Evolution of substrate specificity in a retained enzyme driven by gene loss

Ana L. Juárez-Vázquez,[1] Janaka N. Edirisinghe,[2,8] Ernesto A. Verduzco-Castro,[1] Karolina Michalska,[3,4] Chenggang Wu,[5] Lianet Noda-García,[1,a] Gyorgy Babnigg,[3] Michael Endres,[3] Sofía Medina-Ruiz,[1,b] J. Julián Santoyo-Flores,[6] Mauricio Carrillo-Tripp,[6,c] Hung Ton-That,[5] Andrzej Joachimiak,[3,4,7] Christopher S. Henry,[2,8] and Francisco Barona-Gómez[1,*]

[1]Evolution of Metabolic Diversity Laboratory, Unidad de Genómica Avanzada (Langebio), Cinvestav-IPN. Km 9.6 Libramiento Norte, Carretera Irapuato - León, CP 36821, Irapuato, México.

[2]Computing, Environment and Life Sciences Directorate, Argonne National Laboratory, South Cass Avenue, Lemont, IL, USA.

[3]Midwest Center for Structural Genomics, Biosciences Division, Argonne National Laboratory, Lemont, IL, USA.

[4]Structural Biology Center, Biosciences Division, Argonne National Laboratory, Lemont, IL, USA.

[5] Department of Microbiology and Molecular Genetics, University of Texas Health Science Center, Houston, TX, USA.

[6]Laboratorio de la Diversidad Biomolecular, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, México.

[7]Departments of Biochemistry and Molecular Biology, University of Chicago, Chicago, IL, USA.

[8]Computation Institute, University of Chicago, Chicago, IL, USA.

[a]Current address: Department of Biological Chemistry, Weizmann Institute of Science, Rehovot, Israel.

[b]Current address: Department of Molecular & Cell Biology, University of California, Berkeley, USA.

[c]Current address: Ciencias de la Computación. Centro de Investigación en Matemáticas, A.C. Guanajuato, México.

[*]Corresponding author: Francisco Barona-Gómez; Fax: +52-462-1663000; Tel: +52-462-1663017; email: francisco.barona@cinvestav.mx

Abstract

The connection between gene loss and the functional adaptation of retained proteins is still poorly understood. We apply phylogenomics and metabolic modeling to detect bacterial species that are evolving by gene loss, with the finding that *Actinomycetaceae* genomes from human cavities are undergoing sizable reductions, including loss of L-histidine and L-tryptophan biosynthesis. We observe that the dual-substrate phosphoribosyl isomerase A or *priA* gene, at which these pathways converge, appears to coevolve with the occurrence of *trp* and *his* genes. Characterization of a dozen PriA homologs shows that these enzymes adapt from bifunctionality in the largest genomes, to a monofunctional, yet not necessarily specialized, inefficient form in genomes undergoing reduction. These functional changes are accomplished via mutations, which result from relaxation of purifying selection, in residues structurally mapped after sequence and X-ray structural analyses. Our results show how gene loss can drive the evolution of substrate specificity from retained enzymes.

Introduction

Genome dynamics, or the process by which an organism gains or loses genes, plays a fundamental role in bacterial evolution. Acquisition of new functions due to horizontal gene transfer (HGT) or genetic duplications is broadly documented (1,2). Gene loss has also been implicated in rapid bacterial adaptation after experimental evolution (3), but this process has not yet been confirmed in natural populations. Phylogenomics involves the comparative analysis of the gene content of a set of phylogenetically related genomes to expose new insights into genome evolution and function, and this approach has been classically applied to study how gene gain is associated with functional divergence in bacteria (4). Here, we propose that bacterial phylogenomics can be similarly applied to study evolution by gene loss (5), specifically where enzymes are evolving within bacterial species that are undergoing genome decay (6,7).

The current bias toward in-depth functional analysis of proteins from genomes that are undergoing gene gain by HGT versus gene loss by decay is likely due to two factors. First, in genomes that are undergoing decay, there is a relaxation in the selection pressure that increases mutation rates in functioning proteins as these proteins begin to contribute less to cell fitness (8,9). As a result, these proteins display higher-than-normal mutation rates, making *in vitro* analysis of protein function a challenge (10). Second, there is only a brief window of opportunity to study the evolution of most proteins during genome decay in bacteria. This is because most proteins are monofunctional, and they are rapidly removed from the bacterial genome once they become dispensable due to gene loss. To overcome this limitation, we propose to use a bifunctional enzyme to study the evolution of substrate specificity after gene loss, as these enzymes may continue to

operate when only one of their associated metabolic pathways becomes dispensable. We use genome-scale metabolic models to determine when each pathway is lost as well as when they become non-functional (11).

The phylum *Actinobacteria,* Gram-positive organisms with high (G+C)-content, are ubiquitous and show one of the highest levels of bacterial metabolic diversity (12). This phylum is known to display significant metabolic specialization, and phylogenomics has been previously applied to correlate genome dynamics with metabolic pathway evolution and enzyme specialization (13, 14, 15). Moreover, within the deep-rooted family *Actinomycetaceae*, phylogenetic analyses have suggested the occurrence of genome decay (16). Furthermore, we have observed that many actinobacterial species lack a *trpF* gene, while retaining a copy of the potentially bi-functional *priA* gene. The PriA enzyme is capable of operating in the L-histidine biosynthesis pathway as HisA, while also functioning in the L-tryptophan biosynthesis pathway as TrpF (13, 14, 17). As such, we suggest PriA is an ideal candidate to study protein evolution during the process of genome decay.

The product of the *priA* gene, which is a *hisA* homolog, catalyzes two analogous isomerizations of structurally similar substrates: (i) the conversion of *N*-(5'-phosphoribosyl)-anthranilate (PRA) into 1-(*O*-carboxyphenylamino)-1'-deoxyribulose-5'-phosphate (CdRP) (TrpF or PRA isomerase activity); and (ii) the conversion of *N*-[(5-phosphoribosyl) formimino]-5-aminoimidazole-4-carboxamide ribonucleotide (ProFAR) into *N*-[(5-phosphoribulosyl)formimino]-5-aminoimidazole-4-carboxamide ribonucleotide (PRFAR) (HisA or ProFAR isomerase activity) (17). Moreover, evolution of PriA in response to genome dynamics has lead to the appearance of the SubHisA and

PriB subfamilies, which have been shown to have different substrate specificities (13, 14) (Figure 1). While these enzyme subfamilies were respectively discovered using phylogenetics in the genera *Corynebacterium* and *Streptomyces*, no study has ever centered on the metabolic genes associated with *priA* in the deep-branching organisms belonging to the phylum *Actinobacteria*, such as those of the family *Actinomycetaceae* where the transition from HisA into PriA must have taken place (18, 19).

Here we exploit the intrinsic features of PriA to explore the link between the evolution of enzyme function and gene loss within the family *Actinomycetaceae*, which includes many human oral cavity commensal and pathogenic organisms (16, 20, 21). We classify this bacterial family into four major evolutionary lineages, including two specific for the genus *Actinomyces*. One of these lineages shows extensive gene loss, including *his* and *trp* biosynthetic genes. After the loss of constrictions imposed by the retention of biosynthetic pathways, we found that evolutionary patterns correlate with the sub-functionalization, yet not necessarily specialization, of PriA into two new subfamilies, which we named SubHisA2 and SubTrpF. We support this classification by comprehensive *in vivo* and *in vitro* biochemical characterization of a dozen PriA homologs from *Actinomyces* and closely related taxa. X-ray structural analysis and molecular docking simulations were further used to start investigating the evolution of substrate specificity by gene loss in structural grounds. Our results demonstrate that gene loss can drive functional protein divergence, and provide unprecedented insights into the evolution of enzyme substrate specificity in retained enzymes after gene loss in the bacterial genome.

Results

Phylogenomic resolution of the Actinomycetaceae *family*.

To find evidence of gene loss in deep-branching organisms of the phylum *Actinobacteria*, specifically within genera belonging to the order *Actinomycetales*, we selected 133 representative organisms from 18 families with available genome sequences (Figure 2A – Figure 2-source data 1). We then aimed at resolving their taxonomic relationships using 35 single-copy proteins that are conserved among all 133 genomes analyzed (see Methods and Figure 2A – Figure 2-source data 2). We concatenated these proteins to reconstruct their phylogeny, and supported the resulting tree by significant Bayesian posterior probabilities. The phylogenetic tree shows several long branches, which correspond to the families *Actinomycetaceae*, *Micrococcaceae*, *Propionibacteriaceae* and *Coriobacteriaceae*, and to the genus *Tropheryma*. The tree also includes a clade with the family *Bifidobacteriaceae* as the root of six different sister families, including the *Actinomycetaceae* (blue branch and grey box in Fig. 2A**)**.

As expected, all of the organisms contained in the rapidly evolving lineages trended towards smaller genomes and lower (G+C)-content (Fig. 2B). The *Actinomycetaceae* genomes were characterized by particularly broad variances in genome size and (G+C)-content, with the variation being most apparent for organisms belonging to the genus *Actinomyces*. Representative organisms of this genus, e.g. *A.* sp. oral taxon 848 str. F0332, *A. oris* MG-1, *A. neuii,* and *A. odontolyticus*, are distributed throughout the *Actinomycetaceae* clade (blue dots in Fig. 2B). Given these observations, we selected

144 the genus *Actinomyces* as the ideal target for a deeper analysis of rapid evolution by gene
145 loss.

146 We then carried out a phylogenomic analysis using the genome sequences of 33
147 organisms from the family *Actinomycetaceae* (Figure 3 - Figure 3-source data 1), from
148 which 27 are classified as *Actinomyces* (16, 20), including the model strain *A. oris* MG-1
149 sequenced in this study. The remaining sequences came from the genera *Actinotignum*,
150 formerly *Actinobaculum* (21), *Trueperella, Varibaculum* and *Mobiluncus*. As an out-
151 group we used the genera *Bifidobacterium*, which included 8 genome sequences. We
152 identified a total of 205 single-copy proteins shared among all these 41 organisms, which
153 were used for constructing a concatenated phylogenetic tree by Bayesian (Fig. 3 - Figure
154 3-source data 2), and maximum likelihood approaches (Fig. 3 - supplement Fig. 1). Based
155 in this analysis, the family *Actinomycetaceae* separated into four evolutionary lineages
156 contained in three sub-clades: Lineage I, which includes *A.* sp. oral taxon 848 str. F0332
157 (Org10); lineage II, which includes *A. oris* MG-1 (Org21); and lineages III and IV, which
158 form a monophyletic group and include *A. neuii* (Org27) and *A. odontolyticus* (Org41),
159 respectively. Remarkably, these lineages group depending on their mammalian hosts and
160 human body niches from which they were isolated (Figure 3 - Figure 3-source data 1).

161 Our phylogenetic analysis also shows that 25 of the 27 *Actinomyces* species
162 analyzed have a paraphyletic origin leading to lineages II and IV. These two lineages can
163 be distinguished not only according to their genome size and (G+C)-content, but also to
164 the number of coding sequences (CDS) and metabolic functions or subsystems (Fig. 3 -
165 supplement Fig. 2). Specifically, as revealed by genome annotation using RAST (22),
166 lineage II, which has the highest (G+C)-content (68.32% on average) and the biggest

genome size (3.04 Mbp on average), has the largest number of amino acid biosynthetic pathways (see next section). This observation contrasts with the results obtained for lineage IV, which shows reduced (G+C)-content (60.66% on average) and genome size (2.19 Mbp on average), as well as less amino acid biosynthetic pathways. Indeed, the genomic differences between lineages II and IV were found to be statistically significant, including the presence or absence of the *his* and *trp* biosynthetic genes (Fig. 3 - Figure 3-figure supplement 2-source data 1). We explore this observation in more detail by constructing metabolic models for all of the analyzed genomes in the following section.

Metabolic evolution of the* Actinomycetaceae *family

In order to reduce the risk of overreaching conclusions based only in homology sequence searches, we constructed genome-scale metabolic models of all 33 organisms comprising this family, plus the 9 outgroup *Bifidobacterium* species (see Methods). Next, flux balance analysis was applied to predict the minimal nutrients required to support growth for each genome. Finally, after automated curation of the metabolic reconstructions (23), which includes not only homologous but also analogous enzymes, we classified each reaction in each model as: (i) essential for growth on predicted minimal media; (ii) functional but not essential; and (iii) nonfunctional. All model results, which represent the highest quality functional annotation available for metabolism, are provided as source data of Fig. 4: model overview (Figure 4-source data 1), reaction content and classifications (Figure 4-source data 2) and predicted minimal media (Figure 4-source data 3).

189 The lineage II models were generally the largest with an average of 1019 gene-
190 associated reactions, which is to be expected since the lineage II genomes are also the
191 largest. These models also had the fewest predicted essential nutrients with an average of
192 19 nutrients required. This result indicates that most biosynthetic pathways for essential
193 biomass precursors are complete in the lineage II models. The lineage I and IV models
194 were substantially smaller with an average of 850 and 843 gene-associated reactions,
195 respectively. Although similar in size, the lineage I models had more required nutrients
196 (25 on average) compared with the lineage IV models (22 on average). Finally, the
197 lineage III models were the smallest of all, with an average of 817 gene-associated
198 reactions. Surprisingly, these models still had fewer required nutrients than the lineage I
199 models (23 on average). These results provide a meaningful biochemical context in which
200 biosynthetic enzymes are evolving.
201 To study the metabolic diversity of each lineage in more detail we performed a
202 comparative analysis of the gene-associated reactions of our models (Fig. 4A - 4D).
203 Given the large metabolic and genetic diversity, we used less stringent parameters than
204 those used for our core genome analysis sustaining our phylogenomics of previous
205 section (see Methods). This comparative analysis revealed that the lineage II genomes
206 were the least diverse, with a very large fraction of reactions present in all models,
207 including those for amino acid biosynthesis (Fig. 4 - supplement Fig. 1). All other
208 lineages were more diverse. Interestingly, a comparative analysis of our models found
209 that all models across all lineages share a common conserved core of 695 reactions.
210 When we similarly compute a conserved core for each individual lineage (Fig. 4E), we

find that the 89% of reactions in the conserved core for each lineage are contained in the conserved core across all lineages.

