BMC Biology

BioMed Central

Research article

Open Access

Inter-genomic displacement via lateral gene transfer of bacterial *trp* operons in an overall context of vertical genealogy

Gary Xie[1], Carol A Bonner[2], Jian Song[1], Nemat O Keyhani[2] and Roy A Jensen*[1,2]

Address: [1]Los Alamos National Laboratory, Los Alamos, New Mexico, 87544, USA and [2]Department of Microbiology & Cell Science, University of Florida, PO Box 110700, Gainesville, Florida, 32611, USA

Email: Gary Xie - xie@lanl.gov; Carol A Bonner - cbonner@ufl.edu; Jian Song - jian@lanl.gov; Nemat O Keyhani - keyhani@ufl.edu; Roy A Jensen* - rjensen@ufl.edu

* Corresponding author

Published: 23 June 2004

BMC Biology 2004, **2**:15 doi:10.1186/1741-7007-2-15

Received: 27 January 2004
Accepted: 23 June 2004

This article is available from: http://www.biomedcentral.com/1741-7007/2/15

Abstract

Background: The growing conviction that lateral gene transfer plays a significant role in prokaryote genealogy opens up a need for comprehensive evaluations of gene-enzyme systems on a case-by-case basis. Genes of tryptophan biosynthesis are frequently organized as whole-pathway operons, an attribute that is expected to facilitate multi-gene transfer in a single step. We have asked whether events of lateral gene transfer are sufficient to have obscured our ability to track the vertical genealogy that underpins tryptophan biosynthesis.

Results: In 47 complete-genome *Bacteria*, the genes encoding the seven catalytic domains that participate in primary tryptophan biosynthesis were distinguished from any paralogs or xenologs engaged in other specialized functions. A reliable list of orthologs with carefully ascertained functional roles has thus been assembled and should be valuable as an annotation resource. The protein domains associated with primary tryptophan biosynthesis were then concatenated, yielding single amino-acid sequence strings that represent the entire tryptophan pathway. Lateral gene transfer of several whole-pathway *trp* operons was demonstrated by use of phylogenetic analysis. Lateral gene transfer of partial-pathway *trp* operons was also shown, with newly recruited genes functioning either in primary biosynthesis (rarely) or specialized metabolism (more frequently).

Conclusions: (i) Concatenated tryptophan protein trees are congruent with 16S rRNA subtrees provided that the genomes represented are of sufficiently close phylogenetic spacing. There are currently seven tryptophan congruency groups in the *Bacteria*. Recognition of a succession of others can be expected in the near future, but ultimately these should coalesce to a single grouping that parallels the 16S rRNA tree (except for cases of lateral gene transfer). **(ii)** The vertical trace of evolution for tryptophan biosynthesis can be deduced. The daunting complexities engendered by paralogy, xenology, and idiosyncrasies of nomenclature at this point in time have necessitated an expert-assisted manual effort to achieve a correct analysis. Once recognized and sorted out, paralogy and xenology can be viewed as features that enrich evolutionary histories.

Background

The process of LGT

Lateral gene transfer (LGT) undoubtedly occurs with high frequency [1]. But what is required for any given LGT event to escape transience and actually survive as a non-vertical contributor to the evolutionary history of a species? (**i**) Initially the incoming gene(s) must be replicated in the original recipient cell and its immediate progeny. (**ii**) Selective advantages of the LGT-gene(s) must be sufficient to foster eventual domination of the species population. (**iii**) During this time a crucial factor influencing survival will be whether the considerable demands imposed by the recipient genome for amelioration of alien genes can be met [2]. Obviously, there will be less amelioration pressure if the donor and recipient genomes are phylogenetically close and have similar guanine/cytosine base ratios, dinucleotide frequencies, promotor motifs, and so on. Even here a LGT insertion into a genome might be contra-selected for other reasons, for example, if the location of the insertion disrupts the symmetry and physical balance between the origin and terminus points of replication [3]. There is a balance in that one can expect alien sources of greatest novelty to be phylogenetically distant, yet genes from such remote sources will usually confront the greatest amelioration pressures. The most obvious candidates as successfully imported alien genes will encode novel functions that confer clear selective value, such as resistance to threatening environmental agents (e.g. antibiotics) or ability to utilize a new source of carbon and energy.

