An atomic-resolution view of neofunctionalization in the evolution of apicomplexan lactate dehydrogenases

Jeffrey I. Boucher[1], Joseph R. Jacobowitz[1], Brian C. Beckett[1], Scott Classen[2] and Douglas L. Theobald[1]

[1]*Brandeis University, Department of Biochemistry, Waltham MA 02454, USA.*
[2]*Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.*

Malate and lactate dehydrogenases (MDH and LDH) are homologous, core metabolic enzymes that share a fold and catalytic mechanism yet possess strict specificity for their substrates. In the Apicomplexa, convergent evolution of an unusual LDH from MDH resulted in a difference in substrate preference exceeding 12 orders of magnitude. The molecular and evolutionary mechanisms responsible for this extraordinary functional shift are currently unknown. Using ancestral sequence reconstruction, we find that the evolution of pyruvate specificity in apicomplexan LDHs arose through a classic neofunctionalization mechanism characterized by long-range epistasis, a promiscuous intermediate, and relatively few gain-of-function mutations of large effect. Residues far from the active site determine specificity, as shown by the crystal structures of three ancestral proteins that bracket the key gene duplication event. This work provides an unprecedented atomic-resolution view of evolutionary trajectories resulting in the *de novo* creation of a nascent enzymatic function.

Background

The common ancestor of the eukaryotic Apicomplexa evolved nearly one billion years ago (1), and its modern descendants comprise a large phylum of intracellular parasites that are currently responsible for numerous devastating metazoan diseases, including malaria (*Plasmodium*), toxoplasmosis (*Toxoplasma*), cryptosporidiosis (*Cryptosporidium*), cyclosporiasis (*Cyclospora*), and babesiosis (*Babesia*). A key event in the early evolution of the Apicomplexa was the acquisition of a malate dehydrogenase (MDH) via lateral gene transfer from α-proteobacteria (2-4). Following a gene duplication event roughly 700-900 Mya, one copy of this MDH evolved a novel substrate specificity to become a highly specific lactate dehydrogenase (LDH) that is now essential to the life cycle of many modern apicomplexans (5). As a core metabolic enzyme that evolved independently of metazoan LDH, the unique apicomplexan LDH has attracted significant attention as a potential drug target (6-9). However, the molecular and evolutionary mechanisms that drove this switch in substrate specificity are currently unknown.

LDH and MDH are homologous, 2-ketoacid oxidoreductases that share both a protein fold (10) (**Figure 1-figure supplement 1**) and a common catalytic mechanism (11-16) (**Figure 1**). Both enzymes are found in central metabolism: MDH catalyzes the interconversion of oxaloacetate and malate in the citric acid cycle, and LDH converts pyruvate to lactate in the final step of anaerobic glycolysis. Despite their structural and catalytic similarities, modern apicomplexan LDHs and MDHs have extraordinarily strict substrate specificity. For example, *Plasmodium falciparum* (*Pf*) MDH and LDH each prefer their respective substrates by over six orders of magnitude. The biophysical basis for this extraordinary substrate preference is presently an unresolved question.

A conspicuous structural difference between apicomplexan MDHs and LDHs is an insertion within the active site loop of the LDHs (17, 18) (**Figure 2**). In the LDH/MDH

superfamily, closure of this loop over the active site is rate-limiting during catalysis (16), and mutations within this loop have a large effects on activity and substrate specificity (19). For example, simply mutating Gln102 to Arg in the specificity loop of *Bacillus stearothermophilus* (*Bs*) LDH converts the enzyme into an MDH, shifting specificity from a 10^3-fold preference for pyruvate to a 10^4-fold preference for oxaloacetate (19) (**Figure 1**, residue numbering is based on the dogfish LDH convention (20)). In fact, all known MDHs have an Arg at position 102, while canonical LDHs have a Gln, and consequently residue 102 has been called the "specificity residue" (21). Residue 102 is thought to contribute to substrate discrimination by balancing the substrate charge within the active site: the positively charged Arg in MDHs forms a salt bridge with the C4 carboxylate of oxaloacetate, whereas the neutral Gln in canonical LDHs packs with the C3 methyl of pyruvate (**Figure 1**). Yet, attempts to convert an MDH into an LDH by mutating Arg102 to Gln have met with limited success (22, 23). In the apicomplexan LDHs, residue 102 is not a Gln but a Lys, a relatively conservative substitution compared to the MDH Arg. It is currently not understood why *Plasmodium* LDHs lack activity towards oxaloacetate, despite having a positively charged sidechain at residue 102 similar to MDHs (8, 24-28).

Apicomplexan LDH evolved from the duplication of an ancestral MDH gene (3, 4). Gene duplication is widely considered the major force that has driven the evolutionary diversity of protein functions (29). There are three general ways duplicated genes can be fixed in a population by selection: 1) "dosage selection", beneficial increase in dosage due to multiple copies, 2) "subfunctionalization", specialization of previously existing functions, or 3) "neofunctionalization", creation of a novel function through the accumulation of beneficial, gain-of-function mutations (30). Most mutations, however, are either neutral or detrimental. A new duplicated gene typically degrades to a crippled pseudogene before it can acquire the rare beneficial mutations needed to confer a selectable function (31, 32). Hence, classical neofunctionalization has fallen out of favor in preference for models that begin with the duplication of a multifunctional protein, such as "specialization" and "subfunctionalization"

models. Currently the molecular and evolutionary mechanisms that create novel functions in gene duplicates are fiercely debated (29, 33-35), and there are no clear examples of classic neofunctionalization or gain-of-function mutations (36-38).

The apicomplexan LDH and MDH enzyme family provides an exceptional model system for investigating several long-standing questions in molecular evolution, including the mechanisms available to convergent evolution, the number of mutations required to produce a nascent function, the role of promiscuous intermediates during evolution of function, and the effects of epistasis on evolutionary irreversibility. In order to identify the biophysical and evolutionary mechanisms responsible for pyruvate specificity in apicomplexan LDHs, we have reconstructed ancestral proteins along the evolutionary trajectories leading to modern apicomplexan MDHs and LDHs (**Figure 3B**). We kinetically and structurally characterized the ancestral proteins together with multiple evolutionary intermediates. This work provides a clear example of neofunctionalization in protein evolution and the first crystal structures documenting the evolution of a new enzyme. We show that apicomplexan LDHs evolved as the result of few mutations of large effect via the classic neofunctionalization of a duplicated MDH gene.

Results

LDH enzymes have evolved independently at least four times

A maximum likelihood phylogeny of representatives of all known LDH and MDH proteins provides strong support for five distinct protein clades (**Figure 3A, Figure 3-figure supplement 1**): canonical LDHs, "LDH-like" MDHs, mitochondrial-like MDHs, cytosolic-like MDHs, and the poorly characterized HicDHs (hydroxyisocaproate-related dehydrogenases), confirming previous phylogenetic analyses (2-4, 39).

The HicDH clade are close sequence homologs of a known hydroxyisocaproate dehydrogenase. They all possess a residue other than a Gln or an Arg at the "specificity" position 102, as well as insertions of varying lengths within the catalytic loop between residues 102 and 109. Despite these alterations within the catalytic loop, all other catalytic residues (Arg109, Asp168, Arg171, and His195) are conserved. Only one taxon within the HicDH clade has been functionally characterized, DHL2_LACCO, which is a specific hydroxyisocaproate dehydrogenase (85). These observations suggest that the clade features dehydrogenases with altered substrate specificity.

Except for the HicDHs, which are exclusively eubacterial, both eukaryotic and eubacterial enzymes are found in all major clades (**Figure 3-figure supplement 2**). The "LDH-like" MDH clade additionally contains archaeal dehydrogenases, which are basal and group to the exclusion of the bacterial MDHs.

Intriguingly, three different groups of LDH proteins cluster with high confidence outside of the canonical LDH clade. A set of trichomonad LDHs found in the cytosolic-like MDH clade are thought to have evolved from a recent gene duplication of an MDH (40). The Trichomonads appear to lack a canonical LDH. A prominent eukaryotic group of LDH and MDH proteins from the Apicomplexa nests deep within the bacterial "LDH-like" MDHs, sister to many Rickettsiales sequences, signifying a horizontal gene transfer event from α-proteobacteria

to the eukaryotic Apicomplexa. We find no evidence that the Apicomplexa have canonical LDH or conventional eukaryotic-type MDH (either cytosolic- or mitochondrial-like MDHs), despite searching in many available complete apicomplexan genomes (multiple Eimeria, Neospora, Toxoplasma, Plasmodium, and Cryptosporidium species) (41-43). In the Apicomplexa, LDH activity has apparently evolved independently twice (**Figure 3B, Figure 3-figure supplement 3**), once in a lineage leading to *Plasmodium*-related species and once in *Cryptosporidium*. The apicomplexan portion of the LDH/MDH gene phylogeny is consistent with recent apicomplexan species phylogenies constructed from concatenated protein sequences (44).

We rooted the MDH/LDH phylogeny using the Rossmann fold domain of the distantly related α/β-glucosidases and aspartate dehydrogenases as outgroups. The ML root position apparently splits the tree into two large groups: (1) one which contains the cytosolic- and mitochondrial-like MDHs, which are largely dimeric, and (2) another which contains the canonical LDHs, "LDH-like" MDHs, and HicDHs, which are primarily tetrameric (**Figure 3-figure supplement 1**). While the ML root position is robust to variation in taxon coverage, the exact location is poorly supported. Nevertheless, there is strong support for a root position within the central MDH section of the tree and outside of the five identified clades, including the canonical LDH clade (confidence level > 0.99985 according to the aLRT), indicating that the canonical LDHs evolved from an ancestral MDH. The global rooting and the location of the three separate LDH groups, deep within MDH clades, indicate that LDH enzymes have evolved convergently from MDHs at least four times in the superfamily.

An insertion in the catalytic loop of apicomplexan LDHs

In the present work, our focus is on the convergent evolution of the unusual apicomplexan LDHs. With the α-proteobacterial "LDH-like" MDHs as the closest outgroup, the apicomplexan enzymes are split into two main groups: (1) LDHs belonging to *Toxoplasma*,

142 *Plasmodium,* and related protists, and (2) MDHs belonging to *Plasmodium* and
143 *Cryptosporidium*. Apart from their atypical phylogenetic position, the apicomplexan MDHs
144 appear as typical α-proteobacterial "LDH-like" MDHs, containing all the key catalytic residues
145 including Arg102. The *Cryptosporidium* LDHs are an exception, being nested within the
146 apicomplexan MDH clade partitioned from the rest of the apicomplexan LDHs.
147 *Cryptosporidium* LDHs have a Gly at position 102 and are thought to be a product of an
148 independent, convergent duplication event (39).
149 In contrast, the large apicomplexan LDH clade is demarcated by a unique, conserved
150 five-residue insertion in the active site loop. While the apicomplexan LDH and MDH proteins
151 are moderately divergent, with about 45% sequence identity, the differences are largely confined
152 to exterior residues removed from the active sites. One important difference is that the
153 apicomplexan LDHs have Lys102 for the "specificity residue", rather than a Gln as found in the
154 canonical LDHs (**Figure 2-figure supplement 1**). Apicomplexan proteins frequently contain
155 numerous insertions relative to proteins from other species (45, 46), a characteristic thought to
156 result from various factors, including high AT genome content, DNA strand slippage, double
157 strand break repair, high recombination rates, and selection pressure for parasite antigenic
158 variation. Except for Met106, the amino acid and coding sequence immediately flanking the
159 apicomplexan LDH loop insertion is largely conserved with α-proteobacterial MDHs (**Figure 2-**
160 **figure supplement 1**). It is therefore likely that a mutation "expanded" the Met106 codon to
161 code for six residues, resulting in the observed five-residue insertion and the Met106Lys
162 mutation. Henceforth we will refer to this expansion mutation as the "six-residue loop
163 insertion".

164 ***Trp107f is the modern apicomplexan LDH specificity residue***

165 In the modern apicomplexan enzymes, the six-residue insertion in the LDH specificity
166 loop (positions 99-112) induces two significant structural changes relative to MDH (**Figure 2**).

First, LDH residue Lys102 is excluded from of the active site, unlike the corresponding Arg102 in MDH, which is enclosed within the active site and participates in functionally important interactions with the substrate. Second, LDH Trp107f, which is part of the novel insertion, occupies the same space as Arg102 in MDH (by convention, residues in the insertion are labeled using numbers and letters to maintain consistency with homologous positions in the dogfish LDH, **Figure 2**).

The only prominent structural difference between the active sites of the LDH and MDH proteins is the replacement of MDH Arg102 with LDH Trp107f. Trp107f is positioned where it could presumably interact with the distinguishing C3 methyl of the pyruvate substrate, while MDH Arg102 interacts with the C4 carboxylate of oxaloacetate (21). As a bulky, hydrophobic residue, Trp107f could recognize pyruvate in preference to oxaloacetate by two mechanisms: (1) a hydrophobic interaction with the pyruvate C3 methyl vs the negatively charged oxaloacetate methylene carboxylate and (2) steric occlusion of the methylene carboxylate of oxaloacetate. Furthermore, Trp107f is conserved in all apicomplexan LDHs (**Figure 2-figure supplement 1**), suggesting negative selection and functional importance. We therefore hypothesized that Trp107f plays an important role in pyruvate recognition.

We tested the functional importance of residues in the specificity loop in *Pf*LDH with an "alanine scan" by individually mutating each residue in positions 101-108 to an alanine (**Figure 2-figure supplement 2**, note Ala103 was mutated to a serine). We assessed the activity of the mutants using k_{cat}/K_m, a measure of enzymatic specificity and catalytic efficiency, as determined from steady state kinetic assays. Mutating Trp107f to Ala reduced pyruvate activity by five orders of magnitude, whereas mutations at all other positions had effects less than a single order of magnitude, including the canonical specificity residue at position 102. The Trp107fAla mutation affects both k_{cat} (1500-fold decrease) and K_m (50-fold increase).

To assess the effects of Trp107fAla mutation on the specificity loop conformation, we solved the crystal structure of *Pf*LDH-W107fA (1.1Å) in the presence of oxamate and NADH.