From these modeling results, we clearly see that lineages I, III and IV are all undergoing the process of gene loss, resulting in a reduction towards a common set of core metabolic pathways. This explains the rapid development of diversity within each lineage, as well as the variability in minimal required nutrients predicted by our models. We can also apply our models to study the gene loss process from a mechanistic perspective by looking for patterns in the presence and absence of genes and reactions for two specific pathways of interest: L-tryptophan and L-histidine biosynthesis. Our models predicted genomes in lineages I, III, and IV (but none from lineage II) that required these amino acids as a supplemental nutrient, indicating the loss of these biosynthetic pathways in these organisms. We also observed that the presence of the *priA* gene, which takes part in both L-tryptophan and L-histidine biosynthesis, closely tracked with the presence of these pathways in these genomes (Fig. 5A**)**. This observation suggests that gene losses could have an effect on the evolution of the retained PriA enzymes.

***Molecular evolution of PriA within the family* Actinomycetaceae**.

To bring down these observations at the enzyme level, we carried out comparable phylogenetic analyses of PriA (Fig. 5), and we measured the evolutionary rate of its gene by estimating the d_N/d_S ratio (ω value) for each resulting clade (Table 1). The PriA phylogeny was complemented with an analysis of the occurrence of the *his* and *trp* biosynthetic genes, including *priA* for both pathways (Fig. 5 - Figure 5-source data 1). Excluding the out-group, our phylogenetic reconstructions show that PriAs from different

lineages are grouped in three sub-clades, highlighted in purple, orange, and yellow boxes in Fig. 5A, which have distinguishable selection pressures operating upon them. This analysis also shows that PriA coevolves with the presence or absence of the *his* and *trp* biosynthetic genes (Fig. 5B).

The purple box denotes a paraphyletic clade that includes PriAs from lineage II, as well as the PriAs from the genus *Bifidobacterium* used as an out-group. The d_N/d_S value of this lineage, which retains the entire set of *his* and *trp* genes, is 0.0636, consistent with purifying selection. The orange and yellow boxes denote polyphyletic groups that include PriAs from lineages I, III, and IV. Interestingly, the included taxa within these lineages lost their extant *his* or *trp* genes differentially (Fig. 4A), and their d_N/d_S values are 0.0901 and 0.1459, respectively, which is suggestive of relaxation of purifying selection. Moreover, the higher d_N/d_S values in the clade shown in yellow seem to be due to accumulation of nonsynonymous substitutions, in other words, higher values of d_N that may relate to changes in enzyme specificity (Table 1).

Thus, in the basis of these evolutionary observations we proposed three functional and testable hypotheses related with the emergence of novel PriA enzyme subfamilies in the bacterial family *Actinomycetaceae* (Fig. 6A). In H1 (purple box) we assume that PriA homologs are conserved as enzymes with dual-substrate specificity, capable of converting both PRA and ProFAR substrates. In H2 (orange box) and H3 (yellow box) the PriA homologs are expected to be monofunctional isomerases, yet not necessarily specialized enzymes, capable of converting ProFAR or PRA as substrates, respectively. Moreover, given that relaxation of purifying selection is associated with the latter two hypothetical scenarios, H2 and H3, our model predicts monofunctional, yet not necessarily specialized

enzymes capable of supporting growth. Representative enzymes of each hypothesis were selected for further biochemical characterization, as described.

Biochemical confirmation of the evolution of PriA by gene loss.

Before evaluating the functional implications of our evolutionary hypotheses from previous section, we confirmed that the *priA* gene is functional in *Actinomyces*. For this purpose we used allelic exchange to delete the *priA* gene from the chromosome of *A. oris* MG-1 (Org21) (24, 25), a model strain that belongs to lineage II and whose genome was sequenced as part of this study. Mutation of *priA* in this organism was confirmed by sequencing the entire genome of the resulting Δ*priA* mutant strain (Supplementary file 1). Determination of the growth requirements of this strain, termed Δ*priA*_Org21, showed that *priA* mutation leads to L-tryptophan auxotrophy, demonstrating the physiological relevance of PriA in this organism. Unexpectedly, however, the Δ*priA* mutant remains prototrophic for L-histidine, which could not be explained in the basis of current data. Thus, it is tempting to speculate that this phenotype may found an explanation in the previously reported association between enzyme promiscuity and genome decay (6,7).

To biochemically evaluate the functional implications of our evolutionary hypotheses (Fig. 6A), we characterized nine selected PriAs, both *in vivo,* by complementation assays using *trpF* and *hisA Escherichia coli* mutants; and *in vitro*, by estimation of their Michaelis Menten steady-state enzyme kinetic parameters, as we have previously done (13, 14, 26). The results of these experiments are included in Table 2. First, *in vivo* complementation assays using appropriate *priA* constructs, showed that PriA homologs from Org15, Org21, and Org22 (H1) were able to rescue growth of both HisA

and TrpF deficient strains. Second, *priA* homologs from Org34 and Org36 (H2) complemented the HisA activity and, to a lesser extent, the TrpF activity. Third, those *priA* homologs from Org10, Org13, Org39 and Org41 (H3) were able to complement the TrpF activity but not the HisA activity.

The *priA* homologs were then heterologously expressed and purified to homogeneity in *E. coli* (see Methods). Only five enzymes out of nine were found to be soluble and could be purified as needed, which agrees with the high mutation rate encountered in previous section. Fortunately, we obtained Michaelis Menten enzyme kinetics parameters for representative enzymes of all three evolutionary hypotheses, namely, three enzymes belonging to H1 and one enzyme each for H2 and H3, with the following results (Fig. 6B and Table 2). First, enzymes from Org15, Org21, and Org22 (H1) showed dual-substrate specificity but also poor catalytic efficiencies, namely, k_{cat}/K_M^{ProFAR} from 0.01 to 0.1 $\mu M^{-1}s^{-1}$ and k_{cat}/K_M^{PRA} around 0.01 $\mu M^{-1}s^{-1}$. Second, only ProFAR isomerase activity could be detected *in vitro* using pure enzyme from Org36 (H2), with a catalytic efficiency of k_{cat}/K_M^{ProFAR} of 0.002 $\mu M^{-1}s^{-1}$, but not PRA isomerase activity, as suggested by our highly sensitive *in vivo* complementation assay. Third, PRA isomerase activity as the sole activity present in H3 was confirmed in the enzyme purified from Org42, with a k_{cat}/K_M^{PRA} of 0.02 $\mu M^{-1}s^{-1}$.

The obtained enzyme kinetics parameters suggest that mutations that accumulate during relaxation of purifying selection, which make these enzymes difficult to work with, affect the turnover (k_{cat}). In the case of the H1 enzymes, the poor turnovers are compensated for by relatively high substrate affinities (K_M), mainly for ProFAR.

However, this does not seem to be the case for the enzymes belonging to H2 and H3, which have poor K_M parameters not only for the substrate of the missing activity but also for the substrates they are active against, ProFAR and PRA, respectively. Therefore, PriA homologs from *Actinomyces* have poor catalytic efficiencies when compared with *bona fide* PriAs from its closely related genus *Bifidobacterium* (Table 2). This suggests that enzyme evolution–from bifunctionality to monofunctionality–under relaxation of purifying selection does not necessarily express itself in the same way as recorded during purifying or positive selection, where specialization and enzyme proficiency come together.

The case of the *in vivo* PRA isomerase activity detected for the enzyme from Org36, which could not be confirmed *in vitro*, may be related to the different resolutions of our enzyme assays. For instance, the detection limits for the PRA and ProFAR isomerase assay used in the present study are 0.0001 $\mu M^{-1}s^{-1}$ and 0.001 $\mu M^{-1}s^{-1}$, respectively (13,14,26). However, despite the poor catalytic efficiency of all *Actinomyces* enzymes investigated, these detection limits guarantee that our enzyme parameters are in agreement between them and with our hypotheses. Based on these results the family related to H1, which has both activities, is referred to as PriA, whereas the latter two enzyme subfamilies, related to H2 and H3, were renamed as SubHisA2 and SubTrpF, respectively. These names, together with the name of the organism from which the enzymes were obtained, are used in Table 2 and in the following sections.

Structural insights into the evolution SubHisA2 and SubTrpF.

To potentially identify mutations in active-site residues that may affect k_{cat} and K_M parameters, we attempted to elucidate the structure of the five PriA homologs that we were able to *in vitro* characterize. However, we were only able to crystallize and solve the structure of PriA_Org15 (H1) at atomic resolution of 1.05 Å (PDB: 4X2R, Fig. 7 - Figure 7-source data 1). To compare this structure with SubHisA2_Org36 (H2) and SubTrpF_Org41 (H3), we opted for the construction of structural homology models. Since the ability of PriA to accept both ProFAR and PRA as substrates requires productive conformations, we also explored these interactions using molecular docking. This was complemented with detailed structure-based multiple sequence alignments taking into account all available PriA functional and structural data (Fig. 7B). This combined approach allowed us to identify mutations that may be driving the evolution of PriA into SubHisA2 and SubTrpF enzyme subfamilies.

Changes of conserved residues from PriA (H1) into SubHisA2 (H2) enzymes include Ile47Leu and Ser79Thr. Previous independent mutation of these two residues, even into similar amino acids, in SubHisA from *Corynebacterium* abolished the PRA isomerase activity of this monofunctional enzyme (13). Analogously, the SubHisA2_Org36 has a change of Ser79 into Thr79 (Fig. 7B). In this mutation, the methyl group of the threonine residue may affect the contact between PRA and the hydroxyl group common to these residues (Fig. 7A), thus abolishing PRA isomerase activity. This effect agrees with the estimated binding affinities for ProFAR (-9.5 kcal/mol) and PRA (-9.2 kcal/mol) obtained after molecular docking (Supplementary file 2). The energy-minimized docking model of the productively bound PRA, in agreement with the kinetic parameters from the preceding section, indicates that the catalytic residue

347 Asp11 does not interact with the 2′-hydroxyl group from the substrate. A precedent for
348 this contact is found in previous X-ray structural and mutagenesis analysis of *bona fide*
349 PriA enzymes (26, 27).

350 Comparison of PriA (H1) with SubTrpF (H3) revealed the mutations Gly126Cys
351 and Trp139Gly. In PriA, Gly126 faces the active site near the catalytic residue Asp128.
352 The introduction of the Cys side-chain in SubTrpF could influence the positioning of
353 Arg137 with respect to Asp128, obstructing the accommodation of ProFAR, as this
354 region interacts with a large phosphosugar moiety that is absent from PRA (Fig. 7B).
355 Furthermore, Trp139, which is catalytically important for conversion of ProFAR by PriA,
356 is mutated into several different amino acid residues in SubTrpFs. Trp139 is important
357 for the correct positioning of the catalytic residues present in loop 5, and for substrate
358 binding through stacking interactions (14, 27). Indeed, the indole group of Trp139 in
359 PriAs can form a hydrogen bond with Asp128, stabilizing the knot-like conformation
360 observed during ProFAR binding. Thus, mutation of this residue in SubTrpF is in
361 agreement with the loss of ProFAR isomerase activity. Arg83 is also interesting as it is
362 differentially missing from SubTrpF, and/or the fragment preceding it contains a two-
363 residue insertion (Fig. 7B). Arg83 interacts with the second phosphate group of ProFAR,
364 allowing its correct position in the substrate-binding pocket of PriA. Overall, these
365 modifications in key residues disfavor the ProFAR binding affinities, a result that is in
366 agreement with the enzyme kinetic parameters and the estimated binding affinities for
367 ProFAR (-9.5 kcal/mol) and PRA (-9.7 kcal/mol) obtained after molecular docking
368 (Supplementary file 2).

Although further research will be needed to confirm the exact mutations, and their roles, leading to SubTrpF and SubHisA2 sub-families, our results provide a promising first step towards deciphering at the atomic level how relaxation of purifying selection influences the evolution of substrate specificity. At this point in time, when PriA, SubTrpF and SubHisA2 sequences and structures are still scarce, the effects of genetic drift, i.e. mutations related to taxonomic distance rather than functional divergence (as previously shown for the evolution of PriB from PriA [14]) cannot be ruled out. An extra factor potentially hampering sequence and structural analysis is the higher-than-normal mutation rates of these protein sub-families, which translates into lack of sequence conservation and disordered regions in X-ray crystal structures. Our structural data, including the estimates for molecular binding affinities, can therefore only be used to support other biochemical and evolutionary evidence.

Discussion

Our study highlights the use of phylogenomics and metabolic models to identify and investigate gene loss in bacteria. Our results indicated that the distinctive reactions retained in each *Actinomyces* genome reflect the preservation of some full biosynthetic pathways over others, conferring a capacity to grow on different sets of environmental nutrients. This result in turn implies an exposure of these genomes to a diverse range of environmental conditions and selection pressures, while the phylogenetic proximity of these functionally diverse genomes speaks to a strong capacity for rapid adaptation to the diverse conditions present in the human body. The process of gene loss, associated with

392 relaxation of purifying selection, is the key driver of this adaptation strategy. Thus,
393 metabolic diversity in complex systems as the human microbiome might be characterized
394 after reconstruction of evolutionary trajectories, which may reflect different bacterial
395 functions and ecological sub-niches. The pattern of reaction conservation seen in our
396 metabolic modeling analysis exemplifies a likely signature for gene loss, which could be
397 used to identify these phenomena among other genome families. Remarkably, in this
398 context, enzyme specialization does not necessarily means catalytic proficiency.

