**, Distance to better-sampled ancestor grouped by size within tool. **d**, Distance to better-sampled ancestor grouped by rate parameter within data set size. All p-values were determined by a two-tailed Student's t-test. Only significant comparisons are shown (* means $p < 0.05$, ** means $p < 0.01$, *** means $p < 0.001$, **** means at limits of precision of test). All parameter settings are from Supplementary Fig. 2.

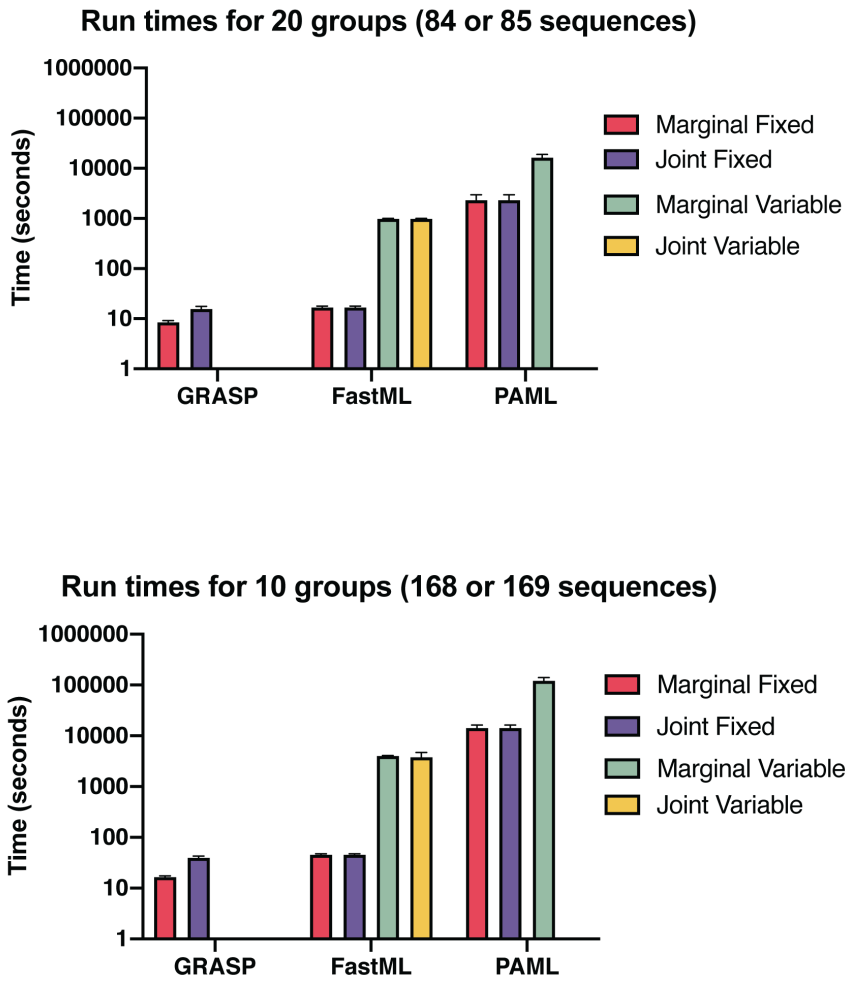

Figure 5: Run times of GRASP, FastML, and PAML at different parameter combinations and group sizes on the KARI data set. Parameter combinations are joint and marginal reconstruction; and fixed or variable evolutionary rates (FastML and PAML only).