193 The protein crystallizes in the same space group as the wild-type *Pf*LDH, with nearly identical
194 cell dimensions (**Figure 2-figure supplement 3A**). In the W107fA mutant, the specificity
195 loop is disordered between residues Leu86 and Arg99, as is often seen in structures in which the
196 loop is in the open conformation. In the mutant, residues 102-105 are in a linear α-helical
197 conformation, in contrast to the wild-type *Pf*LDH closed state which has a very prominent 60°
198 kink in the α-helix at Pro104. Thus, the only significant difference between the wild-type and
199 mutant structures is that the *Pf*LDH-W107fA specificity loop is found in the open conformation,
200 consistent with weaker binding of substrate (**Figure 2-figure supplement 3B**). These results
201 indicate that Trp107f is necessary for pyruvate activity in apicomplexan LDHs, and that it has
202 become the new "specificity residue" despite the fact that Trp107f does not align in sequence
203 with the canonical specificity residue at position 102 (**Figure 2-figure supplement 1**).

204 ***The loop insert fails to swap specificity in modern LDH and MDH***

205 During evolution, the six-residue insertion displaced the canonical specificity residue at
206 position 102 and apparently switched substrate preference in apicomplexan LDHs. If this
207 insertion is sufficient for pyruvate recognition, then adding the insertion to a modern
208 apicomplexan MDH should convert the enzyme to an LDH. To test this hypothesis, we
209 incorporated the six-residue insertion from *Pf*LDH into the catalytic loop of *Pf*MDH (*Pf*MDH-
210 INS) and the *Cryptosporidium parvum* (*Cp*) MDH (*Cp*MDH-INS). The chimeric proteins
211 showed a >100-fold reduction in oxaloacetate activity with no significant gain in pyruvate
212 activity (**Figure 4**). Like other MDHs, the apicomplexan MDHs have an Arg at position 102
213 that is important for oxaloacetate recognition; in the modern apicomplexan LDHs position 102
214 is a Lys. The Arg102Lys mutation may be necessary to eliminate oxaloacetate activity and
215 increase pyruvate activity. Therefore, we also mutated Arg102 to Lys in the *Pf*MDH chimera
216 (*Pf*MDH-R102K-INS). However, this mutation reduced activity towards oxaloacetate by
217 another 100-fold, with no increase in pyruvate activity (**Figure 4**).

Alternatively, it may be possible to revert a modern apicomplexan LDH to MDH-like specificity by deleting its six-residue loop insertion. To test this hypothesis we removed the insertion from *Pf*LDH and from the *Toxoplasma gondii* (*Tg*) LDH2 (constructs *Pf*LDH-DEL and *Tg*LDH2-DEL). However, deleting the insertion from the modern LDHs abolishes pyruvate activity with no significant gain of oxaloacetate activity (**Figure 4**). Both of these deletion mutants retain a Lys at position 102, but a specific MDH likely requires an Arg at position 102. Mutating Lys102 to Arg in *Pf*LDH-DEL results in a two order-of-magnitude gain in oxaloacetate activity (**Figure 4**). However, this mutant fails to recapitulate the level of oxaloacetate activity seen in modern apicomplexan MDHs. In the modern enzymes, substrate specificity cannot be switched with mutations involving the loop insert and position 102, indicating that additional residues govern substrate preference.

The ancestral MDH and LDH enzymes are specific and highly active

The apicomplexan LDH and MDH phylogeny strongly suggests that after (or coincident with) the crucial gene duplication event, the nascent LDH branch gained pyruvate activity due to the six-residue insertion in the specificity loop. This presents a conundrum, as our mutation trials in the modern enzymes failed to recapitulate the historical swap in specificity. However, the modern apicomplexan LDH and MDH enzymes differ by over 200 residues in addition to the loop insert and Arg102Lys, differences that have accumulated in the descendants of the ancestral MDH and LDH. Any of these differences may detrimentally affect the ability to switch substrate specificity with the insertion in the modern enzymes. We therefore reasoned that the ancestral background may be necessary for swapping specificity with the loop insertion. To test this, we reconstructed and characterized four key ancestral enzymes: (1) AncMDH1, the ancestral protein that was transferred from α-proteobacteria to the archaic Apicomplexa, (2) AncMDH2, the last common ancestor of all apicomplexan MDHs and LDHs, found at the critical

duplication event, (3) AncMDH3, the last common ancestor of all modern apicomplexan MDHs, and (4) AncLDH, the last common ancestor of modern apicomplexan LDHs (**Figure 3B**).

All four ancestral proteins are highly active in steady state kinetic assays, with substrate preferences and catalytic efficiencies that are similar to their modern apicomplexan descendants (**Figure 5**), despite sharing only 49-71% sequence identity with the modern apicomplexan proteins (**Figure 5-figure supplement 1**). AncMDH1, AncMDH2, and AncMDH3 are highly specific MDHs with negligible pyruvate activity, having even greater activity towards oxaloacetate than modern *Plasmodium* and *Cryptosporidium* MDHs (**Figure 5**). AncLDH is a highly active and specific LDH, with very low activity towards oxaloacetate (**Figure 5**).

The loop insert successfully swaps specificity in both ancestral LDH and MDH

AncLDH differs from AncMDH2 by 66 residues, including the six-residue insertion and Arg102Lys. We investigated the evolutionary trajectory from AncMDH2 to AncLDH by characterizing three different mutations in the AncMDH2 background: (1) the addition of AncLDH's six-residue insertion to the AncMDH2 specificity loop, (2) Arg102Lys, which assesses the effect of changing the canonical specificity residue, and (3) the remaining 59 residues that separate AncLDH from AncMDH2, simultaneously changed to their AncLDH identities.

Incorporating the loop insertion into AncMDH2 confers significant pyruvate activity with minimal effect on oxaloacetate activity, resulting in a highly active, bifunctional enzyme (AncMDH2-INS, **Figure 6**). In contrast, the Arg102Lys mutation in the AncMDH2 background (AncMDH2-R102K, **Figure 6**) reduces oxaloacetate activity by more than a 100-fold, with no increase in pyruvate activity. The 59 mutations in the AncMDH2 background have a minimal effect on the activity towards both substrates (AncMDH2-59Mut, **Figure 6**). Note that the AncMDH2-59Mut construct is equivalent to a modified AncLDH construct with the Lys102Arg mutation and the insertion deleted from the loop. Therefore, only two changes — Lys102Arg and the loop deletion — are sufficient to convert the AncLDH construct to a highly active and specific MDH.

Combinations of these mutations confirm that the insertion is primarily responsible for the evolution of pyruvate activity. Adding the 59 mutations to AncMDH2-INS (resulting in a construct that differs from AncLDH by only one residue) has little additional effect (AncMDH2-INS-59Mut, **Figure 6**). Surprisingly, the combination of Arg102Lys and the 59 mutations, a construct that differs from AncLDH by just the six-residue insertion, yields a crippled MDH enzyme with 1,000-fold less oxaloacetate activity than AncMDH2 (AncMDH2-R102K-59Mut, **Figure 6**). However, the combination of Arg102Lys and the loop insertion in the AncMDH2 background is sufficient to confer pyruvate activity and specificity comparable to AncLDH (AncMDH2-INS-R102K, **Figure 6**).

Ancestral kinetics are robust to reconstruction uncertainty

Ancestral sequence reconstruction is a difficult statistical problem that strongly relies on evolutionary assumptions, which may be unrealistic, and on available sequence data, which is inherently incomplete. The likelihood and Bayesian ancestral reconstruction methodology that we use produces the most probable ancestral sequence given certain evolutionary model assumptions, along with a posterior probability for alternative amino acids at each position (**Figure 6-figure supplement 1-6**). Ambiguous residues are generally associated with positions of low conservation and presumably less functional importance. The reconstructed AncMDH2 and AncLDH sequences have 31 and 48 ambiguous positions, respectively, all of which are located outside of the "first active site shell" (defined as within 6 Å of the substrate). In order to verify that these sequence ambiguities do not affect our kinetic results, alternative ancestral sequences were reconstructed and assayed. We tested the robustness of our ancestral proteins by constructing alternative ancestors based on perturbed sequence data, evolutionary assumptions, and phylogenetic methodology. Both phylogenies give very similar relationships, and Figure 2B summarizes both equally well. The alternative AncMDH2 (AncMDH2*) differs from AncMDH2 by 27 residues; the alternative AncLDH (AncLDH*) differs from AncLDH by 19 residues.

The alternative ancestral reconstructions behave very similar to the prior reconstructions. AncMDH2* is a strict MDH, and AncLDH* is a strict LDH (**Figure 6/Figure 6-figure supplement 7**). Addition of the six-residue insertion from AncLDH* to AncMDH2* confers pyruvate specificity without adversely affecting oxaloacetate activity (AncMDH2*-INS, **Figure 6-figure supplement 7**). In the AncMDH2* background, mutating Arg102 to Lys together with the 58 mutations from AncLDH* yields a poor enzyme with little pyruvate activity (AncMDH2*-R102K-58Mut). The kinetic behavior of these AncMDH2* constructs closely matches those seen with the corresponding AncMDH2 constructs (AncMDH2, AncMDH2-INS, and AncMDH2-R102K-59Mut, **Figure 6**).

Crystal structures of ancestral MDH, LDH, and an evolutionary intermediate

In order to understand the structural changes during evolution that shifted the enzymatic substrate specificity of the apicomplexan dehydrogenases, we determined the high-resolution crystal structures of three ancestral proteins bracketing the key duplication event: AncMDH2 (1.9 Å), AncLDH* (1.3 Å), and AncMDH2-INS (1.8 Å). All three ancestral proteins adopt the same overall fold and conformation as the modern, descendant enzymes. In particular, the ancestral active sites and specificity loops are highly similar to their modern counterparts.

Ancestral malate dehydrogenase: AncMDH2

The AncMDH2 structure superposes closely with the modern *Cp*MDH structure (47) (0.56 Å RMSD), although differing at ~119 residue positions (62% sequence identity, **Figure 7A**). In the modern and ancestral MDHs, all residues within the first shell of the active sites (within 6 Å of the substrate) are identical, and the active site conformations are correspondingly highly similar (**Figure 7B**). The first shell active site residues comprise Arg102, Arg109, Leu112, Asn140, Leu167, Asp168, Arg171, His195, Met199, Gly236, Gly237, Ile239, Val240, Ser245, Ala246, and Pro250.

Compared to the modern MDH, only slight differences are seen in the substrate loop backbone and the positioning of the Arg102 and Arg109 sidechains, which are the only residues from the specificity loop that directly interact with the substrate. However, these modest conformational differences are largely within coordinate error, as the loop residues have some of the highest B-factors in the structures. Furthermore, AncMDH2 was crystallized with lactate and NADH while *Cp*MDH was crystallized with citrate and ADPR (an NADH analog lacking the nicotinamide ring). Citrate is roughly three times larger than lactate and has likely affected the position of substrate loop in the *Cp*MDH structure.

Ancestral lactate dehydrogenase: AncLDH*

The ancestral AncLDH* and modern apicomplexan LDH structures are likewise highly similar (6, 26, 28) (RMSD ~0.8 Å, **Figure 7C**), while sharing only 63-71% sequence identity. The first shell active site residues are identical in the AncLDH* and modern *Toxoplasma* LDHs, comprising the same residues as the apicomplexan MDH active site with the sole exception of position 102, which is replaced by Trp107f in the LDHs. The modern *Plasmodium* LDHs have two different residues in the active site first shell: Pro246 and Ala236, rather than Ala246 and Gly236 as found in *Tg*LDH1, *Tg*LDH2, and AncLDH*. The conformations of the ancestral and modern active sites are nearly indistinguishable, with only small differences in the specificity loop conformation (**Figure 7D**).

In both the ancestral and modern LDH structures, Trp107f and Arg109 are the only residues from the specificity loop that interact with the substrate. As in the modern LDH structures, ancestral Lys102 does not interact with the substrate but points away from the active site into solution. In contrast, Trp107f is buried within the active site, with the edge of the indole ring interacting with the pyruvate C3 methyl, which is the very chemical moiety that distinguishes pyruvate from oxaloacetate.

The largest differences between the modern and ancestral proteins are confined to two regions: (1) a small shift of the entire C-terminal helix, and (2) a loop opposite the active site

specificity loop (residues 242-244, hereafter called the "opposing loop"). The modern *Plasmodium* LDHs have a two-residue deletion within this opposing loop (highlighted in cyan in **Figure 7C**), while the opposing loop is shared with AncLDH* and the *Toxoplasma* LDHs. The ancestral LDHs also share very modest oxaloacetate activity with the modern *Toxoplasma* LDHs, while the *Plasmodium* LDHs lack oxaloacetate activity (**Figure 5**). This correlation indicates the opposing loop deletion (and perhaps Ala236 and Pro246) may be responsible for the unusually strict substrate specificity of the modern *Plasmodium* LDHs.

Ancestral malate dehydrogenase with loop insertion: AncMDH2-INS

We also crystallized AncMDH2-INS, a bifunctional AncMDH2 construct with the six-residue specificity loop insertion. This AncMDH2-INS construct represents a possible intermediate along the evolutionary trajectory between the MDH duplication event and the ancestral apicomplexan LDH. AncMDH2-INS was successfully co-crystallized with both oxamate/NADH and lactate/NADH. In the closed form, the specificity loop adopts an LDH-like confirmation with Trp107f occupying the specificity position and Arg102 oriented into solution, similar to how Lys102 is positioned in the modern and ancestral LDH structures (**Figure 7F**). The lactate and oxamate structures are highly similar (RMSD 0.19 Å), and the active site architectures are nearly indistinguishable.

The three ancestral proteins, AncMDH2, AncLDH*, and AncMDH2-INS, are all highly similar (RMSD 1.20 Å) with the main structural differences found in the conformation of the specificity loop (**Figure 7E,** RMSD 0.86 Å excluding residues in the specificity loop). Otherwise the first shell active site residues are identical between AncMDH2-INS and AncLDH*, and the conformations of the active sites are correspondingly similar (**Figure 7F**).