399 Our study of this gene loss process exposed evolutionary patterns of PriA in L-
400 tryptophan and L-histidine biosynthesis pathways, with the potential to unveil the
401 underpinning mechanisms driving the evolution of substrate specificity of retained
402 enzymes. Because multifunctional enzymes may have more than one constraint operating
403 on them, tracking functional evolution promptly after selection is relaxed during genome
404 decay might be done more readily than with monofunctional enzymes. As shown here,
405 only partial selection may be released in the retained bifunctional enzyme PriA. Indeed,
406 the predicted metabolic phenotypes unveiled by flux balance analysis did correlate better
407 with the evolutionary patterns revealed by metabolic gene occurrence and PriA
408 phylogenetic reconstructions than they did with the natural history told by the species
409 tree. To confirm this sort of evolutionary behavior, other instances of well-known
410 multifunctional proteins, such as moonlighting proteins, may be investigated.

411 The occurrence of SubHisA2 in *Actinomyces*, together with the appearance of
412 SubHisA in *Corynebacterium*, demonstrates that subfunctionalization of PriA leading
413 into HisA-like enzymes has occurred at least twice. Such phenotypic plasticity is a
414 reflection of the intrinsic enzymatic proficiency of PriA upon two related but

topologically dissimilar substrates; but, more interestingly, the evolutionary histories behind these independent subfunctionalization events responded to somehow contrasting evolutionary mechanisms. Whereas SubHisA is the result of positive selection after the acquisition of an entire *trp* operon by HGT (13), SubHisA2 responded to the loss of *trp* genes, and it evolved under relaxation of purifying selection. Consequently, SubHisA has drastic mutations in its catalytic active site, which have been shown to be responsible for its inability to catalyze PRA, whilst SubHisA2 shows some residual PRA isomerase activity, congruent with the observation that its active-site architecture is almost completely conserved.

The subfunctionalization of PriA into SubTrpF, in contrast, has been documented only here. This functional shift had to involve 'non-proficient' enzyme specialization, since the ancestral activity of PriA is ProFAR isomerase (19). Thus, the appearance of SubTrpF with substitutions in its catalytic active-site could be discussed based on previous knowledge about PriA. These mutations actually resulted in the elimination of the ancestral ProFAR activity, which is remarkable because the driving force behind this process relates to the relaxation of purifying selection. In agreement, a recent study of PriA sequences obtained from a diverse metagenome, complemented by some of the SubTrpF sequences studied here, classified this enzyme subfamily at the transition from HisA into PriA (18). Since *Actinomyces* undergoes interspecies recombination with protein functional implications (28), such a mechanism may provide a means to explain the sequence heterogeneity found in these *Actinomyces* PriA homologs.

Our study, therefore, provides experimental evidence that gene loss can drive functional protein divergence. It also shows that, despite the fact that the retained

438 enzymes possess low catalytic activities, they contribute to the maintenance of
439 metabolism, and therefore, to fitness. Taken together, our evolutionary observations
440 backed with metabolic modeling, biochemical and structural data, suggest multiple
441 selection types associated with ecological micro-niches, e.g. environmental cues provided
442 by the human body. Thus, enzyme subfamilies are the result of processes involving
443 different selection types upon proteins with more than one function. Although further
444 examples showing metabolic-driven evolutionary histories need to be identified, our
445 study provides a strategy for the in-depth use of genome sequences for protein and
446 bacterial evolutionary studies to understand enzyme function.
447

448 **Materials and Methods**

449

450 ***Phylogenomic and evolutionary analysis.*** The genomes of the genus *Bifidobacterium*
451 and the family *Actinomyceatceae* were obtained from NCBI (NCBI accession numbers
452 are provided as Figure 3-source data 1). The genomes were annotated by using RAST
453 (22). We identified core orthologous genes using BBHs (29) with a defined e-value of
454 0.001. The sequences were aligned with MUSCLE 3.8.31 (30) and edited with
455 GBLOCKS (31). We concatenated all the orthologous groups for phylogenomic analysis.
456 The phylogenetic analyses were carried out using MrBayes v.3.2.1 (32) and maximum
457 likelihood analysis using RAxML v.8 (33). For MrBayes we used a mixed model, and for
458 the maximum likelihood analysis we used the generalized time reversible (GTR) model.
459 Branch support was measured as the posterior probability of clades in the consensus tree
460 for Bayesian analysis; and with 1,000 bootstrapping replicates in the maximum likelihood
461 analysis. To calculate the nonsynonymous (d_N) and synonymous (d_s) substitution rates
462 between PriA and homologous subfamilies, we aligned all the sequences by codon using
463 RevTrans 1.4 Server (34). To calculate the d_N/d_s ratio we used codeml in the PAML 4
464 package (35). GC content, genome size, CDS content, and number of subsystems
465 between the lineages were compared by using the T-test in the package R. All the
466 boxplots were done with R.

467 The *A. oris* MG-1 strain (25) was sequenced using an in-house Illumina MiSeq
468 sequencing platform. We used Trimmomatic (36) to filter the reads and Velvet v1.2.10
469 (37) to assemble the reads. The Whole Genome Shotgun (WGS) *A. oris* MG-1 project has
470 been deposited at GenBank under the project accession [MAUB00000000].

Metabolic model reconstruction and flux balance analysis. We applied the DOE Systems-biology Knowledgebase (KBase) to construct draft genome-scale metabolic models. The model reconstruction process was optimized as previously (23), and comprised of three steps: (i) genome annotation by RAST (22); (ii) reconstruction of a draft model using the ModelSEED approach (11); and (iii) gapfilling of the model to permit growth and plug holes in mostly complete pathways (38). In the gap-filling process, we identified the minimal set of reactions that could be added to each model to permit biomass production in a media containing every transportable metabolite. We also favored the addition of reactions that would permit more gene-associated reactions in each model to carry flux.

Once models were built, we applied flux balance analysis (FBA) (39) to predict minimal feasible media and classify reactions using a six step process: (i) set the biomass flux to a nonzero value; (ii) minimize the number of active exchange reactions to identify the minimal set of external nutrients that must be provided to permit growth; (iii) constrain exchange fluxes so that only the minimal exchanges are allowed to function; (iv) minimize and maximize each reaction flux to classify each reaction during growth on minimal media (40); (v) maximize biomass flux on minimal media and fix the biomass flux at its maximum value; and (vi) minimize the sum of all fluxes in the model to produce the simplest flux profile possible (e.g. removing all flux loops). Reactions with only positive or negative fluxes are classified as *essential*; reactions with only zero flux values are classified as *nonfunctional*; and reactions with zero and non-zero flux values are classified as *functional*.

494 For construction of the overall model per lineage, we identified all reactions that
495 were associated with genes (i.e. not gapfilled) in at least 75% of the models included in
496 the lineage, using a permissive e-value of 0.01. These reactions formed the basis of our
497 lineage model. Then we applied the same gapfilling algorithm used with our genome
498 models to permit the lineage model to grow. Finally, we applied our FBA pipeline to
499 predict minimal media and classify reactions in the lineage model. All the models,
500 associated genomes, minimal media predictions, reaction classifications, and flux
501 predictions generated in this study are presented using the KBase Narrative Interface and
502 are accessible at https://narrative.kbase.us/narrative/ws.17193.obj.1. See also Figure 4-
503 source data 1, Figure 4-source data 2 and Figure 4-source data 3.
504
505 ***Biochemical analysis of PriA enzymes.*** The *priA* genes from Org15, Org10, and Org41
506 were synthesized by GeneArt (Thermo Fisher Scientific, USA). Additionally, *priA* genes
507 from Org13, Org22, Org34, Org36, and Org39 were synthesized by GenScript
508 (GenScript, USA). Codons were optimized for *E. coli* heterologous expression. The *priA*
509 homologs from *A. oris* MG-1, *B. longum, B. gallicum* and *B. adolescentis* were PCR
510 cloned from our genomic DNA collection. Oligonucleotide sequences of primers used in
511 this study are included in Supplementary file 3. All genes were inserted into pET22b,
512 pET28a (Novagen) for expression and protein purification, and pASK for
513 complementation assays, by using the *Nde*I and *Hind*III restriction sites (18). The *in vivo*
514 *trpF* and *hisA* complementation assays, and *in vitro* determination of the Michaelis-
515 Menten steady-state enzyme kinetics parameters for both PRA and ProFAR as substrates,

were done as previously (13, 14, 26). Lack of enzyme activity *in vitro* was confirmed using active-site saturation conditions, as before (13, 14).

To create a *ΔpriA* mutation in *A. oris* MG1 1.5 Kbp fragments upstream and downstream of this organism were amplified by PCR (Supplementary file 3). The upstream fragment was digested with *EcoR*I and *Nde*I, the downstream fragment with *Nde*I and *Xba*I. The upstream and downstream fragments were ligated together in a single step. The fragment was cloned into pCWU3 precut with *EcoR*I and *Xba*I after digestion with appropriate enzymes. The generated plasmid was then introduced into *A. oris* MG-1 (Org21) by electroporation. Corresponding in-frame deletion mutants were selected by using mCherry fluorescence as a counter-selectable marker and resistance to kanamycin (24). The deletion mutant was confirmed by PCR and by sequencing of the entire genome of the resulting mutant strain.

Crystallization, X-ray data collection, structure determination, and refinement. PriA_Org15 was expressed and produced in BL21 Magic cells bearing the plasmid pMCSG68_PriA_Org15. The protein was purified by immobilized metal-affinity chromatography (IMAC) followed by His6-tag cleavage using recombinant His-tagged TEV protease. A second IMAC step was used to remove the protease, the uncut protein, and the affinity tag. Concentrated protein (37 mg ml^{-1}) was crystallized by sitting-drop vapor-diffusion technique in 96-well CrystalQuick plates (Greiner Bio-One, USA). The crystals appeared at 289 K in conditions consisting of 0.2 M Li_2SO_4, 0.1 M CAPS:NaOH pH 10.5, and 1.2 M NaH_2PO_4/0.8 M K_2HPO_4. Prior to data collection crystals were cryoprotected in 2.4 M K_2HPO_4 and subsequently flash-cooled. Diffraction data were

539 collected at 100 K. Native datasets were collected at 19-ID equipped with an ADSC
540 quantum Q315r CCD detector at 0.979 Å wavelength. The images were processed by
541 using the HKL3000 software suite (41). Molecular replacement was carried out by using
542 the coordinates of PriA from *M. tuberculosis* (27) used as a search probe in Phaser (42).
543 The initial model was then improved by the automatic rebuilding protocol in Arp/wArp,
544 and further modified by iterations of manual rebuilding in COOT (43) and fully
545 anisotropic crystallographic refinement in PHENIX (44) with hydrogen atoms in riding
546 positions. The PriA_Org15 model comprises residues Ser-2-Arg137 and Gly143-Ala247,
547 305 water molecules, 4 phosphate ions, and 1 CAPS moiety. The mFo-DFc difference
548 map reveals two strong positive peaks (near Asp51 and Leu230) that could not be
549 unambiguously assigned. The quality of the refined models was verified using the
550 Molprobity server (45). Data collection statistics and the refinement results are provided
551 as Figure 7-source data 1.
552
553 ***Structural alignment, homology modeling and molecular docking.*** T-coffe package was
554 used for all multiple sequence alignments (46). Protein structural homology models of
555 SubHisA_Org36 and SubTrpF_Org41 were based on the crystal structure of PriA from
556 PriA_Org15 (PDB:4X2R; this study). A standard modeling strategy using Robetta and
557 Rosetta 3.5 (47) was adopted. Molecular models of PRA and ProFAR were built using
558 Molden (48), and optimal atomic configuration of both substrates was obtained using
559 Gaussian 09 (Gaussian Inc., Wallingford CT, USA) through a quantic geometry
560 optimization using a self-consistent field at the Hartree-Fock 6-31G* level. Polar
561 hydrogen atoms and Gasteiger-Marsili empirical atomic partial charges were added using

562 AutoDockTools (49). An extensive configuration sampling of PRA and ProFAR binding
563 biophysical interactions with PriA catalytic site was performed with Autodock Vina (50).
564 Results were merged, refined, clustered, and energy sorted to produce a set of complex
565 configuration predictions.
566
567

Acknowledgments

We acknowledge Nelly Selem-Mójica, Víctor Villa-Moreno, José-Luis Steffani-Vallejo and Christian E. Martínez-Guerrero for bioinformatics support. We thank Sean Rovito and Angélica Cibrián-Jaramillo for useful comments and evolutionary discussions. We thank members of the Structural Biology Center at Argonne National Laboratory for data collection support. This work was supported by Conacyt Mexico, via grants 132376 to MCT and 179290 to FBG, as well as a scholarships to AJV. And by The National Institutes of Health, grant GM094585 to AJ, the US Department of Energy, under contract DE-AC02-06CH11357, and the National Science Foundation, grant 1611952 to CSH, and the National Institute of Dental and Craniofacial Research of the National Institutes of Health, grant DE017382, to HTT.