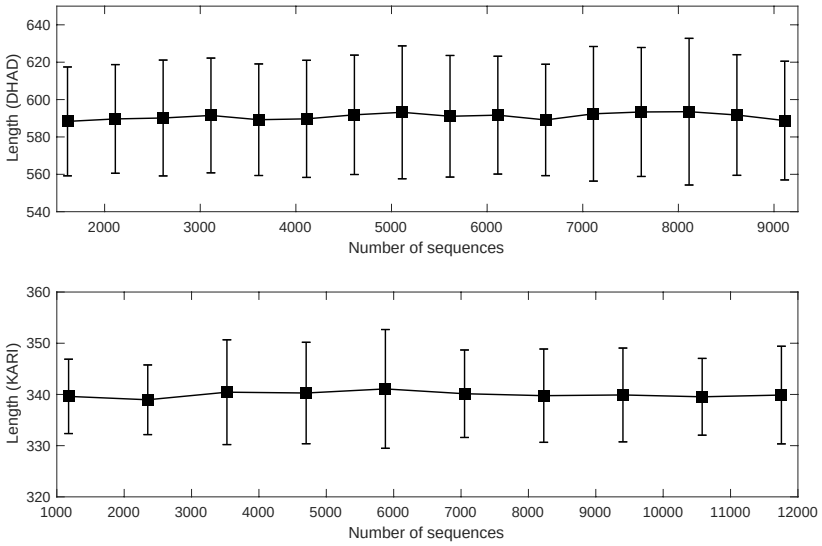

Figure 6: Predicted ancestor sequence lengths are unaffected by size of reconstruction. Mean and standard deviation of the lengths of 50 ancestor sequences mapped are plotted for different reconstructions and data set sizes for DHAD and KARI.

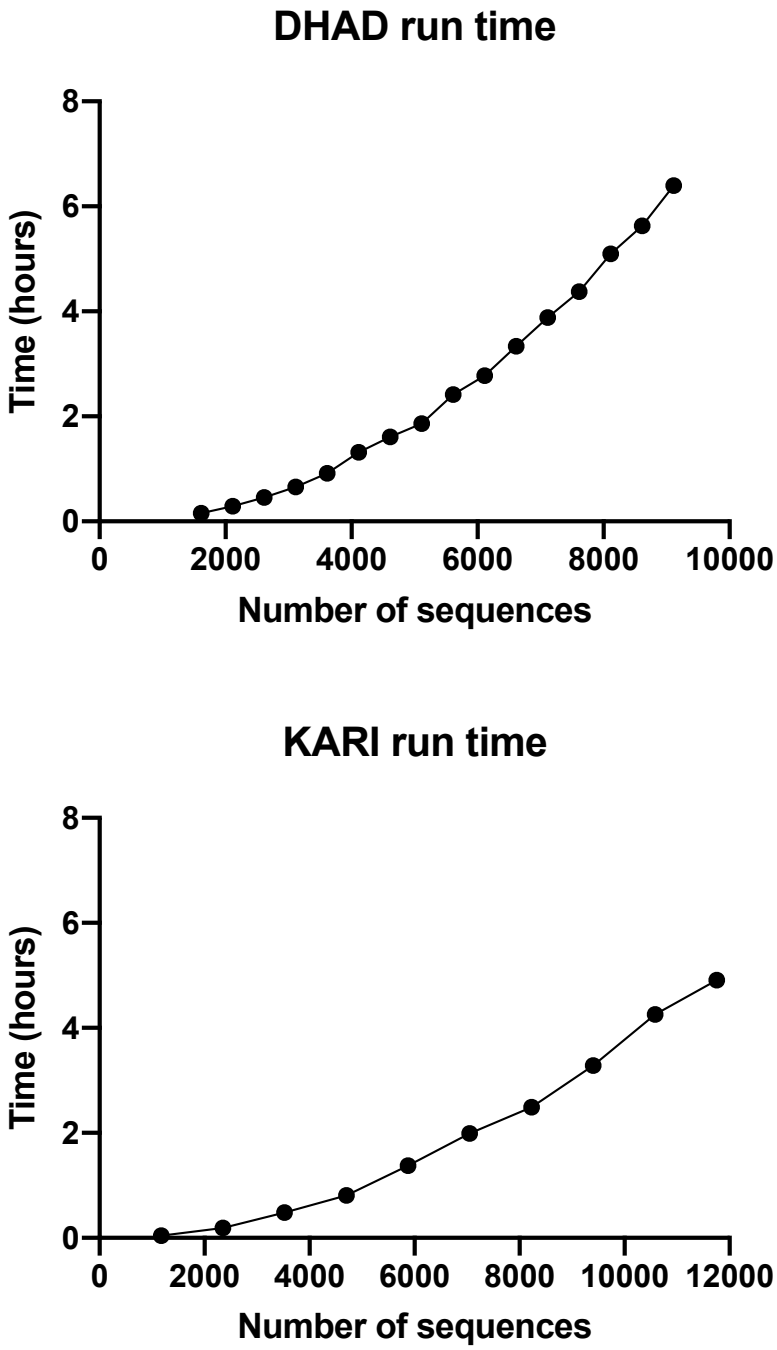

Figure 7: Run times for the DHAD and KARI enzyme families as data set size increases. Reconstructions were performed using GRASP running on 64 GB RAM, 5 threads on 2x 2.6 GHz 14C Xeon VM.