Convergent pathways available to the ancestral MDH

Given the known importance of position 102, the "specificity residue", in substrate recognition, we wondered whether different residues at position 102 could confer pyruvate

activity. Position 102 in fact differs in the four convergent LDH families: Gln in canonical LDHs (19), Lys in the apicomplexan LDHs, Gly in *Cryptosporidium* LDHs (39), and Leu in trichomonad LDHs (40). Could the ancestral apicomplexan MDH have evolved pyruvate specificity by any of these alternative routes? To answer this question, we evaluated the potential of these different amino acids at the 102 position to confer pyruvate specificity in the AncMDH2 background. Each mutation increases pyruvate activity, but none result in a highly specific LDH. The canonical mutation (Arg102Gln) results in the largest gain in pyruvate activity (2,800-fold) and the smallest loss of oxaloacetate activity (2,500-fold) (**Figure 8**). Additionally, we tested whether the full six amino acid insertion was required to confer pyruvate specificity in AncMDH2 or if simply mutating Arg102 to Trp was sufficient. The Arg102Trp mutation all but abolishes activity towards both substrates, indicating that the loop insertion was necessary to switch the specificity residue (**Figure 8**).

Discussion

An alternate mechanism of specificity in the convergent apicomplexan LDH

Substrate recognition in the canonical MDHs and LDHs is thought to be determined by a "specificity residue" in the active site loop at position 102. All known MDHs have Arg at position 102, while canonical LDHs have Gln (21). In the classic explanation of the molecular mechanism of substrate specificity, residue 102 discriminates between pyruvate and oxaloacetate primarily via charge conservation (19). In MDHs, the positively charged Arg interacts with and balances the negatively charged carboxylate of oxaloacetate. If pyruvate were to bind in the active site, loop closure would result in a buried and unbalanced positive charge, which is unfavorable. In canonical LDHs, the neutral Gln interacts with the neutral pyruvate methyl group. Oxaloacetate binding would similarly result in the unfavorable burial of an unbalanced negative charge.

In the apicomplexan LDHs, evolution has converged on pyruvate specificity using an alternative molecular mechanism. Residue 102 is not a Gln but a positively charged Lys, similar to Arg102 of MDHs, leading many researchers to wonder why apicomplexan LDHs lack activity towards oxaloacetate (8, 24-28). However, during the evolution of the apicomplexan LDH from the ancestral MDH, the six-residue insertion in the active site loop shifted both the position and identity of the "specificity residue" from Arg102 to Trp107f. Due to the insertion, residue 102 no longer interacts with the substrate and is extruded from the active site. In contrast, the hydrophobic Trp107f packs against the C3 methyl of the pyruvate substrate. Similar to the canonical LDH, oxaloacetate binding would result in an unbalanced and buried negative charge. As a large bulky residue, Trp107f can also occlude binding of the larger oxaloacetate, in which a methylene carboxylate replaces the pyruvate methyl.

However, as discussed in detail below, this simplistic explanation is complicated by long-range epistatic interactions. When the six-residue insertion is introduced into the modern

apicomplexan MDH, specificity is not switched; both specificity and activity are lost. Similarly, removal of the insertion from the modern apicomplexan LDHs fails to swap specificity and kills the enzymes. Therefore, while Trp107f is necessary for substrate specificity in the apicomplexan enzymes (as indicated by the alanine scan mutations), it is insufficient to confer specificity.

The bifunctionality of AncMDH2-INS and AncMDH2-INS-59Mut also presents a conundrum. Why do these constructs have high activity towards both pyruvate and oxaloacetate substrates? The crystal structure of AncMDH2-INS offers few clues, since the loop insertion, including Trp107f, adopts the same conformation as seen in AncLDH and the modern apicomplexan enzymes. Both the AncMDH2-INS and AncMDH2-INS-59Mut constructs have an Arg at position 102, like the MDHs. In fact, the bifunctional AncMDH2-INS-59Mut enzyme differs from the strict AncLDH by only a R102K mutation, suggesting that Arg102 is responsible for the oxaloacetate activity of AncMDH2-INS and AncMDH2-INS-59Mut. We speculate that perhaps the enzymes change conformation depending upon the substrate. When using pyruvate, these bifunctional enzymes may adopt an LDH-like conformation in which Trp107f interacts with the substrate (as seen in the crystal structure). On the other hand, when presented with oxaloacetate, perhaps Trp107f flips out of the active site, and Arg102 flips in to interact with substrate in a manner similar to the canonical MDHs. We are currently testing this hypothesis.

Apicomplexan LDH evolved by classical neofunctionalization

Our data show that apicomplexan LDHs evolved from a horizontally transferred proteobacterial MDH by a classic neofunctionalization mechanism of gene duplication. Because debasement to a pseudogene is much more likely to occur prior to the evolution of a novel function, neofunctionalization has fallen out of favor as a mechanism for the evolution of novel functions. A variety of alternative specialization models have been proposed that feature a reduced risk of non-functionalization. Though differing in details, all specialization models feature a promiscuous common ancestor of the duplicated proteins.

The reconstructed AncMDH2, which represents the last common ancestor of the apicomplexan MDH and LDHs, is a highly active and specific MDH, preferring oxaloacetate over pyruvate by seven orders of magnitude (**Figure 6**). The activity of AncMDH2 towards pyruvate is barely detectable, requiring a high enzyme concentration to quantify. AncMDH2's k_{cat} for pyruvate is 0.07 s^{-1}, with a K_m of 20 mM, while the physiological concentration of pyruvate is estimated to be about three orders-of-magnitude lower (e.g., ~50 μM in human erythrocytes (48), the *Plasmodium* host during its blood stage). Based on these kinetic parameters, each AncMDH2 reduces one pyruvate molecule per hour. While the enzyme can be forced to reduce pyruvate *in vitro*, this negligible activity is unlikely to have been subjected to selection *in vivo*. Therefore, the various specialization hypotheses, which require a promiscuous ancestor, are poor models for apicomplexan LDH evolution. Activity towards pyruvate increased by over seven orders of magnitude on the evolutionary lineage between AncMDH2 and AncLDH, indicating neofunctionalization.

A highly active, promiscuous intermediate

One of the most favored specialization models is "escape from adaptive conflict" (EAC) (49). EAC holds that functional specialization is driven by an inability to simultaneously optimize multiple functions on a single protein scaffold. Gene duplication relieves this constraint and allows for the independent optimization of conflicting functions. Although the apicomplexan AncMDH2 is highly specific, promiscuous intermediates did play a role in the functional transition between AncMDH2 and AncLDH. AncMDH2-INS and AncMDH2-INS-59Mut have high levels of MDH and LDH activity in a single protein scaffold (**Figure 6**). Both the presence of bifunctional intermediates and the high specificity of AncMDH2 conflict with fundamental predictions of the EAC specialization model.

Convergent evolution of apicomplexan LDH involved long-range epistasis

The evolution of apicomplexan LDHs involved strong epistasis that has profoundly influenced the convergent evolution of pyruvate activity. Epistasis refers to interactions between residues that potentiate the effects of a mutation depending on the presence or absence of other residues (50). Epistasis can constrain the order of mutations and the pathways accessible to evolution, and hence it is of great importance in understanding the evolution of novel functions. In the apicomplexan dehydrogenases, the evolutionary mutations that switched specificity from oxaloacetate to pyruvate (the six-residue insertion and Arg102Lys) are insufficient to confer pyruvate activity in modern apicomplexan MDHs (*Pf*MDH-R102K, *Pf*MDH-INS, *Cp*MDH-INS, *Pf*MDH-R102K-INS, **Figure 4**). However, these mutations are sufficient to confer pyruvate function and specificity in the AncMDH2 background (AncMDH2-INS, AncMDH2-INS-R102K, **Figure 6**). Similarly, removal of the insert from the modern LDHs (*Pf*LDH-DEL and *Tg*LDH2-DEL, **Figure 4**) kills the enzymes, while removal of the insert from the ancestral LDH (AncMDH2-R102K-59Mut, **Figure 6**) results in a weak MDH. The different effects of these mutations, depending upon the sequence of the rest of the protein, provide direct evidence of epistatic interactions.

Why do these historical mutations "work" in the ancestral enzymes, but not in the modern ones? Epistatic interactions are often mediated by direct physical contact. For example, the active site of the ancestral MDH could have certain residues that the modern MDH lacks, residues that interact with the insertion and allow it to preferentially bind pyruvate. However, the active sites of the ancestral and modern MDHs are identical in sequence and virtually indistinguishable in structure (**Figure 7B**), as are the active sites of the ancestral and modern LDHs (**Figure 7D**) and the AncMDH2-INS intermediate (**Figure 7F**). In fact, the active sites of the MDHs and the LDHs are also identical in sequence except for the 102 position, and they are otherwise highly structurally similar. Therefore, residues remote from the active sites necessarily affect the substrate specificity of the enzymes.

In principle, these long-range epistatic residue interactions could differentially modify the structure of the active site. Certain residues found in the ancestral MDH, but not in the modern MDH, could position the active site residues so that they allowed the insertion to confer pyruvate activity. In this scenario the active site residues of the ancestral and modern MDHs would be identical, but their conformations would differ due to interactions with residues in other parts of the protein. However, the crystal structures reveal ancestral, intermediate, and modern active sites that are nearly indistinguishable, suggesting that epistasis has modified the protein dynamics or shifted the energy landscape, effects that are largely invisible to static crystal structures.

Epistasis prevents mechanistic convergence in the LDH/MDH superfamily

Interestingly, *Bacillus subtilis* (*Bs*) LDH reverts to an MDH with only a single mutation, Gln102Arg, indicating a lack of complicating epistatic effects (19). The kinetics of wild-type *Bs*LDH with pyruvate are comparable to those for the Gln102Arg mutant with oxaloacetate (e.g., *Bs*LDH has a k_{cat}/K_M for pyruvate of 4.2×10^6 M^{-1} s^{-1}, and the *Bs*LDH-Q102R mutant has the same k_{cat}/K_M for oxaloacetate). However, *Bs*LDH likely is an exception in the LDH/MDH superfamily, since the reverse mutation (Arg102Gln) fails to switch specificity in MDHs from two other species (22, 23). In *Haloarcula marismortui* (*Hm*) MDH, the Arg102Gln mutation switches specificity, but the mutant's k_{cat}/K_M for pyruvate is 200-fold less than the wild-type's k_{cat}/K_M for oxaloacetate. The Arg102Gln mutation in *Escherichia coli* (*Ec*) MDH is even less effective, as it converts a highly active MDH to an enzyme with low activity on both substrates (10,000-fold lower k_{cat}/K_M). Hence, the strong epistasis observed in apicomplexan LDH and MDHs is likely a general phenomenon within the superfamily.

LDH evolved convergently from MDH four separate times in the superfamily, but did the activity evolve by the same mechanism each time? Each event has resulted in a different change at the specificity residue (position 102) within the catalytic loop. However, the epistatic effects seen in the apicomplexan, *H. marismortui,* and *E. coli* dehydrogenases indicate that in general

507 position 102 is not solely responsible for the transition from MDH to LDH. In order for the
508 historical LDH mutations to confer pyruvate specificity, additional residues must be present to
509 provide a permissive background (**Figure 8**). Due to the presence of different sets of
510 permissive mutations, LDH activity has evolved from an MDH under epistatic constraints by a
511 different mechanism four separate times.

512 ***Large effect, gain-of-function mutation***

513 The evolution of AncLDH from AncMDH2 involves a shift in substrate specificity by
514 twelve orders-of-magnitude. Through the characterization of possible evolutionary
515 intermediates, we have found that just two mutations are responsible for the great majority of
516 this switch: the six-residue insertion and the Arg102Lys point mutation. Mutagenesis within the
517 insertion indicates that only a single position, Trp107f, contributes strongly to pyruvate activity
518 and specificity. Both the insertion and Arg102Lys have a large effect on preference for pyruvate
519 vs oxaloacetate, although by differentially affecting activity towards each substrate.
520 Incorporating the six-residue insertion into AncMDH2's substrate loop results in a 12,000-fold
521 gain in pyruvate activity with little effect on oxaloacetate activity (**Figure 6**). Conversely,
522 mutating Arg102 to Lys reduces oxaloacetate activity by more than 2,500-fold, with minimal
523 effect on pyruvate activity (**Figure 6**).

524 The apicomplexan LDH six-residue insertion is an exceptionally large gain-of-function
525 mutation: it enhances pyruvate activity by more than four orders of magnitude while barely
526 affecting oxaloacetate activity. In contrast, other well-studied mutations of large effect are often
527 predominantly deleterious towards one function while modestly enhancing another. The
528 textbook example of a gain-of-function mutation is Gln102Arg in *Bs*LDH, which causes a 10^7-
529 fold change in the enzyme's specificity (19). The Gln102Arg mutation reduces pyruvate activity
530 by more than 8,000-fold, while enhancing activity towards oxaloacetate by only 1,000-fold.
531 Another example is given by *E. coli* isocitrate dehydrogenase (IDH), where seven mutations are
532 necessary to switch the cofactor specificity from a 7,000-fold preference for NADP to a 200-fold

preference for NAD (51). Within this set of mutations, two reduce specificity for NADP by 6,000-fold, whereas the rest enhance NAD usage 200-fold. Thus, while mutations can have both deleterious and beneficial effects on different functions, the deleterious effects typically appear greater than enhancement.

In previous ancestral sequence reconstruction studies, mutations of large effect are in fact usually loss-of-function rather than gain-of-function (e.g., RNaseA (52), chymase (53), and glucocorticoid receptors (54-57)). In these studies, the modern proteins are generally specific for one substrate, whereas the ancestral proteins are promiscuous. Furthermore, the activity of the ancestral protein is comparable to the modern descendants. Therefore, these proteins specialized by accumulating deleterious mutations, with the modern, specialized activity being the "last function standing". For example, the ancestral glucocorticoid receptor binds three steroid hormones tightly (EC_{50} <10 nM for aldosterone, deoxycorticosterone, and cortisol), while the modern receptors bind only cortisol (EC_{50} ~100 nM) (57). Seven historical mutations produced the modern cortisol preference by completely eliminating aldosterone and deoxycorticosterone sensitivity yet reducing cortisol sensitivity only 50-fold. In other ancestral reconstruction studies, function-enhancing mutations have relatively minor effects, all less than a 50-fold gain in k_{cat}/K_M (37, 38, 58).