References

1. Wiedenbeck J, Cohan FM. (2011). Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol Rev*. 35(5):957-76.
2. Blount ZD, Barrick JE, Davidson CJ, Lenski RE. (2012). Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature*. 489(7417):513-8.
3. Hottes AK, Freddolino PL, Khare A, Donnell ZN, Liu JC, Tavazoie S. (2013). Bacterial adaptation through loss of function. *PLoS Genet*. 9:e1003617.
4. Treangen TJ, Rocha EP. (2011). Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet*. 7(1):e1001284.
5. Albalat R, Cañestro C. (2016). Evolution by gene loss. *Nat Rev Genet*. 17(7):379-91.
6. Adams NE, Thiaville JJ, Proestos J, Juárez-Vázquez AL, McCoy AJ, Barona-Gómez F, Iwata-Peuyl D, de Crécy-Lagard V, Maurelli AT. (2014). Promiscuous and adaptable enzymes fill “holes” in the tetrahydrofolate pathway in *Chlamydia* species. *MBio*. 5(4):e01378-14.
7. Price DR, Wilson AC. (2014). A substrate ambiguous enzyme facilitates genome reduction in an intracellular symbiont. *BMC Biol*. 12:110.
8. Wernegreen JJ, Moran NA. Evidence for genetic drift in endosymbionts (*Buchnera*): analyses of protein-coding genes. (1999). *Mol Biol Evol*. 16(1):83-97.

9. McCutcheon JP, Moran NA. (2011). Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol*. 10(1):13-26.
10. Couñago R, Chen S, Shamoo Y. (2006) *In vivo* molecular evolution reveals biophysical origins of organismal fitness. *Mol Cell*. 22(4):441-9.
11. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. (2010) High-throughput generation, optimization, and analysis of genome-scale metabolic models. *Nat Biotechnol*. 28(9): 977-82.
12. Barka EA, Vatsa P, Sanchez L, Gaveau-Vaillant N, Jacquard C, Klenk H-P, Clément C, Ouhdouch Y, van Wezel GP (2016). Taxonomy, physiology, and natural products of Actinobacteria. *Microbiol Mol Biol Rev*. 80(1):1-43.
13. Noda-García L, Camacho-Zarco AR, Medina-Ruíz S, Gaytán P, Carrillo-Tripp M, Fülöp V, Barona-Gómez F. (2013) Evolution of substrate specificity in a recipient's enzyme following horizontal gene transfer. *Mol Biol Evol*. 30(9):2024-34.
14. Verduzco-Castro EA, Michalska K, Endres M, Juárez-Vazquez AL, Noda-García L, Chang C, Henry CS, Babnigg G, Joachimiak A, Barona-Gómez F. (2016) Co-occurrence of analogous enzymes determines evolution of a novel $(\beta\alpha)_8$-isomerase sub-family after non-conserved mutations in flexible loop. *Biochem J*. 473(9):1141-52.
15. Cruz-Morales P, Kopp JF, Martínez-Guerrero C, Yáñez-Guerra LA, Selem-Mojica N, Ramos-Aboites H, Feldmann J, Barona-Gómez F. (2016) Phylogenomic analysis of natural products biosynthetic gene clusters allows

624 discovery of arseno-organic metabolites in model streptomycetes. *Genome Biol*
625 *Evol*. 8(6):1906-16.
626 16. Zhao K, Li W, Kang C, Du L, Huang T, Zhang X, Wu M, Yue B. (2014)
627 Phylogenomics and evolutionary dynamics of the family *Actinomycetaceae*.
628 *Genome Biol Evol*. 6(10):2625-33.
629 17. Barona-Gómez F, Hodgson DA. (2003). Occurrence of a putative ancient-like
630 isomerase involved in histidine and tryptophan biosynthesis. *EMBO Rep*.
631 4(3):296-300.
632 18. Noda-García L, Juárez-Vázquez AL, Ávila-Arcos MC, Verduzco-Castro EA,
633 Montero-Morán G, Gáytan P, Carrillo-Tripp M, Barona-Gómez F. (2015).
634 Insights into the evolution of enzyme substrate promiscuity after the discovery of
635 $(\beta\alpha)_8$ isomerase evolutionary intermediates from a diverse metagenome. *BMC*
636 *Evol Biol*. 15:107.
637 19. Plach MG, Reisinger B, Sterner R, Merkl R. (2016). Long-term persistence of bi-
638 functionality contributes to the robustness of microbial life through exaptation.
639 *PLoS Genet*. 12(1):e1005836.
640 20. Yeung MK. Molecular and genetic analyses of *Actinomyces spp*. (1999). *Crit Rev*
641 *Oral Biol Med*. 10(2):120-38.
642 21. Könönen E, Wade WG. (2015). *Actinomyces* and related organisms in human
643 infections. *Clin Microbiol Rev*. 28(2):419-42.
644 22. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K,
645 Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL,
646 Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich

C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. (2008). The RAST server: rapid annotations using subsystems technology. *BMC Genomics*. 9:75.

23. Satish Kumar V, Dasika MS, Maranas CD. (2007). Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*. 8:212.
24. Wu C, Ton-That H. (2010). Allelic exchange in *Actinomyces oris* with mCherry fluorescence counterselection. *Appl Environ Microbiol*. 76(17):5987-9.
25. Delisle AL, Nauman RK, Minah GE. (1978). Isolation of a bacteriophage for *Actinomyces viscosus. Infect Immun*. 20(1):303-6.
26. Noda-García L, Camacho-Zarco AR, Verdel-Aranda K, Wright H, Soberón X, Fülöp V, Barona-Gómez F. (2010). Identification and analysis of residues contained on beta alpha loops of the dual-substrate (beta alpha)$_8$ phosphoribosyl isomerase A specific for its phosphoribosyl anthranilate isomerase activity. *Protein Sci*. 19(3):535-43.
27. Düe AV, Kuper J, Geerlof A, von Kries JP, Wilmanns M. (2011) Bisubstrate specificity in histidine/tryptophan biosynthesis isomerase from *Mycobacterium tuberculosis* by active site metamorphosis. *Proc Natl Acad Sci USA*. 108(9):3554-9.
28. Do T, Henssge U, Gilbert SC, Clark D, Beighton D. (2008). Evidence for recombination between a sialidase (*nanH*) of *Actinomyces naeslundii* and *Actinomyces oris*, previously named "*Actinomyces naeslundii* genospecies 1 and 2". *FEMS Microbiol Lett*. 288(2):156-62
29. Tatusov RL, Koonin EV, Lipman DJ. (1997). A genomic perspective on protein families. *Science*. 278(5338):631-7.

30. Edgar RC. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*. 5:113.
31. Castresana J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*. 17(4):540-52.
32. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huetsenbeck JP. (2012). MrBayes 32: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*. 61(3):539-42.
33. Stamatakis A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 30(9):1312-3.
34. Wernersson R, Pedersen AG. (2003). RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res*. 31(13):3537-9.
35. Yang Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24(8):1586-91.
36. Bolger AM, Lohse M, Usadel B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 30(15):2114-20.
37. Zerbino DR, Birney E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. 18(5):821-9.
38. Dreyfuss JM, Zucker JD, Hood HM, Ocasio LR, Sachs MS, Galagan JE. (2013). Reconstruction and validation of a genome-scale metabolic model for the filamentous fungus *Neurospora crassa* using FARM. *PLoS Comput Biol*. 9(7): e1003126

692 39. Orth JD, Thiele I, Palsson BO. (2010). What is flux balance analysis? *Nat*
693 *Biotechnol*. 28(3): 245-248.
694 40. Mahadevan R, Schilling CH. (2003). The effects of alternate optimal solutions in
695 constraint-based genome-scale metabolic models. *Metab Eng*. 5(4): 264-276.
696 41. Minor W, Cymborowski M, Otwinowski Z, Chruszcz M. (2006). HKL-3000: the
697 integration of data reduction and structure solution--from diffraction images to an
698 initial model in minutes. *Acta Crystallogr D Biol Crystallogr*. 62: 859-66.
699 42. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ.
700 (2007). Phaser crystallographic software. *J Appl Cryst*. 40: 658-74.
701 43. Emsley P, Cowtan K. (2004). Coot: model-building tools for molecular graphics.
702 *Acta Crystallogr D Biol Crystallogr*. 60: 2126-32.
703 44. Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, Headd JJ,
704 Hung LW, Kapral GJ, Grosse-Kunstleve RW, McCoy, AJ, Moriarty NW, Oeffner
705 R, Read RJ, Richardson D. C, Richardson JS, Terwilliger TC, Zwart PH.
706 (2010). PHENIX: a comprehensive Python-based system for macromolecular
707 structure solution. *Acta Crystallogr D Biol Crystallogr*. 66: 213-21.
708 45. Chen V, Arendall W, Headd J, Keedy D, Immormino R, MolProbity: all-atom
709 structure validation for macromolecular crystallography. (2010). *Acta Crystallogr*
710 *D Biol Crystallogr*. 66(1):12-21.
711 46. Notredame C, Higgins DG, Heringa J. (2000). T-Coffee: A novel method for fast
712 and accurate multiple sequence alignment. *J Mol Biol*. 302(1):205-17.
713 47. Rohl CA, Strauss CE, Misura KM, Baker D. (2004). Protein structure prediction
714 using Rosetta. *Methods Enzymol*. 383:66-93.

48. Schaftenaar G, Noordik JH. (2000). Molden: a pre- and post-processing program for molecular and electronic structures. *J Comput Aided Mol Des*. 14(2):123-34.

49. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ. (2009) AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem*. 30(16):2785-91.

50. Trott O, Olson AJ. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*. 31(2):455-61.

724

725

726

727 **Table 1.** Selective pressures in PriA homologs from H1, H2 and H3 hypotheses

Hypothesis	d_N/d_s	d_N	d_s
H1	0.0636	0.3151	4.9559
H2	0.0901	1.8687	20.736
H3	0.1459	1.8703	12.8227

728

Table 2. Biochemical characterization of PriA, SubHisA2 and SubTrpF homologs

Enzymes	*In vivo* activity		*In vitro* activity [a]					
			ProFAR isomerase (HisA)			PRA isomerase (TrpF)		
	HisA	TrpF	K_M (μM)	k_{cat} (s^{-1})	k_{cat}/K_M ($s^{-1}\mu M^{-1}$)	K_M (μM)	k_{cat} (s^{-1})	k_{cat}/K_M ($s^{-1}\mu M^{-1}$)
PriA_Org3_*B. longum*	+	+	2.7 ± 0.5	0.4 ± 0.1	0.1	6.1 ± 0.1	2.1 ± 0.5	0.3
PriA_Org1_*B. gallicum*	+	+	1.7 ± 0.3	0.3 ± 0.1	0.2	40 ± 9	3.5 ± 0.1	0.09
PriA_Org6_*B. adolescentis*	+	+	17 ± 4.3	2.3 ± 0.01	0.1	21± 5	0.9 ± 0.2	0.04
PriA_Org15_*A. urogenitalis*	+	+	4.0 ± 0.9	0.2 ± 0.03	0.04	23 ± 6.5	0.5 ± 0.05	0.02
PriA_Org22_*A. sp. oral* taxon 171	+	+	3 ± 0.3	0.3 ± 0.09	0.1	8 ± 2	0.4 ± 0.1	0.04
PriA_Org21_*A. oris* MG-1	+	+	10 ± 2	0.2 ± 0.09	0.02	30 ± 7	0.3 ± 0.03	0.01
SubHisA2_Org34_*A. vaccimaxillae*	+	+						
SubHisA2_Org36_*A. cardiffensis*	+	+	56 ± 17	0.14 ± 0.05	0.002	n.d.	n.d.	n.d.
SubTrpF_Org10_*A. sp. oral* taxon 848	-	+	n.d.	n.d.	n.d.	n.d.	n.d.	0.0001
SubTrpF_Org13_*A. graevenitzii*	-	+						
SubTrpF_Org39_*A. sp. oral* taxon 180	-	+						
SubTrpF_Org41_*A. odontolyticus*	-	+	n.d.	n.d.	n.d.	8.5 ± 0.9	0.15 ± 0.06	0.02

[a] Each data point comes from at least three independent determinations using freshly purified enzyme. n.d., activity not detected, even using active-site saturation conditions. Empty entries reflect our inability to properly express and/or solubilize these proteins. The detection limits for the PRA and ProFAR isomerase assay used in the present study are 0.0001 $\mu M^{-1}s^{-1}$ and 0.001 $\mu M^{-1}s^{-1}$, respectively (13,14,26).