Enhanced probability of LGT success can be expected if only a single gene is needed for the novel function. This does not necessarily rule out the acquisition of a new function that is complex and multi-genic. Sometimes where existing functions are complex and multi-step, a single incoming gene can create a new metabolic linkage. For example, an organism having a multi-step pathway for tyrosine catabolism could acquire a previously unavailable ability to also catabolize phenylalanine through import of a gene encoding phenylalanine hydroxylase (which converts phenylalanine to tyrosine). In cases where the total repertoire of multiple steps that define a novel function are all absent in a given organism, acquisition of that function by LGT would be highly improbable were it not for the existence of operon modules in prokaryotes. Lawrence has proposed, in fact, that gene organization as operons largely exists because of "selfish" properties that operate at the hierarchical level of genes [4,5]. Although genes of *L*-tryptophan (Trp) biosynthesis, a classic operon system, are completely dispersed in some prokaryote genomes (e.g. *Aquifex*, *Chlorobium*, *Wolinella* and unicellular cyanobacteria such as *Synechococcus*, *Synechocystis*, and *Prochlorococcus*), they usually exist either as whole-pathway operons or as a combination of several partial-pathway operons [6].

Approaches for the detection of LGT events are either phylogenetic or parametric. Parametric approaches include the detection of nucleotide composition, dinucleotide frequencies and codon-usage biases in gene segments that are atypical of the recipient genome. Such parametric analysis is well illustrated by a study of the *Escherichia coli* genome [1]. However, such atypical parametric properties will drift with time toward those of the recipient genome (amelioration). Therefore, parametric analysis is limited to detection of genes that were acquired recently. We have found only a single example where *trp* genes exhibited parametric features suggesting LGT. In this case a low-GC gene block in *Xylella fastidiosa* contains seven genes, of which two encode the subunits of anthranilate synthase and one encodes a repressor, TrpR [7]. On the other hand, phylogenetic analysis can detect ancient events of LGT, and this approach has allowed the recognition of a number of whole-pathway and partial-pathway *trp* operons that were acquired by LGT. Presumably, the initial aberrant parametric properties associated with LGT have been ameliorated. Phylogenetic analysis also is a much more powerful indicator of the likely donor lineage following gene acquisition by LGT.

En bloc *importation of primary pathways via LGT*

Ubiquitous pathways of primary biosynthesis, having a long history of genomic optimization and metabolic integration, are generally improbable candidates for replacement by LGT [8]. Among the challenges confronting successful LGT are that key regulatory genes may be spatially separated from the structural genes, and promoters and transcription signals vary between bacterial species, as also is the case for elements required for translation.

Exceptions can be envisioned, among them: (**i**) A gene of primary biosynthesis could, in fact, be synonymous with a gene of antibiotic resistance. Thus, if one or more primary-pathway enzymes is the target of an antimicrobial agent, then genes in nature that encode resistant versions of that enzyme(s) could confer strong selective pressure for replacement. The apparent displacement of an essential enzyme of isoprenoid biosynthesis (hydroxymethylglutaryl coenzyme A reductase) in *Archaea* by a statin-resistant enzyme of bacterial origin exemplifies this [9]. (**ii**) If primary pathway genes are lost by deletion, as often occurs with pathogens or symbionts, re-acquisition of the pathway might later become advantageous. In this case, no native pathway having a sophisticated history of amelioration is present to out-compete a pathway of alien origin. In a sense, the recipient has reverted to a pristine evolutionary state that is probably subject to instability and rapid change. (**iii**) An alien suite of enzymes might

Table 1: Key to nomenclature[a] used

New gene name	Prior gene name	Protein domain encoded
trpAa	*trpE*	Anthranilate synthase: aminase subunit (α)
trpAb	*trpG*	Anthranilate synthase: amidotransferase subunit (β)
trpB	*trpD*	Anthranilate phosphoribosyl transferase
trpC	*trpF*	Phosphoribosyl-anthranilate isomerase
trpD	*trpC*	Indoleglycerol phosphate synthase
trpEa	*trpA*	Tryptophan synthase, α subunit
trpEb	*trpB*	Tryptophan synthase, β subunit

[a]The nomenclature is at the level of catalytic domain in the order of reaction steps in the pathway. See [6] for a detailed rationale supporting the new nomenclature. Overall reactions of tight complexes are assigned one capital letter, with α and β subunits assigned the corresponding lowercase 'a' and 'b' respectively. The convention of a bullet denotes a fusion, for example, *trpD•trpC*.

provide extraordinary properties of catalysis and/or regulation that are of sufficient selective value to offset a lack of ameliorative history.