During the evolution of the malate and lactate dehydrogenase superfamily, pyruvate activity has converged multiple times despite strong constraints due to epistasis. While epistasis may constrain evolutionary options locally, there are nevertheless multiple ways to "skin the cat" in more distant regions of protein sequence space. The apicomplexan enzymes provide a clear example of neofunctionalization in protein evolution and thereby validate the plausibility of this particular mechanism of gene duplication. Specialization mechanisms may be more common, but the evolution of novel function does not require a promiscuous genesis.

Acknowledgements

This work was supported by the National Institutes of Health, NIH grants R01GM096053 and R01GM094468. The crystallographic data collection was conducted at the SIBYLS beamline at the Advanced Light Source (ALS), a national user facility operated by Lawrence Berkeley National Laboratory on behalf of the Department of Energy, Office of Basic Energy Sciences, through the Integrated Diffraction Analysis Technologies (IDAT) program, supported by DOE Office of Biological and Environmental Research. Additional support comes from the National Institute of Health project MINOS (R01GM105404). We would also like to thank Chris Miller, Phillip Steindel, and Catherine Theobald for critical commentary on the manuscript.

Materials and Methods

Modern sequences

Protein sequences used in the phylogenetic analyses were identified through searches of the non-redundant database (59) with the BLASTP algorithm (60) using selected query sequences. All sequences from these searches that returned BLASTP E-values $<10^{-7}$ were downloaded from NCBI (www.ncbi.nlm.nih.gov). Multiple complete apicomplexan genomes (41-43) were also searched for LDH and MDH homologs in order to fill out the apicomplexan portion of the tree (using a more lenient significance cutoff of E-values $<10^{-4}$). Redundant sequences, synthetic constructs, and sequences from PDB files were removed. To reduce phylogenetic complexity, sequences were curated based on character length and pairwise sequence identity within each dataset (as described below).

The dataset used for the construction of the non-redundant phylogeny (**Figure 3A**) was generated using four query sequences, UniProt IDs (61): MDHC_HUMAN, LDH_THEP1, MDHP_YEAST, and LDH6A_HUMAN. Multiple sequences were necessary to generate full coverage, due to the low sequence identity across the superfamily, which can be less than 20% between members. Sequences were removed if their character length was less than 280 or greater than 340. Limits were chosen to remove truncated/partial sequences and those featuring large insertions or terminal extensions. Sequences greater than 97% identical, determined by pairwise alignment within the dataset, were also removed. This level of identity provides a high level of detail within the tree while accelerating computational time by removing redundant taxa. The final dataset contains 1844 taxa.

Residue numbering in the text is based on the dogfish LDH convention (20) for consistency with previous work.

590 ***Primary Phylogeny Construction***

591 A multiple sequence alignment of this dataset was generated using the program
592 MUSCLE (62). A maximum likelihood (ML) phylogenetic tree was inferred with PhyML 3.0
593 (63) using the LG substitution matrix (64) and estimating the gamma parameter (12 categories)
594 and empirical amino acid frequencies. The starting tree was generated by Neighbor-Joining
595 (BIONJ) and searched by Nearest Neighbor Interchange (NNI); tree topology, branch lengths,
596 and rate parameters were optimized. Branch supports were estimated with the approximate
597 likelihood ratio test (aLRT), as implemented in PhyML, represented as either the raw aLRT
598 statistic (roughly > 8 is considered highly significant) or the confidence level that the clade is
599 correct (65).

600 ***Phylogeny Rooting***

601 The outgroup for rooting the L/MDH phylogeny was identified through a profile analysis
602 of the Rossmann fold (66), based on a method used for OB folds and SH3 domains (67). All
603 structurally characterized Rossmann folds with 40% or less sequence identity were identified
604 from ASTRAL SCOP 1.73 protein domain sequence database (68). Each of the 193 domains
605 identified was searched against the SwissProt database (69) using BLASTP. A multiple
606 sequence alignment for each query and SwissProt sequences with BLASTP E-values $<10^{-10}$ was
607 created using MUSCLE. Each alignment was cropped to the limits of the original query.
608 COMPASS (70) was then used to generate an all-against-all scoring matrix for the 193 multiple
609 sequence alignments. The E-values generated by COMPASS were converted to evolutionary
610 distances as described in Theobald & Wuttke (67). A weighted least-squares phylogenetic
611 analysis of the distance matrix was performed using PAUP (71). First order taxon jackknifing
612 (72, 73) was used to determine the robustness of tree topology, with a consensus tree calculated
613 from all analyses.

Rossmann fold domains from α- and β-glucosidases and aspartate dehydrogenases (AspDH) were identified from the profile-profile analysis as grouping with the Rossmann fold domain from L/MDHs. An L/MDH dataset was constructed for use with the outgroup to create a rooted phylogeny. This dataset was generated by querying four sequences, UniProt IDs: MDHC_HUMAN, LDH_THEP1, MDHP_YEAST, and LDH6A_HUMAN, against the SwissProt database using BLASTP. All sequences from these searches that returned BLASTP E-values $<10^{-7}$ were downloaded from NCBI (www.ncbi.nlm.nih.gov). Redundant sequences, synthetic constructs, and sequences from PDB files were removed. Also, four taxa identified as ubiquitin-conjugating enzymes were removed due to sequence length. This SwissProt L/MDH dataset contained 595 taxa.

An outgroup dataset was constructed by querying three sequences, UniProt IDs: LICH_BACSU, AGAL_THEMA, and ASPD_THEMA, against the SwissProt database using BLASTP. All sequences from these searches that returned BLASTP E-values $<10^{-7}$ were downloaded from NCBI (www.ncbi.nlm.nih.gov). Redundant sequences, synthetic constructs, and sequences from PDB files were removed. The outgroup dataset contained 62 taxa. The SwissProt LDH, MDH, AspDH, and glucosidase datasets were combined and a multiple sequence alignment was generated using the program MUSCLE. The C-terminal domain of the glucosidases and AspDHs were removed from the MUSCLE alignment. A ML phylogenetic tree was inferred from the alignment with PhyML using the LG substitution matrix (74) with the gamma parameter estimated over 10 categories, no invariant sites, and estimating empirical amino acid frequencies. The initial tree was obtained by BIONJ and searched by NNI; tree topology, branch lengths, and rate parameters were optimized. Robustness of root positioning was evaluated with two truncated alignments, one with the LDH and "LDH-like" MDH sequences removed and the other with the cytosolic and mitochondrial MDH sequences removed. Truncated alignments were input to PhyML for phylogenetic analysis using the parameters described above.

Alternative phylogeny construction

The dataset for the alternative phylogeny (used in reconstructing alternative ancestors) is smaller and focused on apicomplexan taxa. It was generated by BLASTP searches with four query sequences, UniProt IDs: MDHC_PIG, Q76NM3_PLAF7, C6KT25_PLAF, and MDH_WOLPM for full coverage of the superfamily. Sequences were removed if their length was less than 290 or greater than 340. The dataset was culled to 60% identity, but the apicomplexan clade was filled back to 97% identity to gain resolution within the clade of interest. The final dataset contained 277 taxa. A multiple sequence alignment of this dataset was produced using the program MUSCLE. The ML tree was inferred with PhyML 3.0 using the LG substitution matrix and estimating the gamma parameter (12 categories) and empirical amino acid frequencies. The starting tree was generated by Neighbor-Joining (BIONJ) and searched by Nearest Neighbor Interchange (NNI); tree topology, branch lengths, and rate parameters were optimized.

Ancestral Sequence Reconstruction

Sequences at internal nodes in phylogenies were inferred using the *codeml* program from the PAML software package (75). Posterior amino acid probabilities at each site were calculated using the LG substitution matrix, given the ML tree generated by PhyML. The initial ancestral reconstruction assumed the background amino acid frequencies implicit in the LG matrix, while the alternative reconstruction estimated background frequencies from the sequence alignment of the alternative dataset. N-/C-termini of ancestral sequences were modified manually to match those of the closest modern sequence (determined by branch length).

Plasmid Construction and Mutation

Escherichia coli codon-optimized coding sequences were constructed for the *Plasmodium falciparum* MDH (gi#: 86171227), *Cryptosporidium parvum* MDH (gi#:

32765705), *Toxoplasma gondii* LDH1 (gi#: 237837615), *Toxoplasma gondii* LDH2 (gi#: 2497625), *Rickettsia bellii* MDH (gi#: 91205459), and ancestrally inferred protein sequences. These coding sequences were synthesized and subcloned into pET-24a, bypassing the N-terminal T7-tag but using the C-terminal 6xHis-tag. *Pf*LDH (gi#: 124513266) with six His residues added to the C-terminus was synthesized and subcloned into pET-11b. All gene synthesis and subcloning was performed by Genscript (Piscataway, NJ). All point mutations were made using the QuikChange Lightning kit from Agilent (Santa Clara, CA) and synthesized primers from IDT (Coralville, IA).

Protein Expression and Purification

Plasmids were transformed in BL21 DE3 (pLysS) *E. coli* cells (Invitrogen, Grand Island, NY) for expression. Cells were grown at 37°C with 225 rpm agitation in 2xYT media supplemented with 30 mM potassium phosphate, pH 7.8 and 0.1% (w/v) glucose. Once cultures reached an OD_{600} between 0.5-0.8, cells were induced with 0.5 mM IPTG for 4 hours. Cells were collected by centrifugation at 10,000*xg* for 15 minutes and stored at -80°C.

Cell pellets were thawed on ice, releasing lysozyme produced by the pLysS plasmid from within the cells, and resuspended in 15 mL lysis buffer (50 mM NaH_2PO_4, pH 8.0, 300 mM NaCl, 10 mM Imidazole) with 375 units of Pierce Universal Nuclease (Thermo Scientific, Rockford, IL) per 1.5 L of culture. Once homogeneously resuspended, lysate was sonicated on ice at 35% amplitude (30 sec ON, 20 sec OFF, 2 min total). Insoluble cell debris was separated by centrifugation at 18,000*xg* for 20 min.

Proteins were purified by nickel affinity chromatography. Clarified lysate was applied to a 5 mL HisTrap FF column (GE Healthcare, Piscataway, NJ) and eluted via an imidazole gradient from 10 mM to 500 mM on an AKTA Prime (GE Healthcare, Piscataway, NJ). Fractions were analyzed by SDS-PAGE, pooled, and concentrated using Amicon Ultracel-10K centrifugal filters (Millipore, Billerica, MA). Finally, proteins were desalted into 50 mM Tris, pH

7.4, 100 mM NaCl, 0.1 mM EDTA and 0.01% azide by either PD10 column (GE Healthcare, Piscataway, NJ) or gel filtration over a HiPrep 16/60 Sephacryl S-200 HR column (GE Healthcare, Piscataway, NJ) on an AKTA Purifier (GE Healthcare, Piscataway, NJ). Enzyme concentrations were determined by absorbance at 280 nm, using extinction coefficients and molecular weights calculated by ExPASy's ProtParam tool (http://web.expasy.org/protparam/).

Steady-state Kinetic Assays

Enzymatic reduction of pyruvate and oxaloacetate was monitored at 25°C by following the decrease in absorbance at 340 nm due to NADH oxidation on a Cary 100 Bio (Agilent, Santa Clara, CA) in 50 mM Tris, pH 7.5, 50 mM KCl. All substrates were purchased from Sigma-Aldrich (St. Louis, MO). NADH concentration was held constant at 200 μM while pyruvate/oxaloacetate concentrations were titrated. Enzyme concentrations ranged from 0.28 nM to 2.8 μM, depending on enzyme activity for a particular substrate. All experiments used 1-cm path-length quartz cuvettes with 500 μL final volume of reaction mixture.

Kinetic parameters were estimated by chi-squared fitting to either the Michaelis-Menton equation ($v/[E]_t = k_{cat}\,[S]/(K_M + [S])$) or a substrate inhibition equation ($v/[E]_t = k_{cat}\,[S]/(K_M + [S] + [S]^2/K_i)$) using the KaleidaGraph software. Three datasets were fit using a modified substrate inhibition equation with $K_M = K_i$ for identifiability and to prevent the K_i being less than K_M. These datasets were: AncMDH2-INS-59Mut oxaloacetate and AncMDH2-R102Q for both oxaloacetate and pyruvate.

Aqueous oxaloacetate spontaneously decarboxylates to pyruvate at 25 °C and neutral pH at a rate of ~3×10^{-5} s^{-1} (approximately 10% per hour) (76). As a result, oxaloacetate preparations contain appreciable pyruvate contamination (approximately 1-3% from Sigma-Aldrich, depending on batch) and must be handled with care. All oxaloacetate stock solutions were made fresh before each assay and kept on ice to keep decarboxylation to a minimum. For enzymes with low pyruvate activity, the oxaloacetate decarboxylation has a negligible affect on

measured rates. However, enzymes with appreciable pyruvate activity can display an apparent, artifactual oxaloacetate activity that is due to pyruvate contamination (19, 27, 77). In this work, seven such proteins are *Pf*LDH, *Pf*LDH-K102R, *Tg*LDH1, *Tg*LDH2, AncLDH, AncLDH*, and AncMDH2-INS-R102K. For these proteins, oxaloacetate activity was assayed at high enzyme concentration (600 nM to 1 μM), resulting in a biphasic ΔA_{340} trace with an initial burst in which pyruvate is rapidly consumed followed by a slower linear phase representing oxaloacetate reduction. The post-burst (slow) phase of the ΔA_{340} trace was used to quantify the oxaloacetate catalytic rate (19, 77). This procedure controls for the standing pyruvate contamination but does not account for the relatively slow spontaneous decarboxylation during the assay. Hence, the oxaloacetate k_{cat}/K_m values for the seven enzymes with high pyruvate activity should be considered upper limits on the true oxaloacetate activity. The low or negligible oxaloacetate activities of these seven enzymes were further confirmed by (1) undetectable malate/NAD+ reactions in spectroscopic steady state enzyme assays, and (2) the absence of malate product as determined from 1D proton NMR (3 μM enzyme, 5 mM oxaloacetate, 5 mM NADH in $NaCl/P_i/D_2O$ pH 7.5 over four hour reaction) (27) .