738 **Legends to Figures**
739

740 **Figure 1 - $(\beta\alpha)_8$ barrel isomerases at which L-tryptophan and L-histidine biosynthesis converge.**
741 Selected L-tryptophan (blue) and L-histidine (red) biosynthetic enzymes are shown. The committed
742 reaction catalyzed by PriA and PriB, or phosporibosyl isomerase A or B in *Actinobacteria* (dashed
743 arrows), is independently catalyzed by the enzymes TrpF or PRA isomerase, and HisA or ProFAR
744 isomerase (standard arrows) in most bacteria. Furthermore, the divergent SubHisA enzyme, resulting
745 from divergent evolution after an event of HGT and positive selection in certain *Corynebacterium*
746 species, is also shown.
747

748 **Figure 2. Identification of reduced genomes in *Actinobacteria*. A.** Protein-based phylogeny of 133
749 representative deep-branching *Actinobacteria* using Bayesian reconstruction. The tree shows a clade with
750 the family *Bifidobacteriaceae* as the root of the families *Dermabacteraceae*, *Cellulomonadaceae*,
751 *Demequinaceae*, *Jonesiaceae*, *Promicromonosporaceae* and *Actinomycetaceae*, shown in blue and
752 highlighted with a grey box. **B.** Relationship between genome size and percentage of (G+C) content. The
753 color key used for taxonomic associations is provided at the bottom, and it is the same for both panels.
754

755 **Figure 3. Concatenated phylogenetic tree of the family *Actinomycetaceae* and occurrence of L-**
756 **histidine and L-tryptophan biosynthetic genes.** The tree was constructed using 205 single-copy
757 conserved proteins using Bayesian methods. Only posterior probabilities are shown but significant
758 bootstrap values close to 100 using maximum likelihood were also calculated (Fig. 3 - supplement Figure
759 1). A new classification of the family, into four major groups, is proposed: lineage I (orange); lineage II
760 (blue); lineage III, (green); and lineage IV (red). Based in the species phylogenetic tree of Fig. 2A, we

761 selected as out-group the genus *Bifidobacterium*. Occurrence of L-histidine (His, black) and L-tryptophan
762 (Trp, grey) biosynthetic genes as revealed by standard genome annotation using RAST is shown next to
763 the tree. Each square represents a complete pathway including all expected genes (10 and 7 for the *his*
764 and *trp* genes respectively) up to 90%. The only missing *his* gene refers to the enzyme histidinol-
765 phosphatase (EC 3.1.3.15), which belongs to a broad enzyme family difficult to annotate.
766
767 **Figure 3 - supplement Figure 1. Concatenated phylogenetic tree of the family Actinomycetaceae**
768 **using maximum likelihood.** The phylogenetic tree shows four major groups: lineage I (orange clade),
769 lineage II (blue clade), lineage III (green clade) and lineage IV (red clade). Bootstrap values are shown.
770 Representative sequences from the closely related genus *Bifidobacterium* were used as out-group (purple
771 clade). Name nomenclature of organisms used in the text is provided. Both phylogenetic trees using
772 different algorithms supported the taxonomic relationships between the selected taxa.
773
774 **Figure 3 - supplement Figure 2. Lineage-specific genomic features of the familiy *Actinomycetaceae*.**
775 Lineages and the out-group are as defined in the text. **A.** Percentage of (G+C)-content in whole genomes,
776 where each point in the box-plot represent a genome. **B.** Genome size (Mbp), where each point in the
777 box-plot represents a genome. **C.** Number of coding sequences, where each point in the box-plot
778 represents a CDS. **D.** Number of subsystems based in RASTtk annotation, where each point represents a
779 subsystem. The median is highlighted with a black line. All the box-plots were created with R package.
780

781 **Figure 4**. **Metabolic diversity amongst the genomic lineages of the genus *Actinomyces*.** The Venn
782 diagrams show the overlap in gene-associated reactions included in models of genomes of lineage I (A),
783 lineage II (B), lineage III (C), and lineage IV (D). The diagrams for lineages I and III show the overlap of
784 all models in these lineages, while the lineage II and IV diagrams show the overlap of a subset of models
785 sampled based on their metabolic diversity. Overlap in gene-associated reaction content for each of the
786 core lineage models (E), which are comprised of conserved reactions present in at least 75% of the
787 models in each lineage, is also shown.
788
789 **Figure 4 - supplement Figure 1.** Phylogenetic projection of amino acid biosynthetic pathways
790 throughout the family Actinomycetaceae as confirmed after genome-scale metabolic modeling. The tree
791 is the same as in Fig. 3, but without the names of organisms to facilitate visual inspection. Occurrence of
792 amino acid biosynthetic pathways, denoted with standard nomenclature, is shown. Each square represents
793 a complete pathway including all expected genes up to 90%. His and Trp pathways are highlighted with
794 an asterisk.
795
796 **Figure 5. Phylogenetic reconstruction of PriA and coevolution with L-histidine and L-tryptophan**
797 **biosynthesis. A.** Analysis of the occurrence of *his* and *trp* biosynthetic genes (*priA* is included in both
798 pathways), marked as absent (white) or present (gray), using the phylogenomics species tree of Fig. 3 as a
799 map (same color code). The missing *his* gene, when almost the entire pathway is present, refers to the
800 enzyme histidinol-phosphatase (EC 3.1.3.15), which belongs to a broad enzyme family difficult to
801 annotate. **B**. Same gene occurrence analysis using the PriA phylogenetic tree as a map**.** Three
802 evolutionary scenarios where PriA is coevolving with the occurrence of *his* and *trp* genes, and in
803 agreement with the intensity of purifying selection (Table 1, gradient shown in the left-hand side of the

804 panel), are marked as H1 (purple), H2 (orange), and H3 (yellow). The same color code as in Fig. 3 is
805 used, and the selected enzymes that were biochemically characterized are underlined.
806
807 **Figure 6. Evolutionary hypotheses and steady-state enzyme kinetics of PriA homologs. A.**
808 Evolutionary hypothesis (H1, H2 and H3) with functional implications leading to PriA enzyme
809 subfamilies, expressed as biochemical conversions, as obtained from Fig. 5. **B.** Comparison of the
810 catalytic efficiencies ($kcat/K_M$) of selected enzymes from different scenarios, including the three
811 postulated evolutionary hypotheses. Values for ProFAR (*x* axis) and PRA (*y* axis) isomerase activities,
812 expressed as log10, are compared. Data from PriAs of *Bifidobacterium* (purple circle), PriA from H1
813 (purple triangle), SubHisA2 from H2 (orange), and SubTrpF from H3 (yellow pentagon) is included.
814
815 **Figure 7. PriA from Org15_*A. urogenitalis* active site and sequence alignment of PriA sub-families.**
816 **A.** The structure of PriA from *A. urogenitalis* (purple, PDB: 4X2R) superimposed with PriA from *M.*
817 *tuberculosis* in a complex with rCdRP (cyan, PDB: 2Y85) and PrFAR (pink, PDB: 2Y88) is used to
818 illustrate the position of the respective substrates. The catalytic residues and those critical for divergence
819 into SubHisA2 or SubTrpF are shown. Since the loop contributing Trp139 and Arg137 is mostly
820 disordered, and Arg137 itself does not adopt substrate binding-relevant position in the structure from *A.*
821 *urogenitalis*, only the equivalent elements from the *M. tuberculosis* homolog are shown. **B**. Multiple
822 sequence alignment of PriA (purple), SubHisA2 (red) and SubTrpF (bold) sequences. Catalytic residues,
823 Asp9 and Asp169, are marked in red. PRA and ProFAR binding residues are shown in blue. SubHisA2
824 and SubTrpF loss-of-function residues are framed. The secondary structure is shown below the
825 sequences. Loops are shown in orange, α helixes are shown in gray and β sheets are shown in green.
826 Sequence corresponding to loops 1, 5, and 6 is highlighted in gray.

List of Tables provided as Source Data

Figure 2-source data 1. Actinobacterial genome sequences from early-diverging families used in this study

Figure 2-source data 2. Conserved orthologs in early-diverging actinobacterial families used for phylogenetic reconstruction

Figure 3-source data 1. Genome sequences of the familiy Actinomycetaceae and the genus *Bifidobacterium* used in this study

Figure 3-source data 2. Conserved orthologs between the family Actinomycetaceae and the genus *Bifidobacterium* and best fit model used to construct the phylogenetic tree with Mr.Bayes

Figure 3-figure supplement 2-source data 1. Statistical analysis of the genomic differences between Lineage II and IV

Figure 4-source data 1. Model overview

Figure 4-source data 2. Model reactions

Figure 4-source data 3. Predicted minimal media

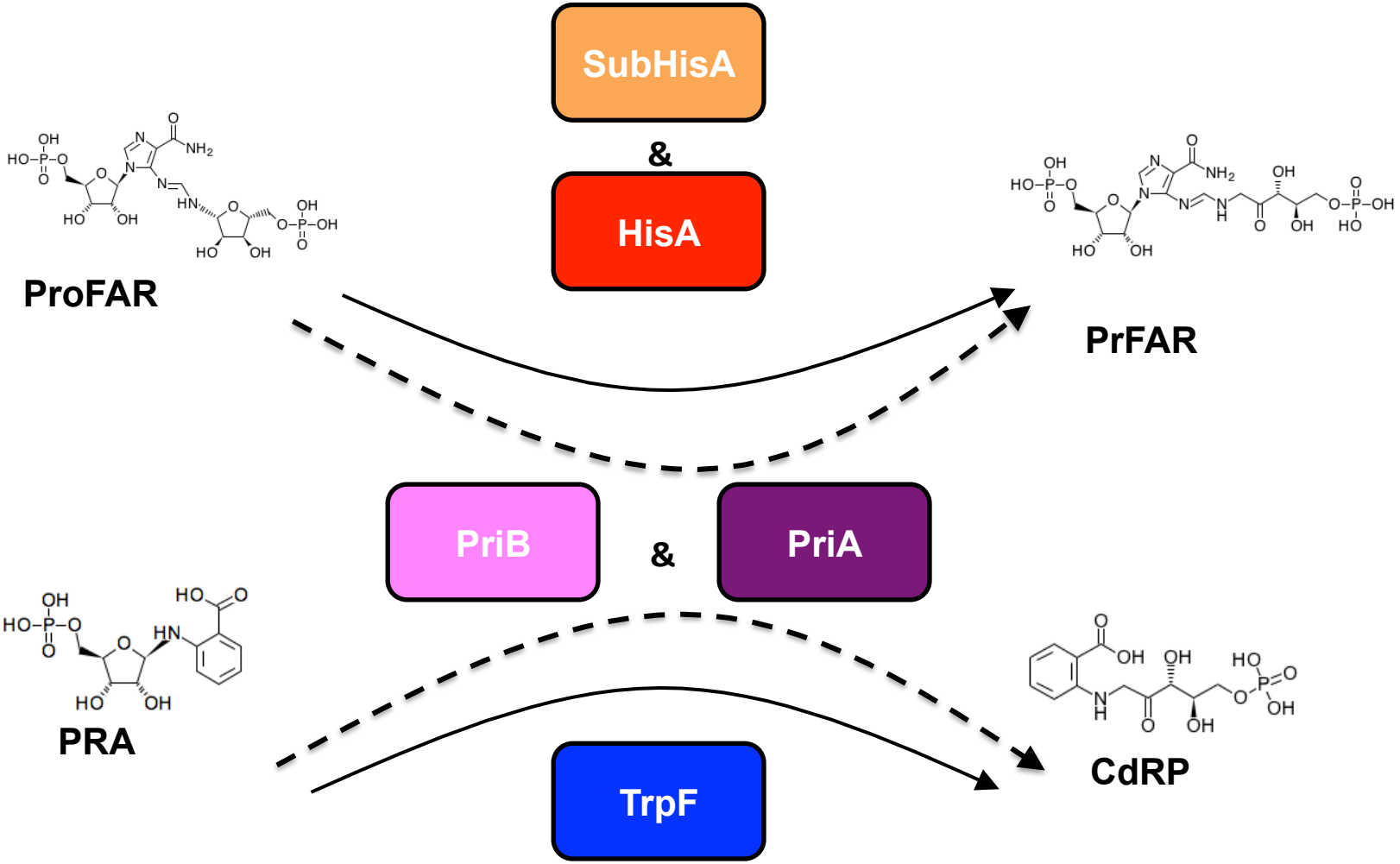
Figure 5-source data 1. Occurrence of L-Histidine and L-Tryptophan biosynthetic enzymes throughout the family Actinomycetaceae and the genus *Bifidobacterium*

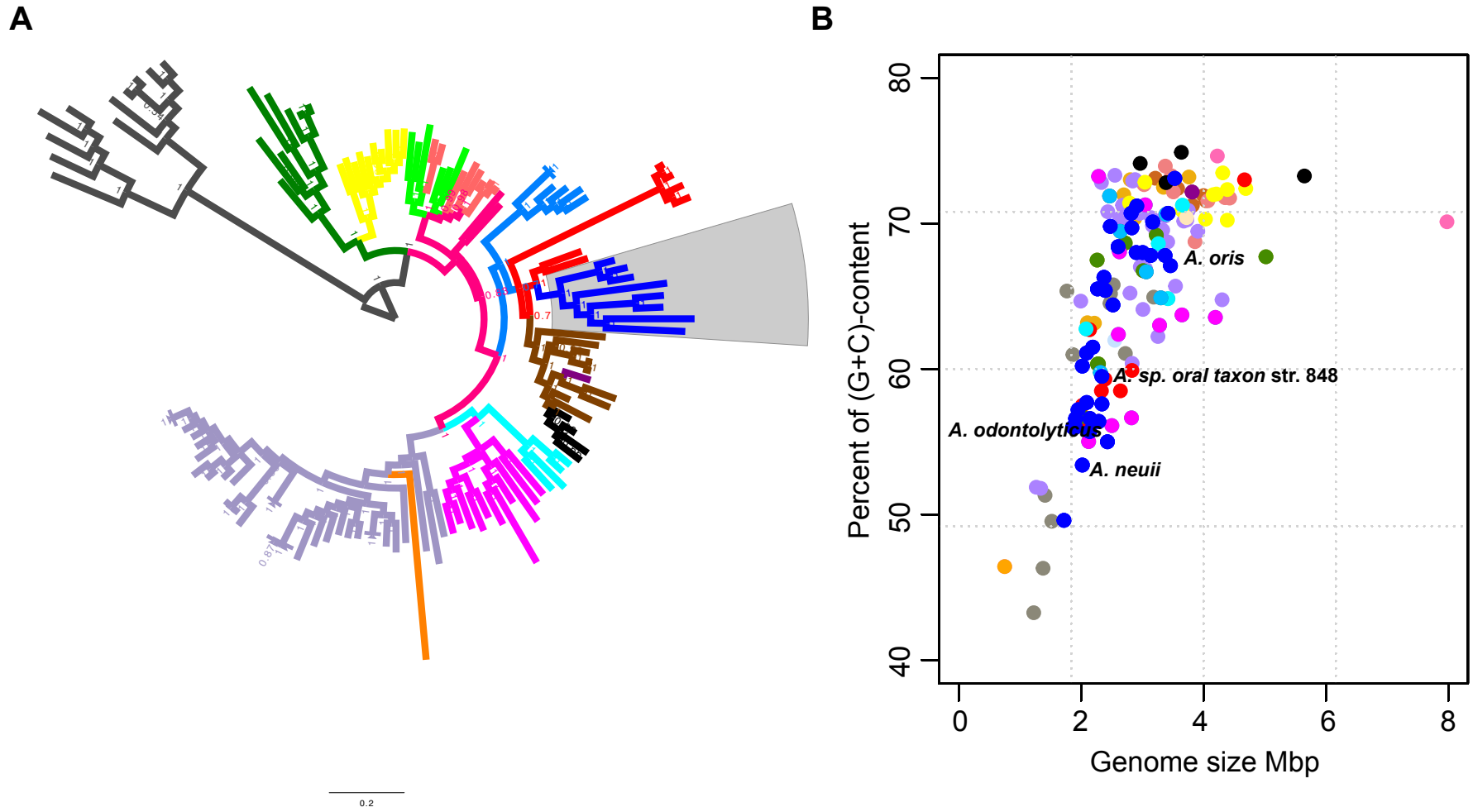
Figure 7-source data 1. X-ray crystalographic data processing and refinement statistics for PriA_Org15.