The pathway of tryptophan biosynthesis

Amino acid biosynthetic pathways, such as the Trp pathway (see Table 1 for nomenclature used), are not novel in the sense that they are ancient and widely distributed. When the Trp pathway is absent in prokaryotes, it is a consequence of gene loss (reductive evolution), usually, if not always, in pathogens or symbionts whose hosts or symbiont partners supply the Trp required. Table 2 provides a current listing of microbial genomes that have lost some or all of the *trp* genes. Such reductive changes can be quite recent, as illustrated by existence of some *trp*-gene degradation in *Yersinia pestis* KIM, but not in the other two completely sequenced *Y. pestis* genomes. Genomes having incomplete *trp* pathways are presumably in an intermediate state of genome reduction. However, see Xie *et al.* [10,11] for novel functional innovations of "incomplete" Trp pathways in chlamydiae. For example, the addition of two genes to the partial-pathway operon of *Chlamydophila psittaci* has been asserted to have created a novel operon responsible for a kynurenine-to-Trp pathway that is important in overcoming a mechanism of host defense during pathogenesis [1]. This hypothesis has recently been confirmed experimentally [12]. Also, see Barona-Gòmez and Hodgson [13] for their resolution of the mystery of why *trpC* is absent in the clade of actinomycete organisms, that is, they identified *priA* as a replacement for the otherwise universal *trpC*. A thorough overview of the phenomenon of "missing genes in metabolic pathways" and how this can be approached via comparative genomics has been provided by Osterman and Overbeek [14].

Trp is biochemically the most expensive of the amino acids synthesized by prokaryotes. Accordingly, it is not surprising that Trp biosynthesis is usually regulated with fine-tuned precision. Different organisms deploy completely different modes of regulation, for example, compare those of *E. coli* and *Bacillus subtilis* [15,16]. *Pseudomonas aeruginosa* exhibits yet a third distinctive system of regulation, part of which involves an activator gene (*trpI*) [17]. The sophisticated, complex, and highly distinctive regulation in each of the latter three organisms appears to be of recent origin based upon the relatively narrow clades possessing these particular regulatory systems. **This is a very important point in support of the thesis** [6] **that modern, sophisticated genomes may be much less prone to displacement by LGT of *trp* genes than were ancient genomes.** In this context, it will be quite important to determine whether organisms that lack known modes of regulation based upon current model organisms really have relatively unsophisticated control mechanisms, or whether unknown regulatory mechanisms are in place. For example, it has been initially surprising that *trp* genes of *Streptomyces coelicolor* are not regulated by feedback repression. However, an excellent and detailed study [18] has demonstrated the existence of regulation that is both growth phase-dependent and growth rate-dependent. This has been attributed to the oligotrophic lifestyle of *Streptomyces*. It would be interesting to know whether the clade defined by the *Streptomyces* mode of regulation is also narrow (and therefore of recent origin).

Identical enzyme steps operating in different pathways

We refer to *trp* genes that produce Trp in general support of protein synthesis as genes of **primary** metabolism. Those *trp* genes responsible for the production of intermediates (or Trp) for any other purpose (e.g. as precursors of antibiotics or pigments) are referred to as genes of **specialized** (or **secondary**) metabolism. There is ample precedent for maintenance in a given genome of co-existing structural genes whose gene products catalyze the same reaction, but which function in differing temporal or spatial modes in different pathways. Such genes may be dif-

Table 2: Complete genomes where Trp-pathway genes were lost via reductive evolution