Protein Crystallization

Crystallization trials were conducted by hanging-drop vapor-diffusion at room temperature using Crystal Screen™ and Crystal Screen 2™ from Hampton Research (Aliso Viejo, CA). Drops consisting of 2 μL reservoir solution and 2 μL protein stock were equilibrated against 1 mL of reservoir solution. Crystals of the ancestral proteins were identified from condition #43 of Crystal Screen™ (30% (w/v) polyethylene glycol 1,500) and further refined by adding 0.1 M sodium HEPES.

Crystals of the ternary complexes were grown at room temperature by hanging-drop vapor-diffusion with 4 μL drops of 1:1 precipitating buffer:protein. Ancestral malate dehydrogenase (AncMDH2, 25 mg/mL) was co-crystallized with 2 mM oxamate/NADH in 35%

740 (w/v) PEG-1500, 0.1 M sodium HEPES, pH 7.5 and with 2 mM L-lactate/NADH in 30% (w/v)
741 PEG-1500, 0.1 M sodium HEPES, pH 7.3. AncMDH2 with insertion (AncMDH2-INS, 18
742 mg/mL) was co-crystallized with 1 mM oxamate/NADH and 1 mM L-lactate/NADH in 25%
743 (w/v) PEG-1500, 0.1 M sodium HEPES, pH 8.1. Alternative ancestral lactate dehydrogenase
744 (AncLDH*, 20 mg/mL) was co-crystallized with 2 mM oxamate/NADH and 2 mM L-
745 lactate/NADH in 20% (w/v) PEG-1500, 0.1 M sodium HEPES, pH 7.5. *Pf* LDH_W107fA
746 (20mg/mL) was co-crystallized with 1.2 mM oxamate/2 mM NADH in 22% (w/v) PEG-1000.
747 All crystals were cryoprotected with a 30% (w/v) dextrose solution (15 mg dextrose
748 dissolved in 50 μL reservoir solution). Crystals were harvested from the drop, soaked in 15%
749 (w/v) dextrose solution for 3 minutes, transferred to the 30% solution, and flash-frozen
750 immediately in liquid N_2.

751 ***Structure Determination***

752 Diffraction datasets were collected at the SIBYLS beamline (12.3.1, Lawrence Berkeley
753 National Laboratory, Berkeley, CA). All datasets were indexed, integrated, and scaled with
754 XDS/XSCALE (78). Structures were solved by molecular replacement using AutoMR in
755 PHENIX (79). Homology models for the AncMDH2 and AncLDH* datasets were generated by
756 the Phyre2 server (80). The AncMDH2 homology model was based on the structure for
757 *Cryptosporidium parvum* MDH (PDB entry: 2hjr, 62% sequence identity, (47)), while the
758 model for AncLDH* was based on the *Toxoplasma gondii* LDH1 structure (PDB entry: 1pzf, 65%
759 sequence identity, (26)). AncMDH2-INS datasets were solved using the AncMDH2 structure as
760 the model. *Pf* LDH_W107fA dataset was solved using the *P. falciparum* LDH structure (pdb id:
761 1t2d) structure as a model. All models were improved by rounds of manual building in Coot (81)
762 and refinement by phenix.refine in PHENIX. Model quality of all structures was validated with
763 MolProbity (82, 83) in PHENIX. All structural alignments were generated using THESEUS (84).
764 Structure images were rendered with PyMOL.

Figure legends

Figure 1. Schematic of M/LDH superfamily active site and catalytic mechanism. MDH reduces oxaloacetate to malate, in which the R-group is a methylene carboxylate. LDH reduces pyruvate to lactate, in which the R-group is a methyl. Key conserved active site residues are shown in black; substrate is shown in blue. The oxidized 2-ketoacid form of the substrate is at left; the reduced 2-hydroxy acid form is shown at right. The R-group of the substrate interacts with Arg102 in MDHs and Gln102 in LDHs. Both Arg109 and position 102 are found in the "specificity loop" that closes over the active site.

Figure 2. Apicomplexan M/LDH active sites. Structures of *Cp*MDH (blue, PDB ID: 2hjr) and *Pf*LDH (vermilion, PDB ID: 1t2d) superposed using THESEUS. The ligands (oxalate and NAD^+) are from 1t2d and colored WHITE. Side chains of important residues are shown as sticks and the six-residue insert of *Pf*LDH is highlighted in YELLOW. Note how the *Pf*LDH Trp107f overlays Arg102 from *Cp*MDH. Residues in the insertion are labeled using numbers and letters to maintain consistency with homologous positions in the dogfish LDH.

Figure 3. Phylogeny of M/LDH superfamily. A. 1844 taxa. The tree is colored according to function (LDH – vermilion; MDH – blue; HicDH – moss). The N-terminal Rossmann-fold of glucosidases and aspartate dehydrogenases (AspDHs) was used to root the phylogeny. Numbers highlight convergent events of LDH evolution from MDHs: 1 - Canonical LDHs, 2 - Trichomonad LDHs, and 3,4 - apicomplexan LDHs. The shaded clades have aLRT supports of 42, 57, 75 and 117, respectively (roughly, an aLRT > 8 is considered highly significant (65)). **B. Apicomplexan M/LDH Clade.** A close-up of the apicomplexan portion of the phylogeny in **A**, similarly colored by function. aLRT supports for each group: α-proteobacteria MDHs, 15; apicomplexan LDHs, 11; *Plasmodium* LDHs, 333; *Cryptosporidium* MDHs, 54; *Cryptosporidium* LDHs, 202. Ancestral reconstructed proteins are labeled at internal nodes

(AncMDH1, AncMDH2, AncMDH3, AncLDH). The focus of the present work is the gene duplication at node 3.

Figure 4. Specificity switching in apicomplexan M/LDHs. Blue horizontal bars (left) quantify activity towards oxaloacetate; red horizontal bars (right) quantify activity towards pyruvate. Error bars are shown as small black brackets and represent 1 SD from triplicate measurements. INS refers to the presence of the six-residue insertion from *Pf*LDH, DEL refers to the removal of the six-residue insertion. Relative specificity (RS) is the ratio of k_{cat}/K_M for pyruvate vs oxaloacetate, with positive $\log_{10}$(RS) representing a preference for pyruvate and negative $\log_{10}$(RS) representing a preference for oxaloacetate. All logarithms are base 10.

Figure 5. Evolution of novel LDHs in Apicomplexa. The y-axis of the bar graphs is $\log(k_{cat}/K_M)$, with oxaloacetate in blue and pyruvate in vermilion. Blue vertical bars (left) quantify activity of the given enzyme towards oxaloacetate; red vertical bars (right) quantify activity towards pyruvate. *Rb*MDH is a representative α-proteobacterial MDH from *Rickettsia bellii*. *T. gondii* has two LDH proteins (TgLDH1 and TgLDH2), each expressed at different stages of the life cycle (25).

Figure 6. Specificity switching in ancestral MDH2. INS refers to the reconstructed six-residue insertion from AncLDH. 59Mut is described in the text. Relative specificity (RS) is described in legend of **Figure 4**.

Figure 7. Ancestral and modern dehydrogenase structures. A. Superposition of *Cp*MDH and AncMDH2. Superposition of AncMDH2 structure (blue) and *Cp*MDH (aquamarine, PDB ID: 2hjr). Ligands from AncMDH2 are shown in gray; ligands from *Cp*MDH are in white. **B. Active site detail of *Cp* MDH and AncMDH2.** Side chains of catalytic

residues highlighted as sticks. **C. Superposition of apicomplexan LDHs and AncLDH*.** Superposition of AncLDH* structure (vermilion) and four apicomplexan LDHs (deep olive, *Pf*LDH (PDB ID: 1t2d), *Plasmodium berghei* (*Pb*) LDH (PDB ID: 1oc4), *Tg*LDH1 (PDB ID: 1pzh), *Tg*LDH2 (PDB ID: 1sow)). Ligands from AncLDH* are shown in gray, ligands from apicomplexan LDHs are in white. The "opposing loop" and residues 236 and 246 (discussed in text) are highlighted in cyan. **D. Active site detail of apicomplexan LDHs and AncLDH*.** Side chains of catalytic residues highlighted as sticks. **E. Superposition of ancestral dehydrogenases.** Superposition of AncMDH2 (blue), AncLDH* (vermilion), and AncMDH2-INS (magenta). Ligands are shown in gray. **F. Active site detail of ancestral dehydrogenases.** Side chains of catalytic residues highlighted as sticks.

Figure 8. Alternative LDH mutations in AncMDH2. Relative specificity (RS) is described in legend of **Figure 4**.

Figure 1-figure supplement 1. Fold architecture in the LDH and MDH superfamily. At left is *Cp*MDH (blue, PDB ID: 2hjr), at right is *Pf*LDH (vermilion and olive, PDB ID: 1t2d). The Rossmann fold domain, which binds the NADH cofactor, is show as light blue in *Cp*MDH and vermilion in *Pf*LDH. The active site is found at the interface of the two domains. In *Cp*MDH, the "opposing loop" is highlighted in yellow (see text). In *Pf*LDH, the six-residue insertion is highlighted in yellow.

Figure 2-figure supplement 1. Sequence alignment of the specificity loop from apicomplexan M/LDHs with ancestral sequences.

Figure 2-figure supplement 2. Alanine scanning of *Pf*LDH specificity loop. Logarithm of pyruvate k_{cat}/K_M of *Pf*LDH and each mutant. Labels on x-axis describe the mutation tested in the WT *Pf*LDH background.

Figure 2-figure supplement 3. Crystal structure of *Pf*LDH-W107fA mutant. A. Crystal lattice of the W107fA mutant (left) compared to the WT *Pf*LDH (right). B. Superposition of the WT *Pf*LDH (olive) and the W107fA mutant (vermilion). The structures are highly similar throughout, expect for the active site loop (at top), which is closed in the WT and partially disordered and open in the mutant.

Figure 3-figure supplement 1. Phylogeny of M/LDH superfamily. Same phylogeny as **Figure 3A** with select branch supports shown (aLRT supports).

Figure 3-figure supplement 2. Phylogeny of M/LDH superfamily. Same phylogeny as **Figure 3A**. The tree is colored by domain of life (eubacterial – vermilion; eukaryotic – blue; archeal – magenta).

Figure 3-figure supplement 3. Apicomplexan M/LDH Clade. Same phylogeny as **Figure 3B** with aLRT branch supports and clades shown in full detail.

Figure 5-figure supplement1. Sequence identity of ancestral and modern proteins.

Figure 6-figure supplement 1-6. Histograms of ancestral reconstructions. Reconstructed residues binned according to posterior probability (PP) of the predicted residue.

Figure 6-figure supplement 7. Alternative ancestral enzymes. INS refers to the reconstructed six amino acid insertion from AncLDH*. 58Mut refers to remaining residue differences between AncMDH* and AncLDH* that are not R102K or INS. Relative specificity (RS) is described in legend of **Figure 4**.

Figure 7-figure supplement 1. Statistics table for AncMDH2 structures. Statistics for highest resolution shell are shown in parentheses.

Figure 7-figure supplement 2. Statistics table for AncLDH* structures. Statistics for highest resolution shell are shown in parentheses.

Figure 7-figure supplement 3. Statistics table for AncMDH2-INS structures. Statistics for highest resolution shell are shown in parentheses.

Figure 2 -source data 1. Source data for figure supplement 2. Kinetic parameters for PfLDH alanine-scan.

Figure 4-source data 1. Kinetic parameters for modern constructs.

Figure 5-source data 1. Kinetic parameters for ancestral/modern phylogeny.

Figure 6-source data 1. Kinetic parameters for ancestral specificity switch mutants.

Figure 6 -source data 2. Source data for figure supplement 7. Kinetic parameters for alternative ancestral proteins.

Figure 8-source data 1. Kinetic parameters for specificity residue mutants.

Supplementary File 1. Sequences, alignments, and trees. Alignments and tree files for both the original (**Figure 3**) and the alternative phylogeny. Alignment for **Figure 5-figure supplement 1**. Ancestral FASTA files and posterior probabilities for each ancestral sequence (parsed in **Figure 6-figure supplement 1-6**.

Supplementary File 2. Molecular weights and extinction coefficients. ExPASy calculated molecular weights and extinction coefficients for all proteins used within this study.