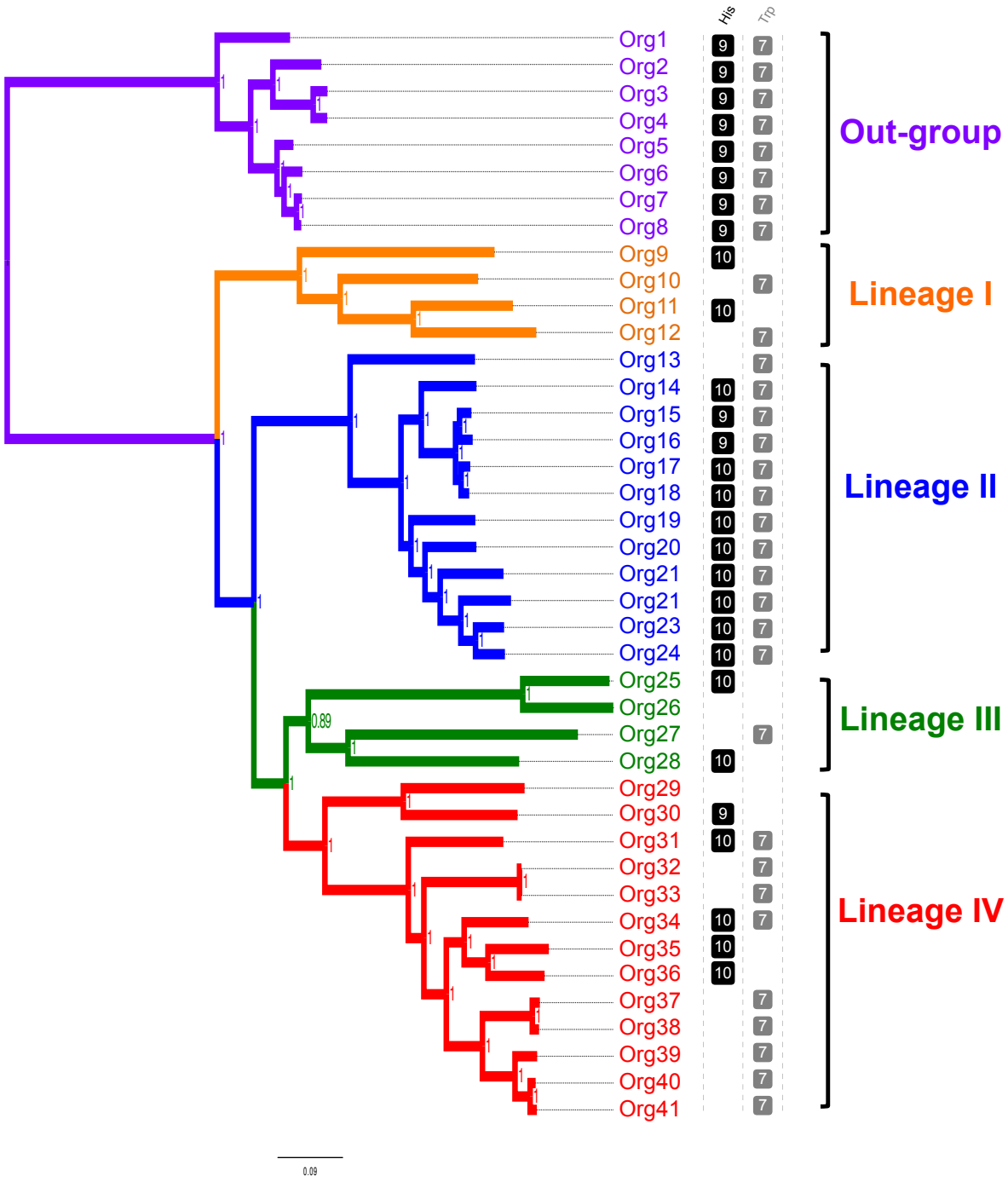
Supplementary file 1. Genome analysis of the priA minus *Actinomyces oris* mutant

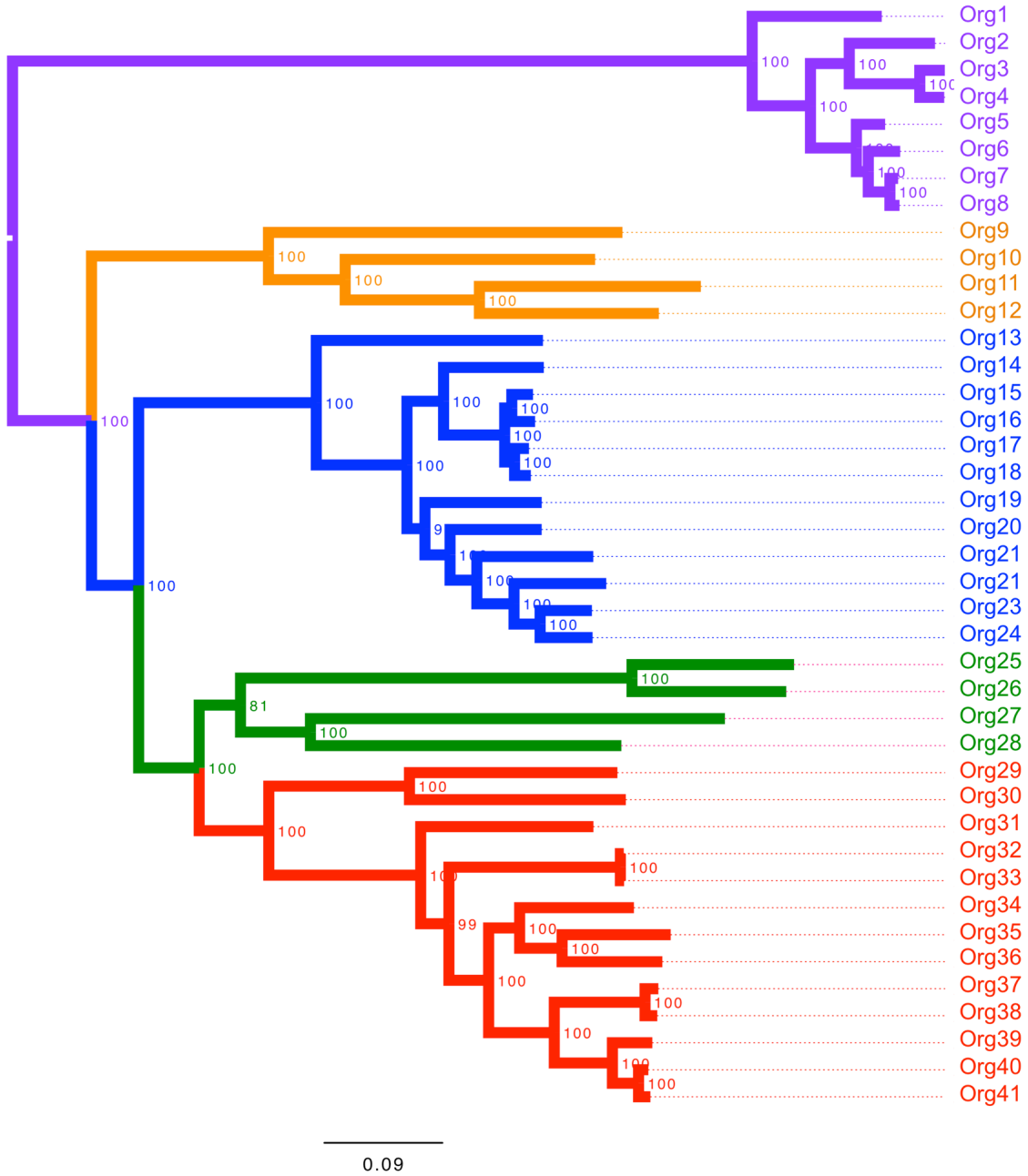
Supplementary file 2. Predicted affinities for PRA and ProFAR

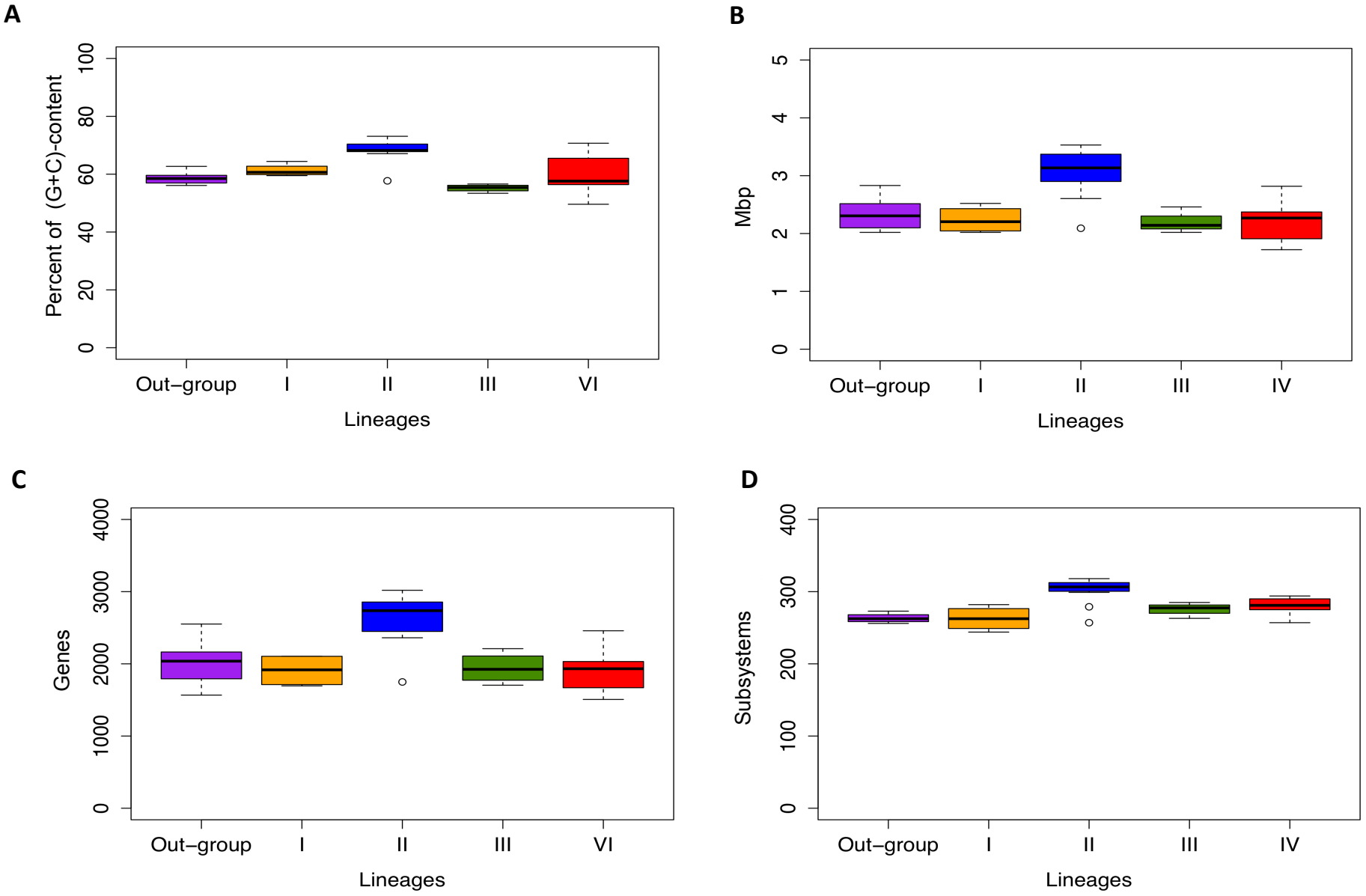
Supplementary file 3. Primers used in this study