Bacterial organisms	No pathway	Incomplete pathway
Bdellovibrio bacteriovorus	√	
Borrelia burgdorferi	√	
Chlamydia muridarum		√
Chlamydia trachomatis		√
Chlamydophila pneumoniae	√	
Chlamydophila psittaci[a]	√	√
Clostridium difficile	√	
Clostridium perfringens	√	
Clostridium tetani	√	
Coxiella burnetti		√
Enterococcus faecalis	√	
Fusobacterium nucleatum		√
Haemophilus ducreyi	√	
Lactobacillus johnsonii	√	
Mycoplasma genitalium	√	
Mycoplasma mycoides	√	
Mycoplasma pneumoniae	√	
Phytoplasma asteris	√	
Porphyromonas gingivalis	√	
Rickettsia prowazekii	√	
Streptococcus agalactiae	√	
Streptococcus equi	√	
Streptococcus pyogenes	√	
Treponema denticola	√	
Treponema pallidum	√	
Tropheryma whipplei	√	
Ureaplasma urealyticum	√	
Wigglesworthia glossinidia[b]	√	
Wolbachia sp.[b]	√	
Yesinia pestis KIM		√

[a] But see [10] for a description of how an incomplete Trp pathway has been joined with other genes to yield a mosaic operon with a novel function.
[b] Insect endosymbiont.

ferentially regulated by distinctive control mechanisms or mechanisms that accomplish spatial separation. For example, *S. coelicolor* possesses *trp* paralog genes of identical catalytic function that exist in separate operons dedicated to primary biosynthesis, on the one hand, or to calcium-dependent antibiotic (CDA) production, on the other hand [6,19]. It is interesting that, in such cases, one or more of the pathway genes sometimes exist as a single copy and, therefore, must be shared between both functions.

Phylogenetic trees for proteins require a continuum of close relatives

This study is primarily a phylogenetic analysis and depends upon protein trees. However, sequences of proteins such as the Trp enzymes are not nearly as conserved as 16S rRNA, and it is well known that they are of limited value for making phylogenetic inferences over wide phylogenetic distances [20]. On the other hand, for a group of very closely related organisms, the 16S rRNA sequences can be so similar that there is a limited basis for discriminating order of branching. Here some protein trees may yield more refined branching relationships over short phylogenetic distances. Until recently, there have been relatively few sequenced genomes available that would provide a critical mass of closely related organisms as a source of Trp-protein sequences. The sequencing of new genomes is now beginning to provide phylogenetic groups of ever more dense genome representation. Health-related organisms heavily influence priorities, and wide gaps in the currently available tree exist. Phylogenetic regions where genome representation is sparse will undoubtedly persist for many years, and it would be helpful if new genomes for sequencing were selected specifically to fill phylogenetic gaps. One can anticipate that a given protein tree will progressively increase its useful phylogenetic span of discrimination as the gaps between regions of sufficient genome representation are filled.

Additional daunting problems (which are minimal for 16S rRNA trees) assail the phylogenetic validity of protein trees. These are: (**i**) unrecognized paralogy and (**ii**) xenology. Ancient paralogs that arose in a common cell will usually have diverged greatly in the contemporary organisms that house them. If one or the other paralog has been lost in various lineages in an erratic fashion and the genealogical history of these losses is not recognized, this situation has been termed "unrecognized paralogy". Differential loss of ancient paralogs in different lineages has the potential to place surviving homolog proteins of distantly related organisms closer to one another than to the surviving proteins of closely related organisms. In this case, xenology can be erroneously inferred. Although LGT has been increasingly recognized for genes in general, genes encoding 16S rRNA are thought to be recalcitrant to LGT (but see Gogarten *et al.* [21] and Doolittle [22].

Congruency and implied cases of LGT

Since particular lineages have optimized Trp biosynthesis to overall demands of primary and secondary metabolism that are both qualitatively and quantitatively individualistic, it seems that abandonment of native *trp* genes in favor of imported alien genes would not occur very often. If correct, then one would expect that the vertical trace of evolutionary descent for *trp* genes engaged in primary Trp biosynthesis would be demonstrable, as has indeed been shown following a very comprehensive analysis [6]. In the latter study, individual Trp-protein trees were found to be generally congruent with 16S rRNA trees in subtree regions that were supported by sufficiently dense phylogenetic representation.