References

1. Douzery EJP, Snell EA, Bapteste E, Delsuc F, Philippe H. The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils? Proceedings of the National Academy of Sciences of the United States of America. 2004 Oct 26;101(43):15386-91.
2. Madern D. Molecular Evolution Within the L-Malate and L-Lactate Dehydrogenase Super-Family. Journal of Molecular Evolution. 2002;54(6):825-40.
3. Zhu G, Keithly JS. Alpha-proteobacterial relationship of apicomplexan lactate and malate dehydrogenases. The Journal of eukaryotic microbiology. 2002 May-Jun;49(3):255-61.
4. Golding GB, Dean AM. The structural basis of molecular adaptation. Molecular biology and evolution. 1998 Apr;15(4):355-69.
5. Royer RE, Deck LM, Campos NM, Hunsaker LA, Vanderjagt DL. Biologically-Active Derivatives of Gossypol - Synthesis and Antimalarial Activities of Peri-Acylated Gossylic Nitriles. Journal of Medicinal Chemistry. 1986 Sep;29(9):1799-801.
6. Cameron A, Read J, Tranter R, Winter VJ, Sessions RB, Brady RL, et al. Identification and activity of a series of azole-based compounds with lactate dehydrogenase-directed anti-malarial activity. Journal of Biological Chemistry. 2004 Jul 23;279(30):31429-39.
7. Conners R, Schambach F, Read J, Cameron A, Sessions RB, Vivas L, et al. Mapping the binding site for gossypol-like inhibitors of Plasmodium falciparum lactate dehydrogenase. Molecular and Biochemical Parasitology. 2005 Aug;142(2):137-48.
8. Gomez MS, Piper RC, Hunsaker LA, Royer RE, Deck LM, Makler MT, et al. Substrate and cofactor specificity and selective inhibition of lactate dehydrogenase from the malarial parasite P-falciparum. Molecular and Biochemical Parasitology. 1997 Dec 1;90(1):235-46.
9. Read JA, Wilkinson KW, Tranter R, Sessions RB, Brady RL. Chloroquine binds in the cofactor binding site of Plasmodium falciparum lactate dehydrogenase. Journal of Biological Chemistry. 1999 Apr 9;274(15):10213-8.
10. Rossmann MG, Liljas, A., Branden, C-I., Banaszak, L. J. Evolutionary and Structural Relationships among Dehydrogenases. In: Boyer PD, editor. The Enzymes. 3rd ed: Academic Press; 1975. p. 61-102.
11. Birktoft JJ, Banaszak LJ. The presence of a histidine-aspartic acid pair in the active site of 2-hydroxyacid dehydrogenases. X-ray refinement of cytoplasmic malate dehydrogenase. J Biol Chem. 1983 Jan 10;258(1):472-82.
12. Clarke AR, Wigley DB, Chia WN, Barstow D, Atkinson T, Holbrook JJ. Site-directed mutagenesis reveals role of mobile arginine residue in lactate dehydrogenase catalysis. Nature. 1986 Dec 18-31;324(6098):699-702.
13. Clarke AR, Wilks HM, Barstow DA, Atkinson T, Chia WN, Holbrook JJ. An investigation of the contribution made by the carboxylate group of an active site histidine-aspartate couple to binding and catalysis in lactate dehydrogenase. Biochemistry. 1988 Mar 8;27(5):1617-22.
14. Hart KW, Clarke AR, Wigley DB, Chia WN, Barstow DA, Atkinson T, et al. The importance of arginine 171 in substrate binding by Bacillus stearothermophilus lactate dehydrogenase. Biochem Biophys Res Commun. 1987 Jul 15;146(1):346-53.
15. Hart KW, Clarke AR, Wigley DB, Waldman AD, Chia WN, Barstow DA, et al. A strong carboxylate-arginine interaction is important in substrate orientation and recognition in lactate dehydrogenase. Biochim Biophys Acta. 1987 Aug 21;914(3):294-8.
16. Waldman ADB, Hart KW, Clarke AR, Wigley DB, Barstow DA, Atkinson T, et al. The Use of a Genetically Engineered Tryptophan to Identify the Movement of a Domain of B-Stearothermophilus Lactate-Dehydrogenase with the Process Which Limits the Steady-State Turnover of the Enzyme. Biochemical and Biophysical Research Communications. 1988 Jan 29;150(2):752-9.

17. Bzik DJ, Fox BA, Gonyer K. Expression of Plasmodium-Falciparum Lactate-Dehydrogenase in Escherichia-Coli. Molecular and Biochemical Parasitology. 1993 May;59(1):155-66.
18. Dunn CR, Banfield MJ, Barker JJ, Higham CW, Moreton KM, Turgut-Balik D, et al. The structure of lactate dehydrogenase from Plasmodium falciparum reveals a new target for anti-malarial design. Nature structural biology. 1996 Nov;3(11):912-5.
19. Wilks HM, Hart KW, Feeney R, Dunn CR, Muirhead H, Chia WN, et al. A specific, highly active malate dehydrogenase by redesign of a lactate dehydrogenase framework. Science. 1988 Dec 16;242(4885):1541-4.
20. Eventoff W, Rossmann MG, Taylor SS, Torff HJ, Meyer H, Keil W, et al. Structural Adaptations of Lactate-Dehydrogenase Isozymes. Proceedings of the National Academy of Sciences of the United States of America. 1977;74(7):2677-81.
21. Chapman AD, Cortes A, Dafforn TR, Clarke AR, Brady RL. Structural basis of substrate specificity in malate dehydrogenases: crystal structure of a ternary complex of porcine cytoplasmic malate dehydrogenase, alpha-ketomalonate and tetrahydoNAD. J Mol Biol. 1999 Jan 15;285(2):703-12.
22. Cendrin F, Chroboczek J, Zaccai G, Eisenberg H, Mevarech M. Cloning, sequencing, and expression in Escherichia coli of the gene coding for malate dehydrogenase of the extremely halophilic archaebacterium Haloarcula marismortui. Biochemistry. 1993 Apr 27;32(16):4308-13.
23. Nicholls DJ, Miller J, Scawen MD, Clarke AR, Holbrook JJ, Atkinson T, et al. The importance of arginine 102 for the substrate specificity of Escherichia coli malate dehydrogenase. Biochem Biophys Res Commun. 1992 Dec 15;189(2):1057-62.
24. Brown WM, Yowell CA, Hoard A, Jagt TAV, Hunsaker LA, Deck LM, et al. Comparative structural analysis and kinetic properties of lactate dehydrogenases from the four species of human malarial parasites. Biochemistry. 2004 May 25;43(20):6219-29.
25. Dando C, Schroeder ER, Hunsaker LA, Deck LM, Royer RE, Zhou XL, et al. The kinetic properties and sensitivities to inhibitors of lactate dehydrogenases (LDH1 and LDH2) from Toxoplasma gondii: comparisons with pLDH from Plasmodium falciparum. Molecular and Biochemical Parasitology. 2001 Nov;118(1):23-32.
26. Kavanagh KL, Elling RA, Wilson DK. Structure of Toxoplasma gondii LDH1: Active-site differences from human lactate dehydrogenases and the structural basis for efficient APAD(+) use. Biochemistry. 2004 Feb 3;43(4):879-89.
27. Shoemark DK, Cliff MJ, Sessions RB, Clarke AR. Enzymatic properties of the lactate dehydrogenase enzyme from Plasmodium falciparum. Febs Journal. 2007 Jun;274(11):2738-48.
28. Winter VJ, Cameron A, Tranter R, Sessions RB, Brady RL. Crystal structure of Plasmodium berghei lactate dehydrogenase indicates the unique structural differences of these enzymes are shared across the Plasmodium genus. Molecular and Biochemical Parasitology. 2003 Sep;131(1):1-10.
29. Innan H, Kondrashov F. The evolution of gene duplications: classifying and distinguishing between models. Nature Reviews Genetics. 2010 Feb;11(2):97-108.
30. Ohno S. Evolution by Gene Duplication. New York: Springer; 1970.
31. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000 Nov 10;290(5494):1151-5.
32. Walsh JB. How Often Do Duplicated Genes Evolve New Functions. Genetics. 1995 Jan;139(1):421-8.
33. Conant GC, Wolfe KH. Turning a hobby into a job: How duplicated genes find new functions. Nature Reviews Genetics. 2008 Dec;9(12):938-50.
34. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 1999 Apr;151(4):1531-45.
35. Soskine M, Tawfik DS. Mutational effects and the evolution of new protein functions. Nat Rev Genet. 2010 Aug;11(8):572-82.

36. Bridgham JT, Brown JE, Rodriguez-Mari A, Catchen JM, Thornton JW. Evolution of a new function by degenerative mutation in cephalochordate steroid receptors. PLoS Genet. 2008;4(9):e1000191.
37. Voordeckers K, Brown CA, Vanneste K, van der Zande E, Voet A, Maere S, et al. Reconstruction of Ancestral Metabolic Enzymes Reveals Molecular Mechanisms Underlying Evolutionary Innovation through Gene Duplication. Plos Biology. 2012 Dec;10(12).
38. Zhang JZ, Rosenberg HF. Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. Proceedings of the National Academy of Sciences of the United States of America. 2002 Apr 16;99(8):5486-91.
39. Madern D, Cai XM, Abrahamsen MS, Zhu G. Evolution of Cryptosporidium parvum lactate dehydrogenase from malate dehydrogenase by a very recent event of gene duplication. Molecular biology and evolution. 2004 Mar;21(3):489-97.
40. Wu G, Fiser A, ter Kuile B, Sali A, Muller M. Convergent evolution of Trichomonas vaginalis lactate dehydrogenase from malate dehydrogenase. Proc Natl Acad Sci U S A. 1999 May 25;96(11):6285-90.
41. Gajria B, Bahl A, Brestelli J, Dommer J, Fischer S, Gao X, et al. ToxoDB: an integrated Toxoplasma gondii database resource. Nucleic acids research. 2008 Jan;36(Database issue):D553-6.
42. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, et al. PlasmoDB: a functional genomic database for malaria parasites. Nucleic acids research. 2009 Jan;37(Database issue):D539-43.
43. Heiges M, Wang H, Robinson E, Aurrecoechea C, Gao X, Kaluskar N, et al. CryptoDB: a Cryptosporidium bioinformatics resource update. Nucleic acids research. 2006 Jan 1;34(Database issue):D419-22.
44. Templeton TJ, Enomoto S, Chen WJ, Huang CG, Lancto CA, Abrahamsen MS, et al. A genome-sequence survey for Ascogregarina taiwanensis supports evolutionary affiliation but metabolic diversity between a Gregarine and Cryptosporidium. Molecular biology and evolution. 2009 Feb;27(2):235-48.
45. Feng ZP, Zhang X, Han P, Arora N, Anders RF, Norton RS. Abundance of intrinsically unstructured proteins in P. falciparum and other apicomplexan parasite proteomes. Mol Biochem Parasitol. 2006 Dec;150(2):256-67.
46. Kissinger JC, DeBarry J. Genome cartography: charting the apicomplexan genome. Trends Parasitol. 2011 Aug;27(8):345-54.
47. Vedadi M, Lew J, Artz J, Amani M, Zhao Y, Dong AP, et al. Genome-scale protein expression and structural biology of Plasmodium falciparum and related Apicomplexan organisms. Molecular and Biochemical Parasitology. 2007 Jan;151(1):100-10.
48. Garrett R, Grisham, C.M. Biochemistry. 3rd ed. Belmont, CA: Thomson Brooks/Cole; 2005.
49. Des Marais DL, Rausher MD. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature. 2008 Aug 7;454(7205):762-U85.
50. Harms MJ, Thornton JW. Analyzing protein structure and function using ancestral gene reconstruction. Curr Opin Struct Biol. 2010 Jun;20(3):360-6.
51. Chen RD, Greer A, Dean AM. A Highly-Active Decarboxylating Dehydrogenase with Rationally Inverted Coenzyme Specificity. Proceedings of the National Academy of Sciences of the United States of America. 1995 Dec 5;92(25):11666-70.
52. Jermann TM, Opitz JG, Stackhouse J, Benner SA. RECONSTRUCTING THE EVOLUTIONARY HISTORY OF THE ARTIODACTYL RIBONUCLEASE SUPERFAMILY. Nature. 1995 Mar;374(6517):57-9.
53. Wouters MA, Liu K, Riek P, Husain A. A despecialization step underlying evolution of a family of serine proteases. Molecular Cell. 2003 Aug;12(2):343-54.

54. Bridgham JT, Carroll SM, Thornton JW. Evolution of hormone-receptor complexity by molecular exploitation. Science. 2006 Apr 7;312(5770):97-101.
55. Carroll SM, Bridgham JT, Thornton JW. Evolution of Hormone Signaling in Elasmobranchs by Exploitation of Promiscuous Receptors. Molecular biology and evolution. 2008 Dec;25(12):2643-52.
56. Carroll SM, Ortlund EA, Thornton JW. Mechanisms for the Evolution of a Derived Function in the Ancestral Glucocorticoid Receptor. Plos Genetics. 2011 Jun;7(6).
57. Ortlund EA, Bridgham JT, Redinbo MR, Thornton JW. Crystal structure of an ancient protein: Evolution by conformational epistasis. Science. 2007 Sep 14;317(5844):1544-8.
58. Risso VA, Gavira JA, Mejia-Carmona DF, Gaucher EA, Sanchez-Ruiz JM. Hyperstability and Substrate Promiscuity in Laboratory Resurrections of Precambrian beta-Lactamases. Journal of the American Chemical Society. 2013 Feb 27;135(8):2899-902.
59. Pruitt KD, Tatusova T, Klimke W, Maglott DR. NCBI Reference Sequences: current status, policy and new initiatives. Nucleic acids research. 2009 Jan;37(Database issue):D32-6.
60. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10.
61. Consortium TU. Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic acids research. 2013 Jan;41(Database issue):D43-7.
62. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 2004;32(5):1792-7.
63. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic biology. 2010 May;59(3):307-21.
64. Le SQ, Gascuel O. An improved general amino acid replacement matrix. Molecular biology and evolution. 2008 Jul;25(7):1307-20.
65. Anisimova M, Gascuel O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Systematic biology. 2006 Aug;55(4):539-52.
66. Rao ST, Rossmann MG. Comparison of super-secondary structures in proteins. J Mol Biol. 1973 May 15;76(2):241-56.
67. Theobald DL, Wuttke DS. Divergent evolution within protein superfolds inferred from profile-based phylogenetics. J Mol Biol. 2005 Dec 2;354(3):722-37.
68. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, et al. The ASTRAL Compendium in 2004. Nucleic acids research. 2004 Jan 1;32(Database issue):D189-92.
69. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic acids research. 2003 Jan 1;31(1):365-70.
70. Sadreyev RI, Baker D, Grishin NV. Profile-profile comparisons by COMPASS predict intricate homologies between protein families. Protein Sci. 2003 Oct;12(10):2262-72.
71. Swofford DL. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4 Sinauer Associatesm Sunderland, Massachusetts. 2003.
72. Lanyon SM. Detecting Internal Inconsistencies in Distance Data. Systematic Zoology. 1985 Dec;34(4):397-403.
73. Siddall ME. Another monophyly index: Revisiting the jackknife. Cladistics-the International Journal of the Willi Hennig Society. 1995 Mar;11(1):33-56.
74. 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. 2001 May;18(5):691-9.
75. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution. 2007 Aug;24(8):1586-91.
76. Wolfenden R, Lewis CA, Jr., Yuan Y. Kinetic challenges facing oxalate, malonate, acetoacetate, and oxaloacetate decarboxylases. J Am Chem Soc. Apr 20;133(15):5683-5.

77. Parker DM, Holbrook JJ. The oxaloacetate reductase activity of vertebrate lactate dehydrogenase. Int J Biochem. 1981;13(10):1101-5.
78. Kabsch W. Xds. Acta Crystallographica Section D-Biological Crystallography. 2010 Feb;66:125-32.
79. Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallographica Section D-Biological Crystallography. 2010 Feb;66:213-21.
80. Kelley LA, Sternberg MJE. Protein structure prediction on the Web: a case study using the Phyre server. Nature Protocols. 2009;4(3):363-71.
81. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallographica Section D-Biological Crystallography. 2010 Apr;66:486-501.
82. Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallographica Section D-Biological Crystallography. 2010 Jan;66:12-21.
83. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic acids research. 2007 Jul;35:W375-W83.
84. Theobald DL, Wuttke DS. Accurate structural correlations from maximum likelihood superpositions. Plos Computational Biology. 2008 Feb;4(2).