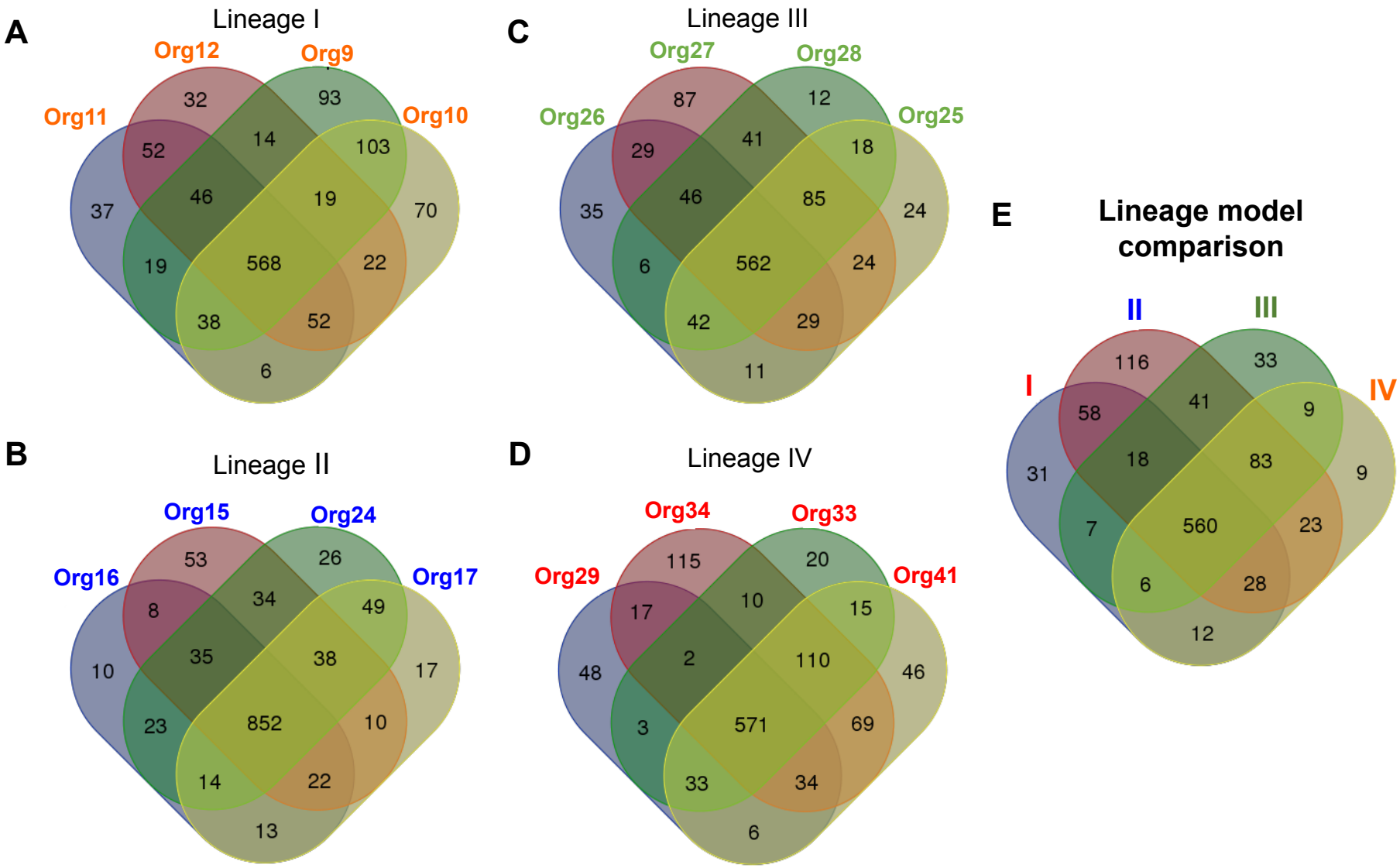

SubHisA
&
HisA
ProFAR
PrFAR
PriB
&
PriA
PRA
CdRP
TrpF

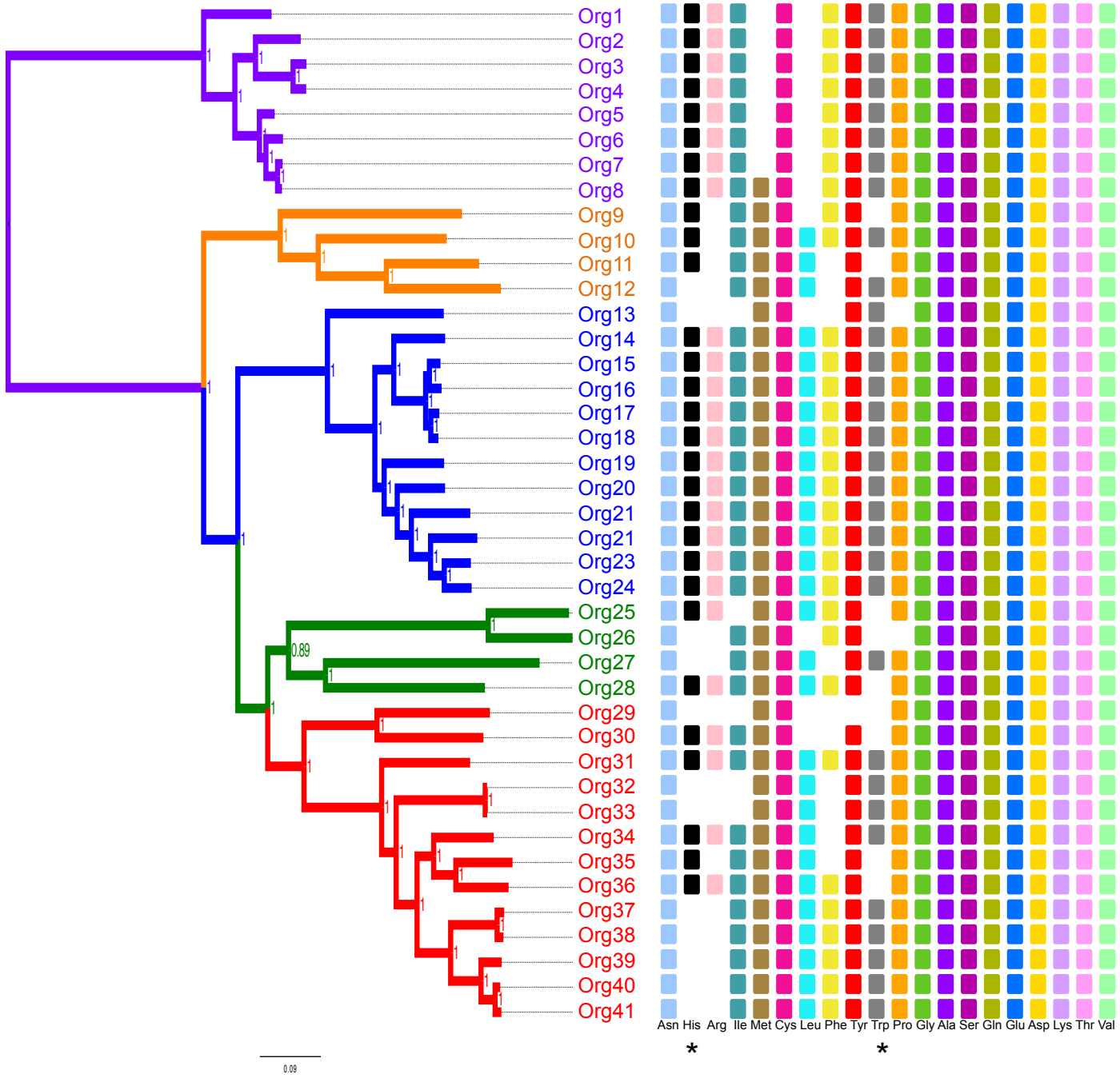

A
B
0.2
Percent of (G+C)-content
Genome size Mbp
40
50
60
70
80
0
2
4
6
8
A. oris
A. sp. oral taxon str. 848
A. odontolyticus
A. neuii
Brevibacteriaceae
Bifidobacteriaceae
Actinomycetaceae
Tropheryma whippelii
Sanguibacteraceae
Kineosporaceae
Intrasporangiaceae
Dermacoccaceae/Dermatophilaceae
Dermabacteraceae
Coriobacteriaceae/Acidothermaceae
Cellulomonadaceae/Demequinaceae/
Timonella/Jonesiaceae
Propionibacteriaceae
Promicromonosporaceae
Nocardiodaceae
Micrococcaceae
Microbacteriaceae

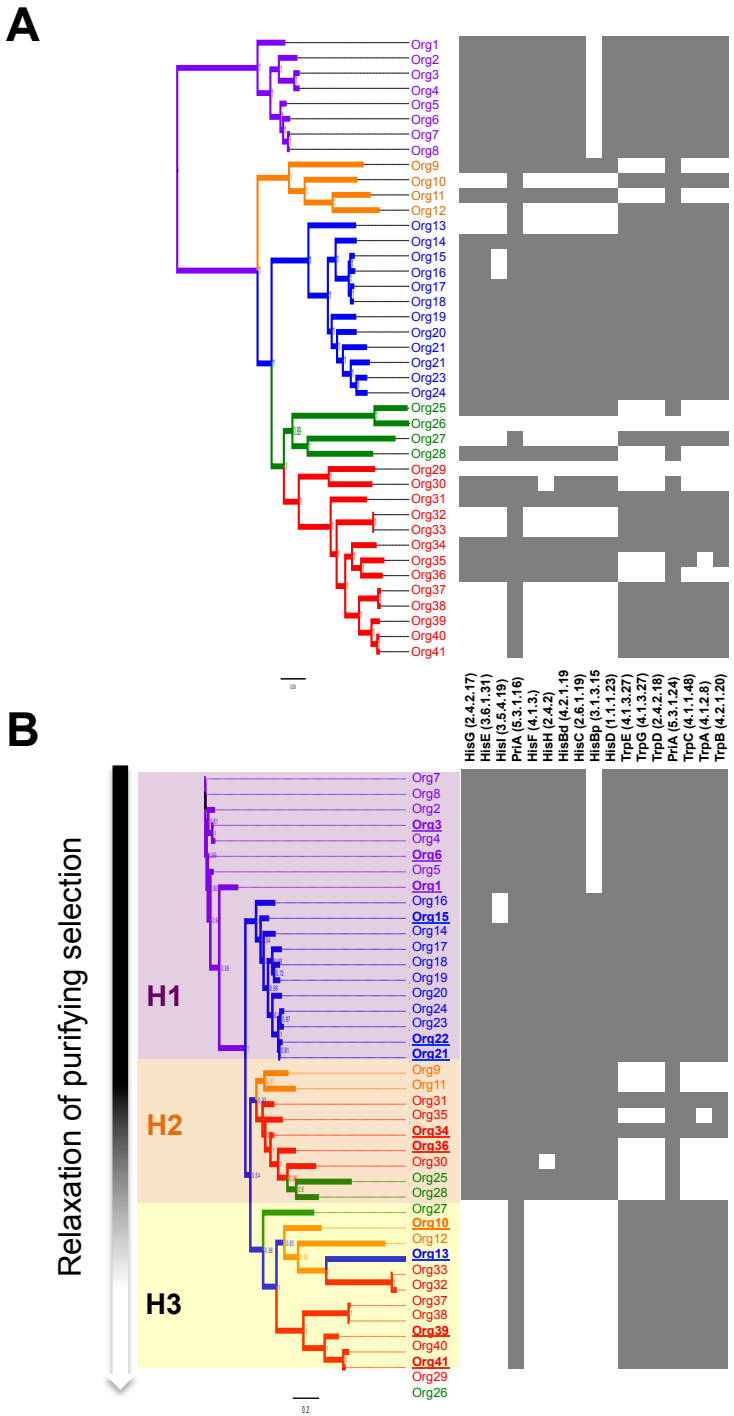

His
Trp
Org1
Org2
Org3
Org4
Org5
Org6
Org7
Org8
Org9
Org10
Org11
Org12
Org13
Org14
Org15
Org16
Org17
Org18
Org19
Org20
Org21
Org21
Org23
Org24
Org25
Org26
Org27
Org28
Org29
Org30
Org31
Org32
Org33
Org34
Org35
Org36
Org37
Org38
Org39
Org40
Org41
0.89
0.09
Out-group
Lineage I
Lineage II
Lineage III
Lineage IV