Against this background of congruency (and, indeed, enabled **because** of the overall congruency), several clear examples were found of LGT of whole-pathway *trp* operons in which all of the native *trp* genes were displaced, occasionally leaving a surviving remnant behind. In addition, examples were found of LGT of partial-pathway *trp* operons that either displaced native genes of primary biosynthesis (one case) or became associated with a secondary function. In the literature relevant to LGT there is frequently little evidence about likely donor genomes, and even less indications of direction of transfer [23]. This paper provides detailed information and analysis not given in [6], and it provides an especially thorough documentation in that the occasional LGT transposition of the entire suite of seven genes responsible for Trp biosynthesis are evaluated. Importantly, we were able to identify donor lineages, to identify a gene remnant of the pre-existing Trp system in three divergent descendants of one of the recipient genomes, and to identify in one genome several evolutionary steps of specific operon perturbation that resumed in the vertical genealogy following LGT.

Results

Figure 1 presents a 16S rRNA tree of 47 finished-genome organisms from the domain *Bacteria*. In preliminary work with the individual protein trees corresponding to the seven catalytic domains of Trp biosynthesis, we noticed that at least seven subtree blocks on the Trp-protein trees tended to be congruent with corresponding subtree blocks (differentially highlighted in Figure 1) of the 16S rRNA tree. The Gram-negative proteobacteria command special attention in this paper simply for the fortuitous reason that the greatest density of sequenced genomes is to be found in proteobacteria. The divisions of the proteobacteria are labelled in the lower-right portion of Figure 1. *Geobacter* and *Desulfovibrio* (delta-proteobacteria) are very divergent lineages and do not contribute to a common tryptophan congruency group. The grey area (epsilon proteobacteria) represents a subtree region that will probably become a tryptophan congruency group when more dense genome representation becomes available in this region. Another subtree region (not shown on Figure 1) in this category is that represented by the chlamydiae. This latter subtree region is temporarily limited for Trp analysis because it is currently represented by genomes having Trp pathways that are incomplete or absent. Beta-proteobacteria fall into a common tryptophan congruency group with some gamma-proteobacteria ("upper" gamma-proteobacteria), whereas the remaining "lower" gamma-proteobacteria (enteric lineage) form a separate and distinct tryptophan congruency group.

Seven tryptophan congruency groups (TCGs)

Individual Trp-protein domains vary in size and degree of conservation and thus confer varied amounts of phylogenetic information. Figure 2 shows a protein tree in which the seven protein domains responsible for the Trp pathway of biosynthesis have been concatenated in the order of the pathway steps (TrpAa/TrpAb/TrpB/TrpC/TrpD/TrpEa/TrpEb) prior to use of the tree-building program. The seven-domain Trp tree was constructed by using 47 organisms having a complete Trp pathway **and by application of an analysis aimed at exclusion of divergent paralogs and xenologs that do not function in primary Trp biosynthesis** (excluded paralogs and xenologs can be viewed in the succeeding figures). Each of the seven protein sections of a given concatenate has an impact on the overall tree that is roughly proportional to the number of amino acids encoded. Each of the seven nodes defining TCGs is supported by a bootstrap value of 100% (in contrast to the weakness of individual Trp-protein trees). Although the **overall** concatenated tree does not parallel the **overall** 16S rRNA tree, seven subtree regions of the concatenated tree do parallel 16S rRNA subtree regions. These seven groupings are not necessarily equivalent in phylogenetic span. TCG-6, for example, is a small, closely related grouping compared with the much