Figure 1

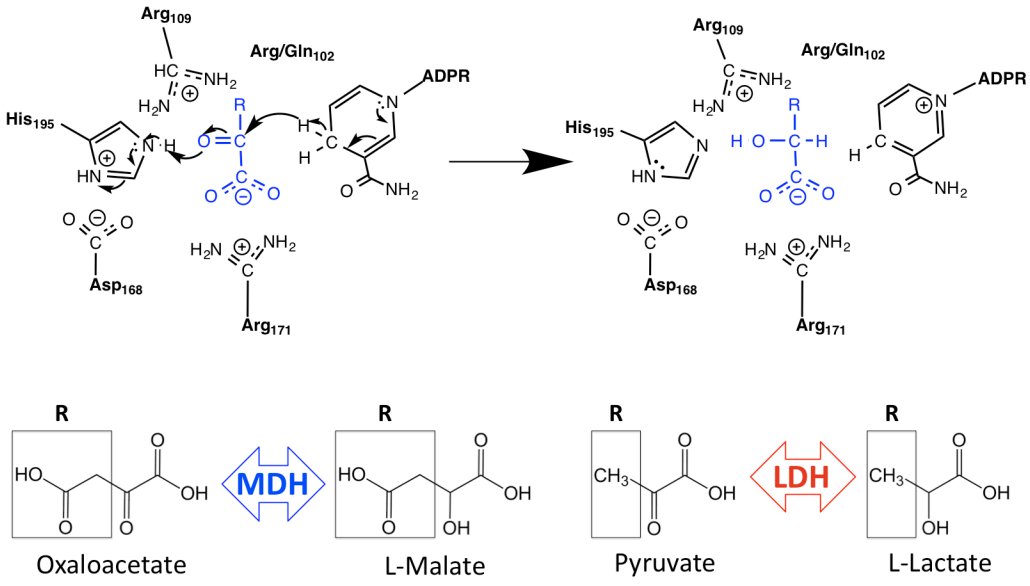

Arg109
Arg/Gln102
ADPR
His195
Asp168
Arg171
R
MDH
LDH
Oxaloacetate
L-Malate
Pyruvate
L-Lactate

Figure 2

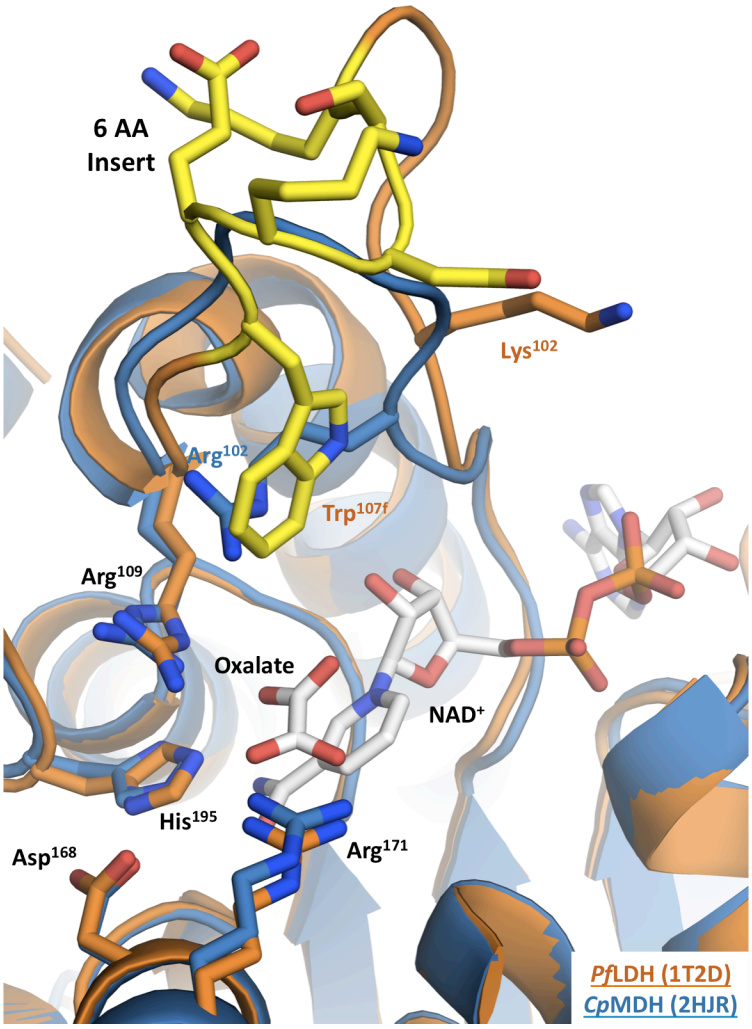
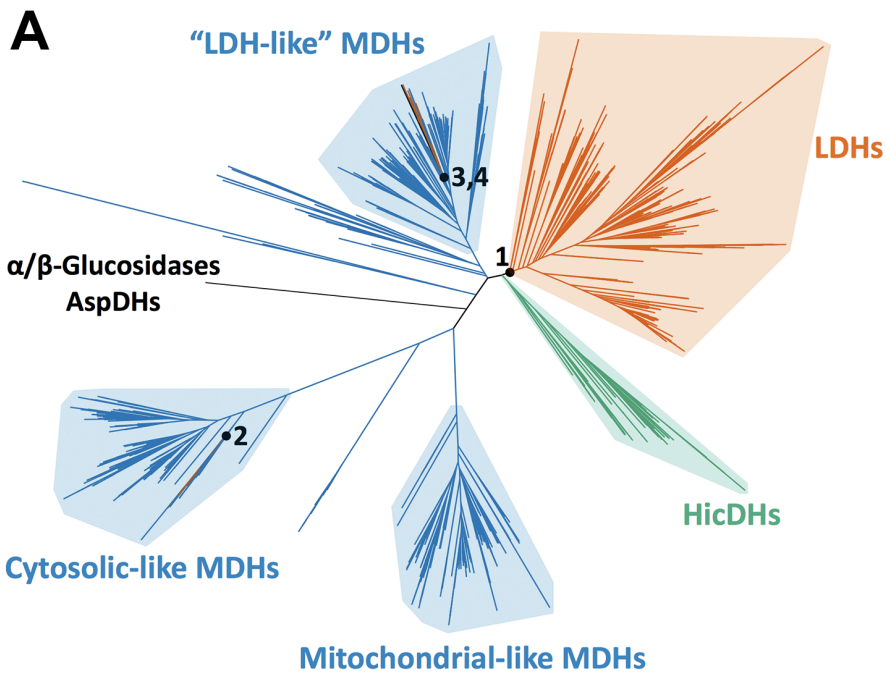

6 AA
Insert
Lys102
Arg102
Trp107f
Arg109
Oxalate
NAD+
His195
Asp168
Arg171
PfLDH (1T2D)
CpMDH (2HJR)

Figure 3


A
"LDH-like" MDHs
LDHs
3,4
1
α/β-Glucosidases
AspDHs
2
HicDHs
Cytosolic-like MDHs
Mitochondrial-like MDHs

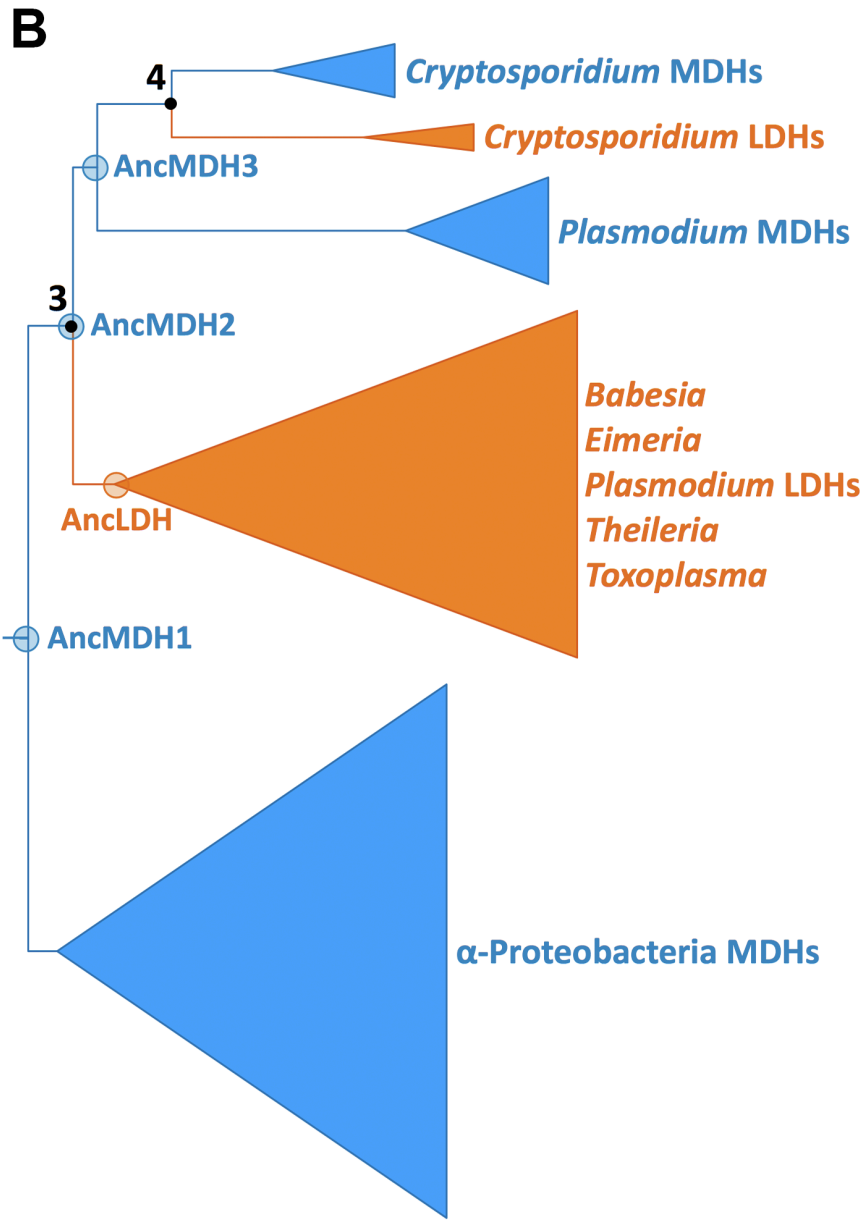

B
4
Cryptosporidium MDHs
Cryptosporidium LDHs
AncMDH3
Plasmodium MDHs
3
AncMDH2
Babesia
Eimeria
Plasmodium LDHs
Theileria
Toxoplasma
AncLDH
AncMDH1
α-Proteobacteria MDHs

Figure 4

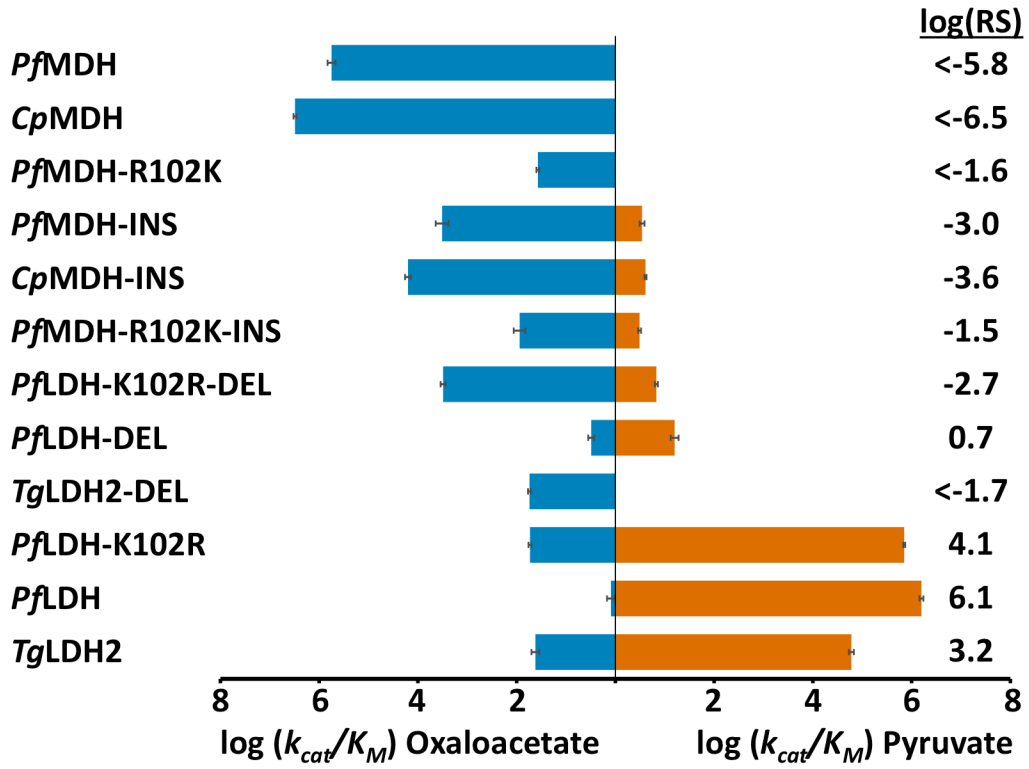

log(RS)
PfMDH
CpMDH
PfMDH-R102K
PfMDH-INS
CpMDH-INS
PfMDH-R102K-INS
PfLDH-K102R-DEL
PfLDH-DEL
TgLDH2-DEL
PfLDH-K102R
PfLDH
TgLDH2
<-5.8
<-6.5
<-1.6
-3.0
-3.6
-1.5
-2.7
0.7
<-1.7
4.1
6.1
3.2
8
6
4
2
2
4
6
8
log (kcat/KM) Oxaloacetate
log (kcat/KM) Pyruvate