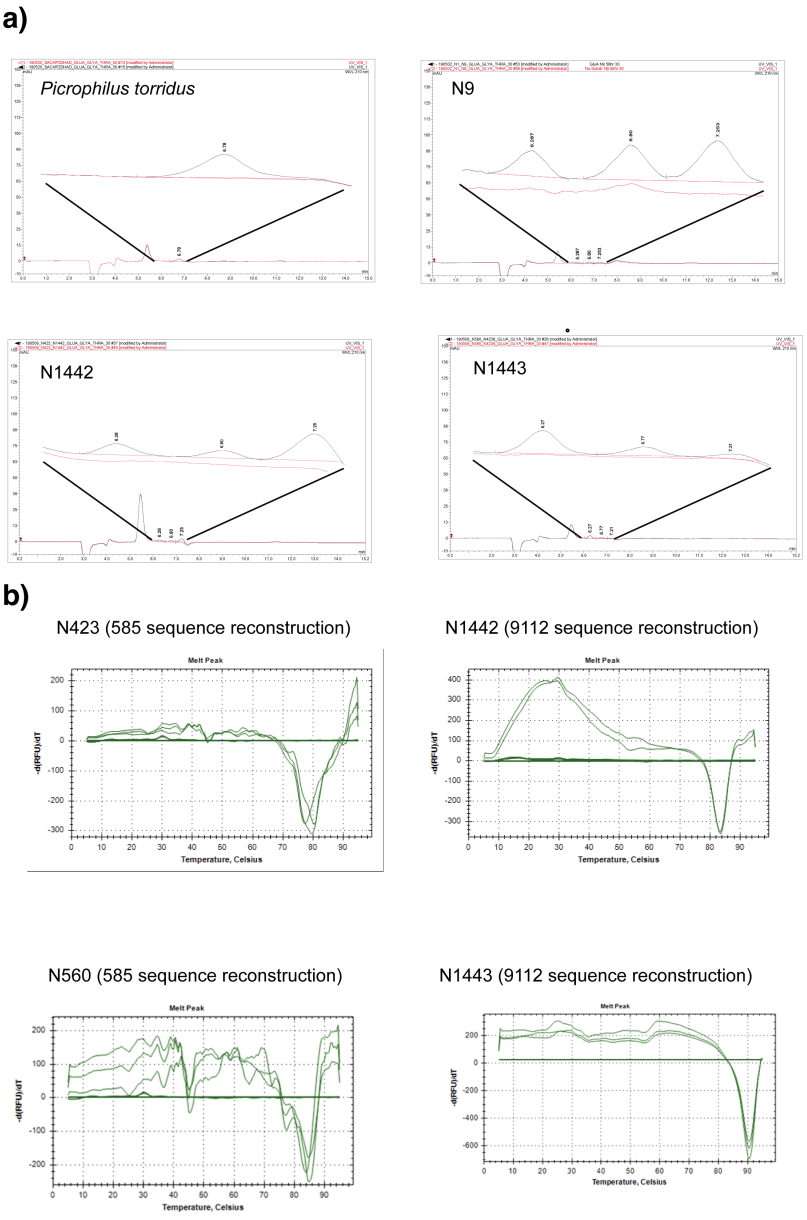

Figure 8: **a**, Chromatogram assays of an extant DHAD showing a typical profile peak when incubated with D-gluconate, and the same peak appearing in assays of ancestral DHAD proteins N9, N1442, and N1443 as inferred in the 9,112 DHAD data set. **b**, Thermal shift assays showing increase in temperature between equivalent ancestral nodes N423 (585 data set size) and N1442 (9,112 data set size), and equivalent ancestral nodes N560 (585 data set size) and N1443 (9,112 data set size).