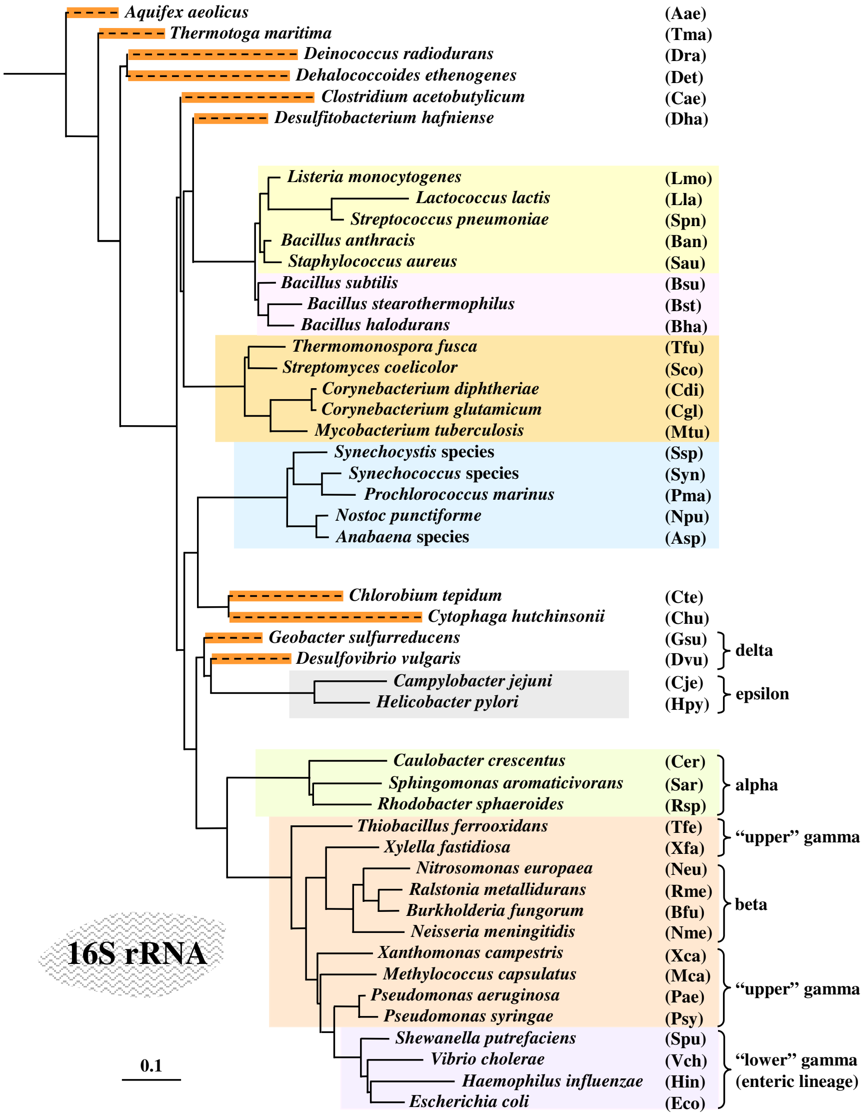
Aquifex aeolicus (Aae)
Thermotoga maritima (Tma)
Deinococcus radiodurans (Dra)
Dehalococcoides ethenogenes (Det)
Clostridium acetobutylicum (Cae)
Desulfitobacterium hafniense (Dha)
Listeria monocytogenes (Lmo)
Lactococcus lactis (Lla)
Streptococcus pneumoniae (Spn)
Bacillus anthracis (Ban)
Staphylococcus aureus (Sau)
Bacillus subtilis (Bsu)
Bacillus stearothermophilus (Bst)
Bacillus halodurans (Bha)
Thermomonospora fusca (Tfu)
Streptomyces coelicolor (Sco)
Corynebacterium diphtheriae (Cdi)
Corynebacterium glutamicum (Cgl)
Mycobacterium tuberculosis (Mtu)
Synechocystis species (Ssp)
Synechococcus species (Syn)
Prochlorococcus marinus (Pma)
Nostoc punctiforme (Npu)
Anabaena species (Asp)
Chlorobium tepidum (Cte)
Cytophaga hutchinsonii (Chu)
Geobacter sulfurreducens (Gsu)
Desulfovibrio vulgaris (Dvu)
delta
Campylobacter jejuni (Cje)
Helicobacter pylori (Hpy)
epsilon
Caulobacter crescentus (Ccr)
Sphingomonas aromaticivorans (Sar)
Rhodobacter sphaeroides (Rsp)
alpha
Thiobacillus ferrooxidans (Tfe)
Xylella fastidiosa (Xfa)
"upper" gamma
Nitrosomonas europaea (Neu)
Ralstonia metallidurans (Rme)
Burkholderia fungorum (Bfu)
Neisseria meningitidis (Nme)
beta
Xanthomonas campestris (Xca)
Methylococcus capsulatus (Mca)
Pseudomonas aeruginosa (Pae)
Pseudomonas syringae (Psy)
"upper" gamma
Shewanella putrefaciens (Spu)
Vibrio cholerae (Vch)
Haemophilus influenzae (Hin)
Escherichia coli (Eco)
"lower" gamma
(enteric lineage)
16S rRNA
0.1