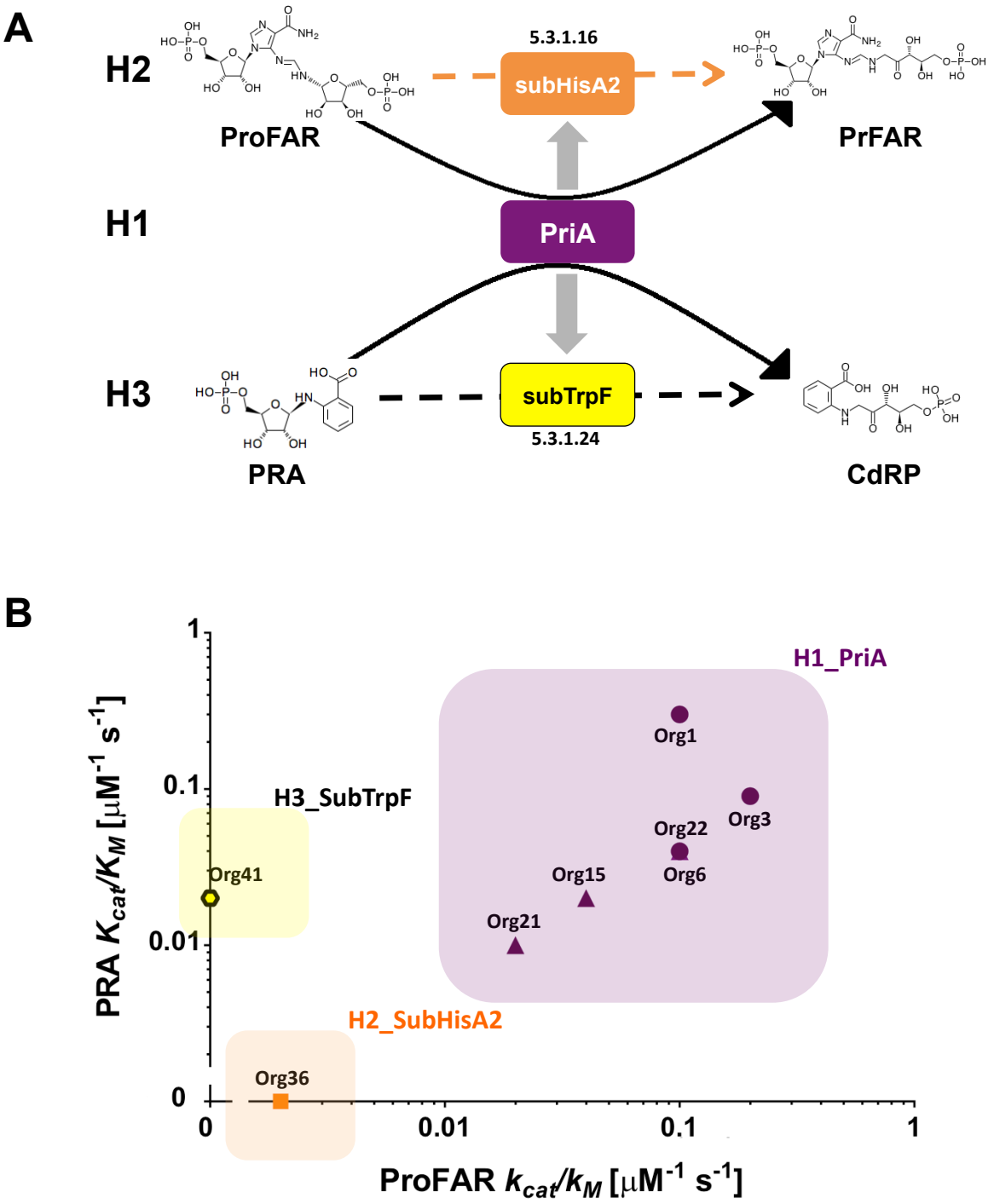

Org1
Org2
Org3
Org4
Org5
Org6
Org7
Org8
Org9
Org10
Org11
Org12
Org13
Org14
Org15
Org16
Org17
Org18
Org19
Org20
Org21
Org21
Org23
Org24
Org25
Org26
Org27
Org28
Org29
Org30
Org31
Org32
Org33
Org34
Org35
Org36
Org37
Org38
Org39
Org40
Org41
0.09


A
Percent of (G+C)-content
0
20
40
60
80
100
Out−group
I
II
III
VI
Lineages
B
Mbp
0
1
2
3
4
5
Out−group
I
II
III
IV
Lineages
C
Genes
0
1000
2000
3000
4000
Out−group
I
II
III
VI
Lineages
D
Subsystems
0
100
200
300
400
Out−group
I
II
III
IV
Lineages

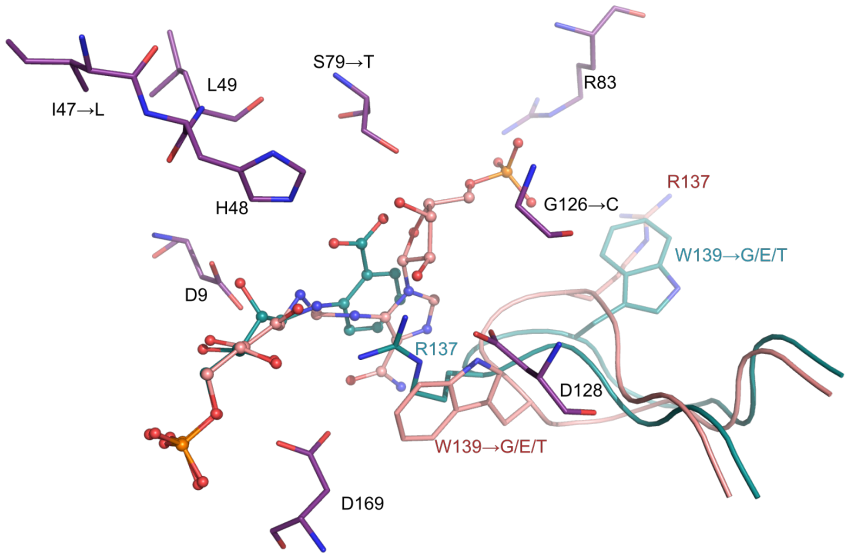

A
Lineage I
Org12
Org9
Org11
Org10
32
93
52
14
103
37
46
19
70
19
568
22
38
52
6
B
Lineage II
Org15
Org24
Org16
Org17
53
26
8
34
49
10
35
38
17
23
852
10
14
22
13
C
Lineage III
Org27
Org28
Org26
Org25
87
12
29
41
18
35
46
85
24
6
562
24
42
29
11
D
Lineage IV
Org34
Org33
Org29
Org41
115
20
17
10
15
48
2
110
46
3
571
69
33
34
6
E
Lineage model comparison
II
III
I
IV
116
33
58
41
9
31
18
83
9
7
560
23
6
28
12


Org1
Org2
Org3
Org4
Org5
Org6
Org7
Org8
Org9
Org10
Org11
Org12
Org13
Org14
Org15
Org16
Org17
Org18
Org19
Org20
Org21
Org21
Org23
Org24
Org25
Org26
Org27
Org28
Org29
Org30
Org31
Org32
Org33
Org34
Org35
Org36
Org37
Org38
Org39
Org40
Org41
0.89
0.09
Asn His Arg Ile Met Cys Leu Phe Tyr Trp Pro Gly Ala Ser Gln Glu Asp Lys Thr Val
*
*


A
Org1
Org2
Org3
Org4
Org5
Org6
Org7
Org8
Org9
Org10
Org11
Org12
Org13
Org14
Org15
Org16
Org17
Org18
Org19
Org20
Org21
Org21
Org23
Org24
Org25
Org26
Org27
Org28
Org29
Org30
Org31
Org32
Org33
Org34
Org35
Org36
Org37
Org38
Org39
Org40
Org41
B
HisG (2.4.2.17)
HisE (3.6.1.31)
HisI (3.5.4.19)
PriA (5.3.1.16)
HisF (4.1.1.)
HisH (2.4.2)
HisBd (4.2.1.19)
HisC (2.6.1.19)
HisBp (3.1.3.15)
HisD (1.1.1.23)
TrpE (4.1.3.27)
TrpG (4.1.3.27)
TrpD (2.4.2.18)
PriA (5.3.1.24)
TrpC (4.1.1.48)
TrpA (4.1.2.8)
TrpB (4.2.1.20)
Relaxation of purifying selection
H1
H2
H3
Org7
Org8
Org2
Org3
Org4
Org6
Org5
Org1
Org16
Org15
Org14
Org17
Org18
Org19
Org20
Org24
Org23
Org22
Org21
Org9
Org11
Org31
Org35
Org34
Org36
Org30
Org25
Org28
Org27
Org10
Org12
Org13
Org33
Org32
Org37
Org38
Org39
Org40
Org41
Org29
Org26


A
H2
5.3.1.16
subHisA2
ProFAR
PrFAR
H1
PriA
H3
subTrpF
5.3.1.24
PRA
CdRP
B
H1_PriA
Org1
Org3
Org22
Org6
Org15
Org21
H3_SubTrpF
Org41
H2_SubHisA2
Org36
PRA K_{cat}/K_M [μM^{-1} s^{-1}]
ProFAR k_{cat}/k_M [μM^{-1} s^{-1}]
1
0.1
0.01
0
0
0.01
0.1
1

A



B

```
1
PriA_Mtuber      ---MPLILLPAVDVVEGRAVRLVQGKAGSQTEYGSAVDAALGWQRDGAEWIHLVDLDAAFGRGSNHELLAEVVG-KL---
PriA_Scoe        -M-SKLELLPAVDVRDGQAVRLVHGESGTETSYGSPLEAALAWQRSGAEWLHLVDLDAAFGTGDNRALIAEVAQ-AM---
PriA_Org15       ----MLTLLPAVDVADGKAVRLLQGEAGSETDYGSPIEAARDWVEAGAEWIHLVDLDAAFGRGSNAPLLERIVG-EV---
PriA_Org21       ----MLTLLPAVDVADGKAVRLLQGEVGSETDYGSPVDAARDWVRAGAEWIHLVDLDAAFGRGSNHELLARIVG-EV---
PriA_Org22       ----MLTLLPAVDVADGKAVRLLQGAIGSETDYGSPVDAARDWVGAGAAWIHLVDLDAAFGRGSNHELLARIVG-EV---
subHisA2_Org34   MS-ANLILLPAVDVVDGQAVRLTQGEAGTETNYGHPLEAARSFVEAGAQWLHLVDLDAAFGRGSNAPLLADITR-EL---
subHisA2_Org36   MS--TLILLPAVDVVNGQAVRLTQGQAGTETVYGTPLEAARSFVEAGAKWLHLVDLDAAFGRGSNAELLQSITA-QL---
subTrpF_Org10    MT-LPLQLLPAVDVADGRSVRLTRGEASSACSFGDPMRAVADFVEAGAAWIHLADIDAAFGRGSNRALLTEIVR-EA---
subTrpF_Org13    MAVGPLRLLPAVDVANGLAVTHRTSAGGDAGAGISALDACLRWVEAGADWIHLVDLDAAFGRGSNAALLAQVIADLARLH
subTrpF_Org39    MN--RIELLPAIDVTGGRAVRLSSGVVD-DRSWADPAQVARSFEEAGARWVHLVDLDRAFGRGNNDELLARVMN-EV---
subTrpF_Org41    MT--ILELLPAIDVTGGQAVRLSSGVID-EGSWGSPIDVARSFDEAGARWVHLVDLDLAFGRGENSELLARVIR-EV---

                 β1          β-α1          α1  α-β1 β2   β-α2   α2   α-β2
80
PriA_Mtuber      ----DVQVELSGGIRDDESLAAALATGCARVNVGTAALENPQWC-ARVIGEHGDQVAVGLDVQIIDGEHRLRG-------
PriA_Scoe        ----DIKVELSGGIRDDDTLAAALATGCTRVNLGTAALETPEWV-AKVIAEHGDKIAVGLDVRG----TTLRG-------
PriA_Org15       ----GIKVELSGGIRDDASLTRALKAGAARVNLGTAALEDPQWT-ARVIAEHGEKIAVGLDVRG----TTLAA-------
PriA_Org21       ----GIKVELSGGIRDDASLARALSAGAARVNLGTAALEDPEWT-ERVIAEHGEKIAVGLDVRG----STLAA-------
PriA_Org22       ----GIKVELSGGIRDDASLARALSAGAARVNLGTAALEDPEWT-ERVIAEHGEKIAVGLDVRG----STLAA-------
subHisA2_Org34   ----PINVELSGGIRDDESLRRALDAGARRVNLGTAALEDPEWT-ERVIAEFGDRIAVGLDVRG----ETLSA-------
subHisA2_Org36   ----PINVELTGGIRDDESLRRALECGARRVNLGTAAIENPEWT-EKVIGEFADRIALGLDVRG----ETLAG-------
subTrpF_Org10    QTRHGVRIEWSGGVRDEESLLAAVASGAARVNLATGALADLEWA-ASAIERFGSQVAVCLDVRG----DVLAA-------
subTrpF_Org13    P---GVSVQWSGGVSSADDVERALAAGAKRVNLGAGALKDLAATTALVGRF-GRHLNVCLDVSAASAAPNPAAPADPATP
subTrpF_Org39    ----DVAIQLSGGIVSRGDVEAALEAGPDRVNIATQALDDLDAV-RDAIDAFGPRVSVCLDVRG----ERLAA-------
subTrpF_Org41    ----PVRVELSGGITSPAAVEAGLAMGPERVNIATQALDDIDAV-CEAVDTFGERVAVCLDVRG----DRLAA-------

                 β3   β-α3  α3 α-β3  β4  β-α4     α4    α-β4  β5    β-α5
160
PriA_Mtuber      ------------------------------RGWETDGGDLWDVLERLDSEGCSRFVVTDITKDGTLGGPNLDLLAGVADR
PriA_Scoe        ------------------------------RGWTRDGGDLYETLDRLNKEGCARYVVTDIAKDGTLQGPNLELLKNVCAA
PriA_Org15       ------------------------------RGWTKEGGDLWQTLDRLNEAGCRRYVVTDVTKDGTLTGPNTELLRQVAAR
PriA_Org21       ------------------------------RGWTKEGGDLWESLERLNAAGCARYVVTDVTRDGTLSGPNTALLTEVCQR
PriA_Org22       ------------------------------RGWTKEGGDLWETLERLNTAGCARYVVTDVTRDGTLSGPNTALLTEVCQR
subHisA2_Org34   ------------------------------RGWTRDGGNLFETIERLDAAGCSRYVVTDVARDGMLSGPNTELLRRVCEA
subHisA2_Org36   ------------------------------RGWTTQGPNLFETIARFDAAGCARYVVTDVSRDGMLSGPNLELLARLCEA
subTrpF_Org10    ------------------------------RGESAEVGRLWDVLPALEEAGCARYVVTDVARDGAMNGPNTELLRKVAAA
subTrpF_Org13    GAAQLAGAQRGAAQPATQPSADLATYVVHPRGQGGPVGPLEPILAALNEAGTGAYVVTDRVRDGALSGPNLPLLGALSGA
subTrpF_Org39    ------------------------------RGTSREGGNVWEVLSALNEAGIARLVVTDVTRDGQMRGANLELLARVADA
subTrpF_Org41    ------------------------------RGGSGEGGNVWEALRVLDEAGVARLVVTDVTRDGQMNGSNRELLARVADQ

                                 β-α5                α5   α-β5 β6  β-α46   α6
```