Figure 5

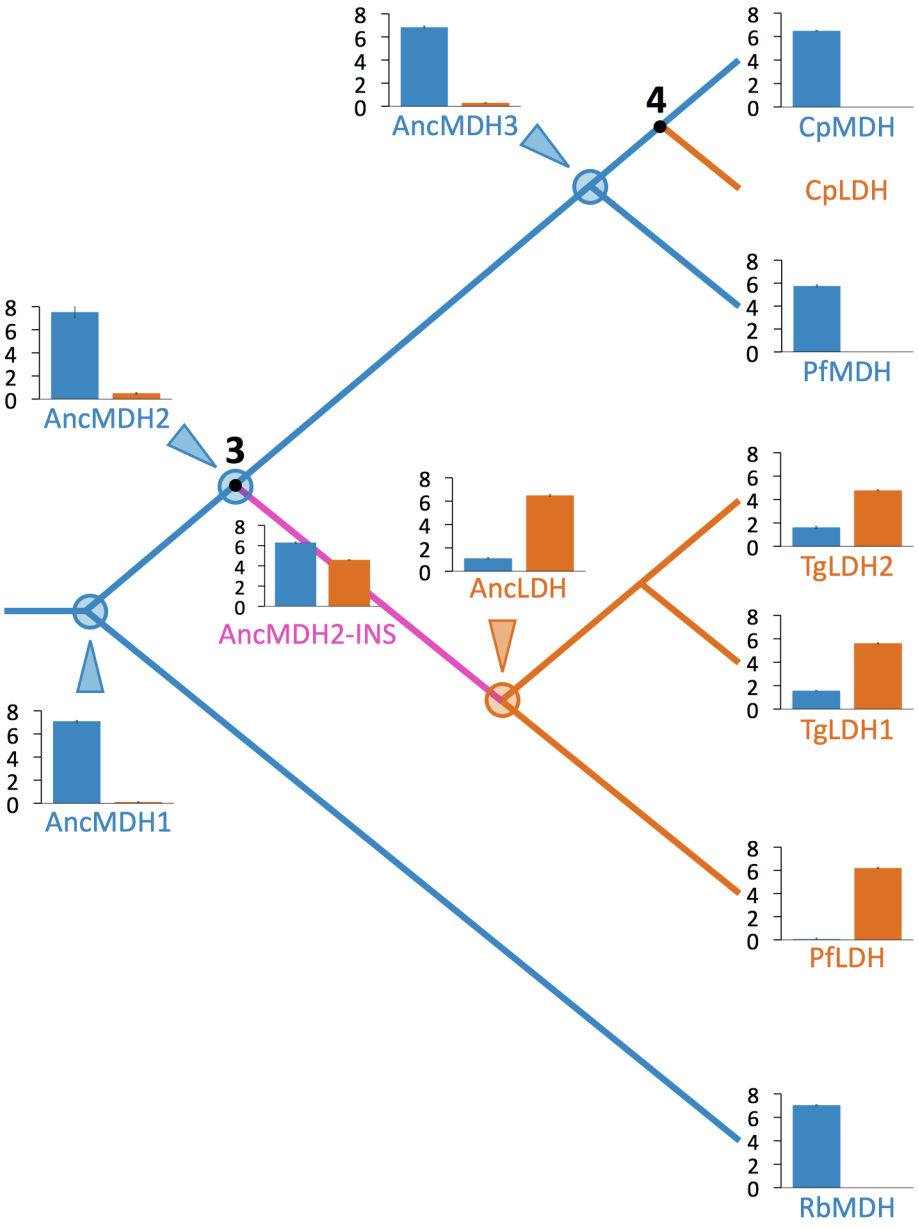

AncMDH3
CpMDH
4
CpLDH
PfMDH
AncMDH2
3
AncLDH
AncMDH2-INS
TgLDH2
TgLDH1
AncMDH1
PfLDH
RbMDH

Figure 6

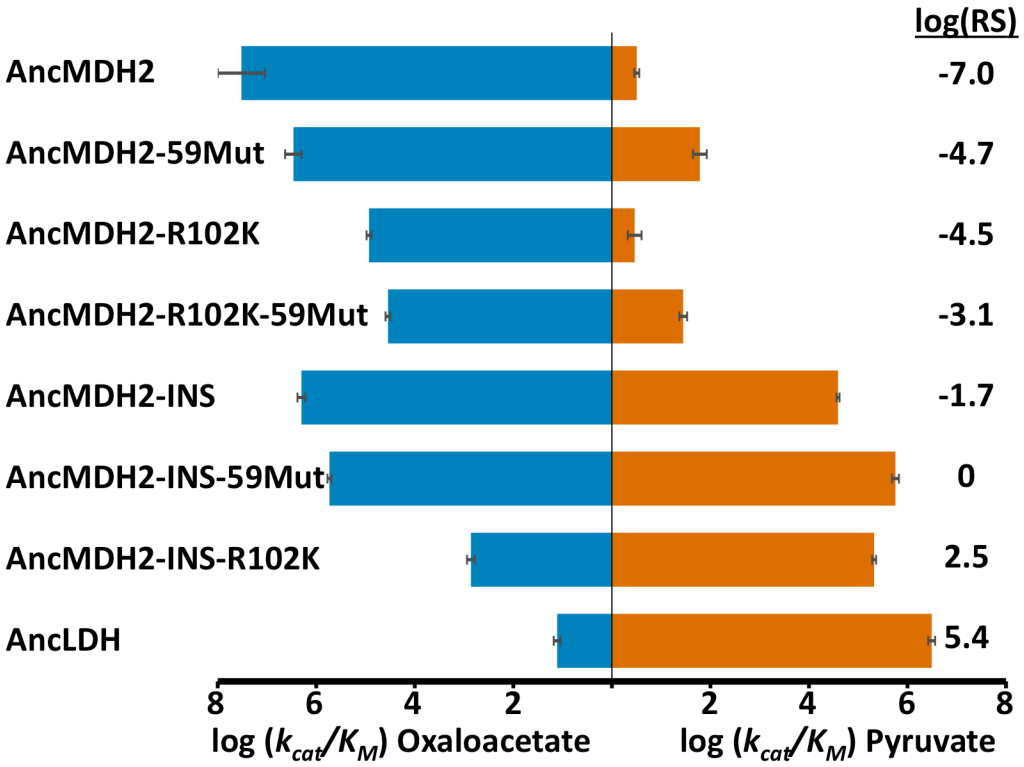

log(RS)
AncMDH2
-7.0
AncMDH2-59Mut
-4.7
AncMDH2-R102K
-4.5
AncMDH2-R102K-59Mut
-3.1
AncMDH2-INS
-1.7
AncMDH2-INS-59Mut
0
AncMDH2-INS-R102K
2.5
AncLDH
5.4
8 6 4 2
2 4 6 8
log (k_{cat}/K_M) Oxaloacetate
log (k_{cat}/K_M) Pyruvate

Figure 7

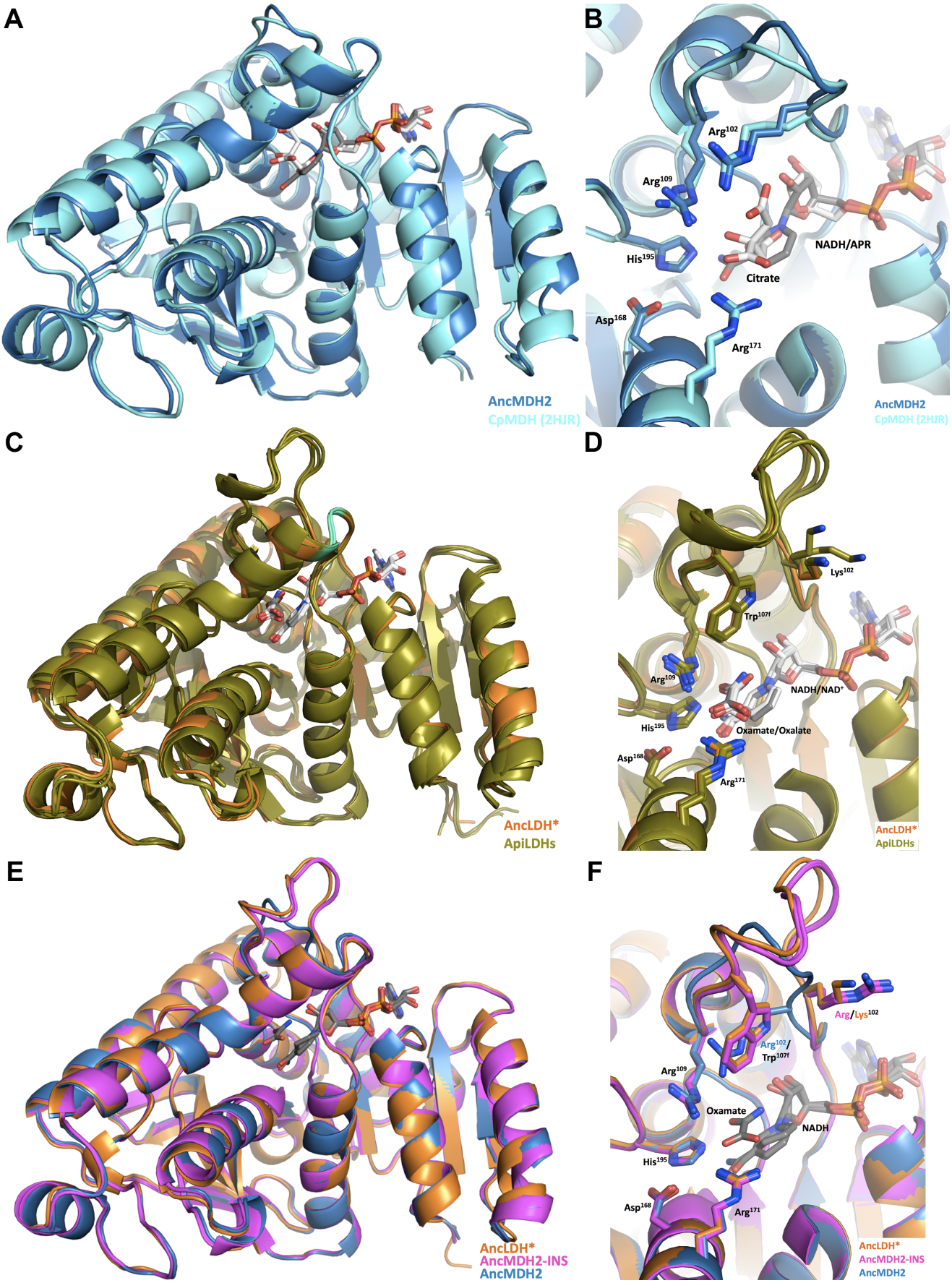
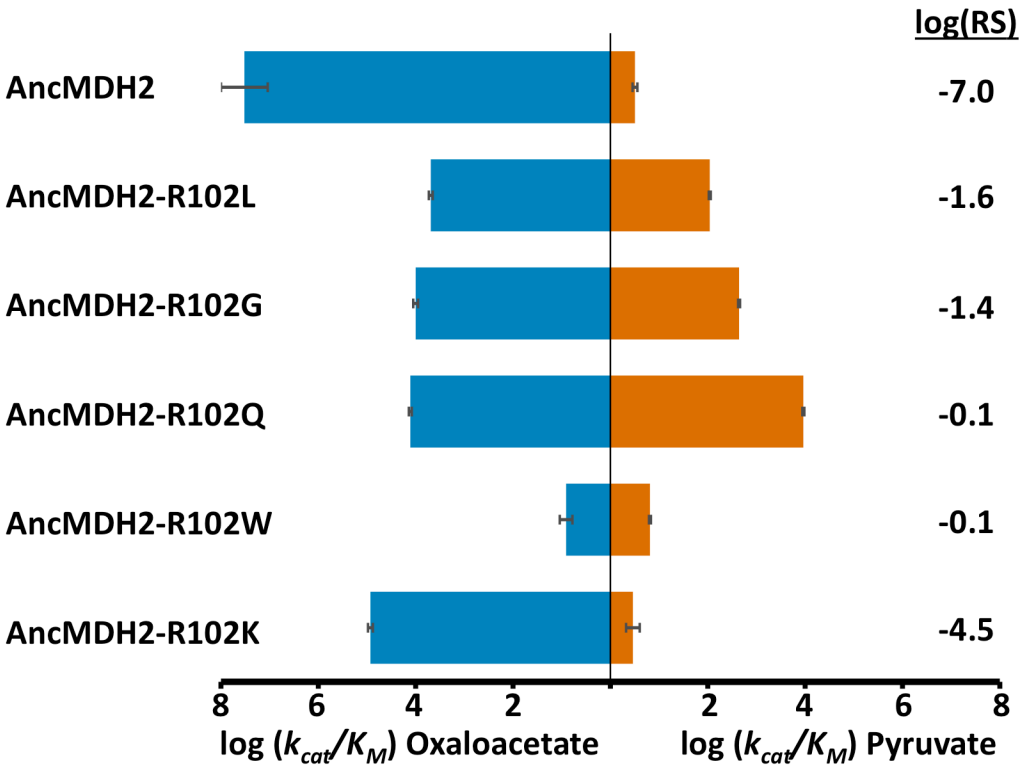

A
AncMDH2
CpMDH (2HJR)
B
Arg102
Arg109
His195
NADH/APR
Citrate
Asp168
Arg171
AncMDH2
CpMDH (2HJR)
C
AncLDH*
ApiLDHs
D
Lys102
Trp107f
Arg109
NADH/NAD+
His195
Oxamate/Oxalate
Asp168
Arg171
AncLDH*
ApiLDHs
E
AncLDH*
AncMDH2-INS
AncMDH2
F
Arg/Lys102
Arg102/
Trp107f
Arg109
Oxamate
NADH
His195
Asp168
Arg171
AncLDH*
AncMDH2-INS
AncMDH2

Figure 8


log(RS)
AncMDH2
-7.0
AncMDH2-R102L
-1.6
AncMDH2-R102G
-1.4
AncMDH2-R102Q
-0.1
AncMDH2-R102W
-0.1
AncMDH2-R102K
-4.5
8
6
4
2
2
4
6
8
log (k_{cat}/K_M) Oxaloacetate
log (k_{cat}/K_M) Pyruvate