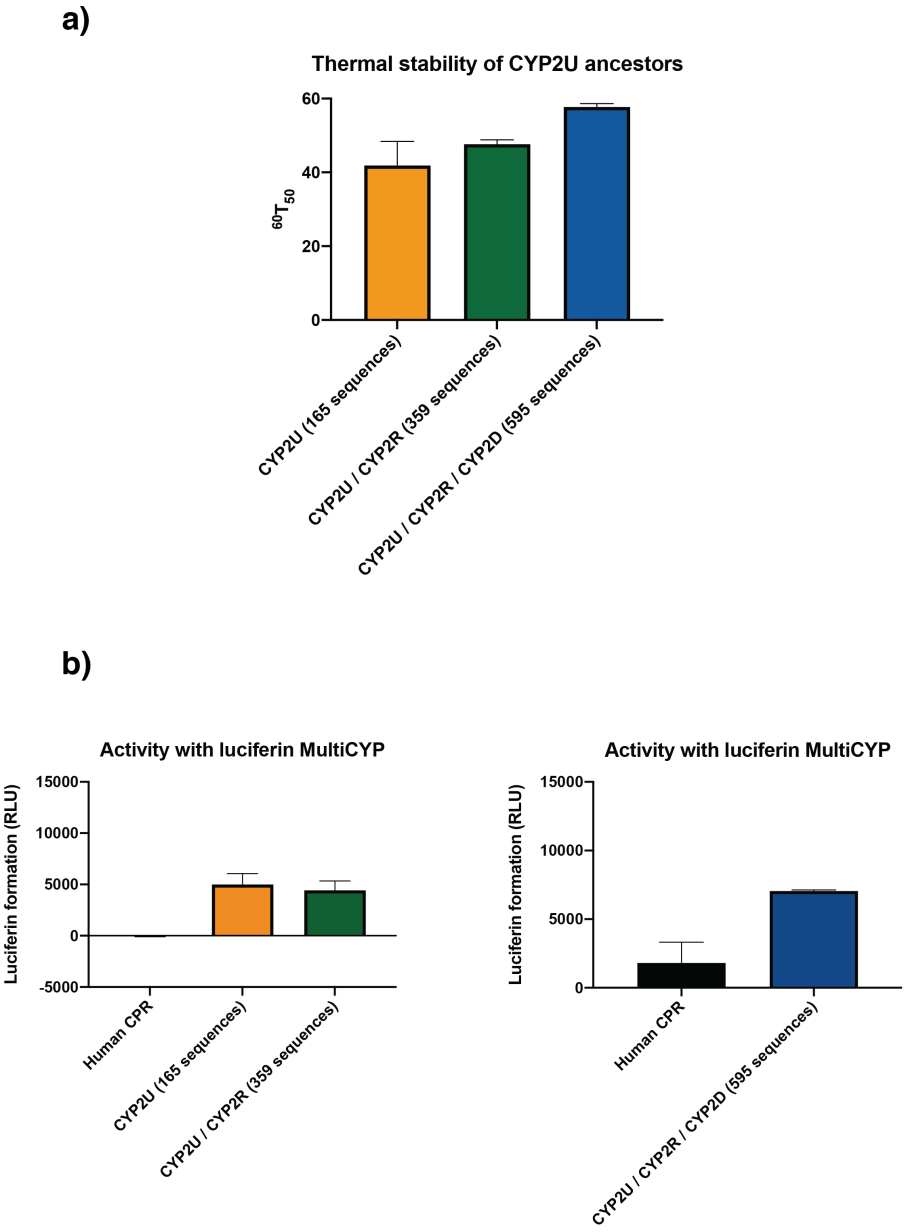

Figure 9: Thermal stability and activity for the CYP2U, CYP2U/CYP2R, and CYP2U/CYP2R/CYP2D ancestors with luciferin MultiCYP. **a**, Comparison of T_{50} values after a 60 minute incubation at a range of temperatures (25-80 °C). Data are means +/- SEM, $N = 2$. **b**, Turnover of luciferin MultiCYP by CYP2U ancestors in *E. coli* membranes, also containing human CPR, after 30 minutes at 37 °C. Data are means +/- SEM, $N = 3$.

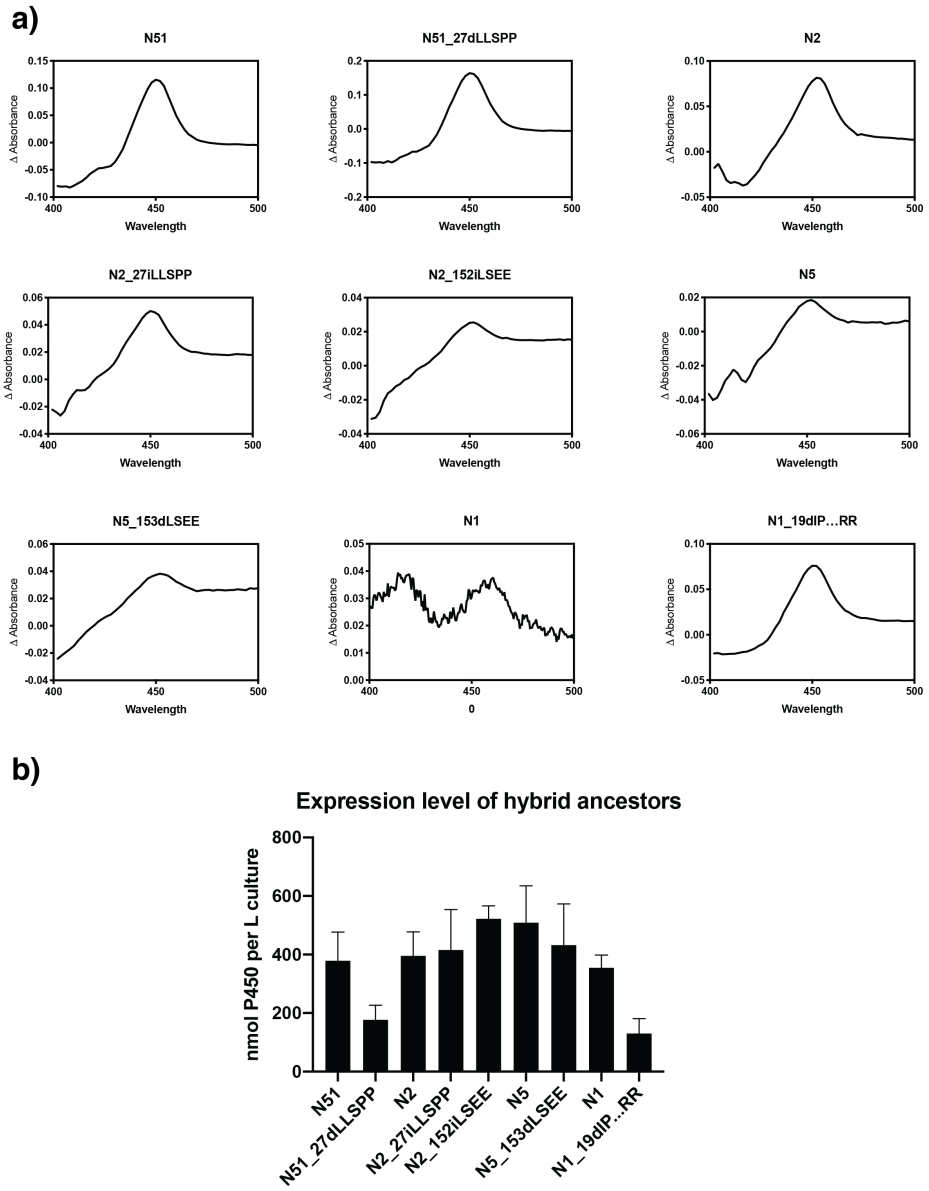

Figure 10: Expression of CYP2U hybrid ancestors. **a**, Fe(II) vs. Fe(II).CO difference spectra for CYP2U ancestors in *E. coli* membranes. **b**, Expression level of CYP2U ancestors in *E. coli* cultures quantified using Fe(II) vs. Fe(II).CO difference spectroscopy. Data are means +/- SEM, $N = 3$.