Figure 1

Positioning of 47 complete-genome organisms on a 16S rRNA tree (phylogram view). Among these, seven 16S rRNA subtree regions are color-coded to facilitate comparison with the TCG regions on the Trp-protein concatenate tree of Fig. 2. Dashed lines in orange indicate organisms representing lineages where genomes of close relatives have not yet been sequenced. A grey box indicates a region of minimal current genome representation that is expected to become a region of subtree congruency. The subdivisions of the Proteobacteria are labelled at the lower right. Note that it is an idiosyncracy of tree presentation that Tfe and Xfa appear to group more closely with beta-proteobacteria than with other gamma-proteobacteria. However, it is a distance tree and close inspection of the distances reveal the identity of the gamma-proteobacteria as a single, cohesive group. The horizontal bar corresponds to 0.1 substitutions per site on the distance tree.

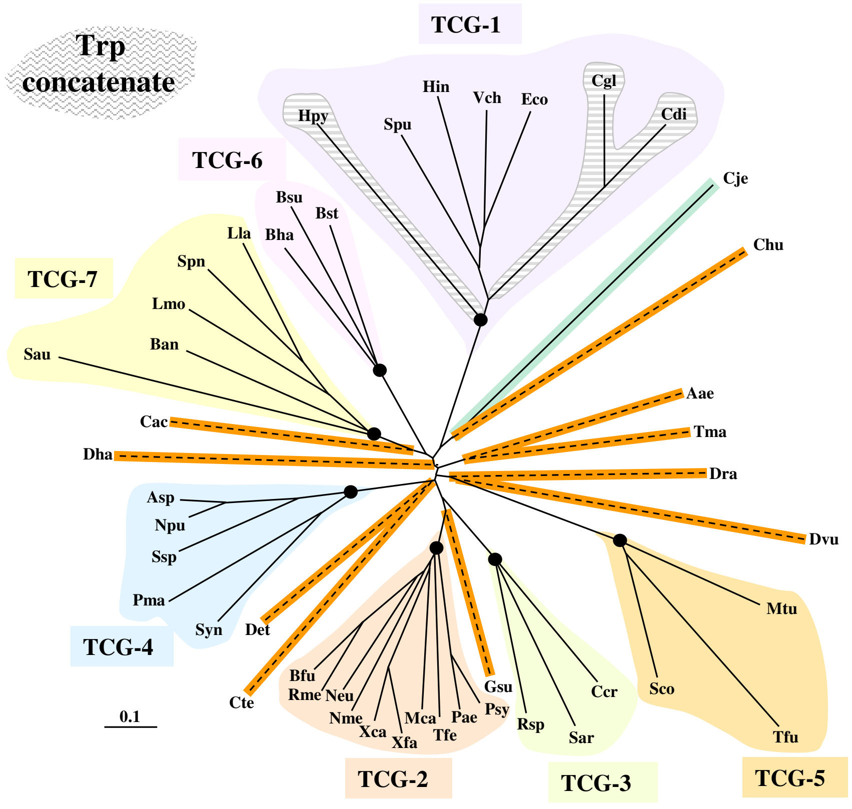

Figure 2

Phylogenetic tree (radial view) constructed using the seven-domain (TrpAa/Ab/B/C/D/Ea/Eb) concatenated sequences of Trp proteins that specifically participate in primary biosynthesis. The Trp-pathway concatenates are from the same 47 organisms shown in Fig. 1. Dashed, orange lines indicate the lineage positions of concatenated sequences from organisms that presently lack any close relatives whose genomes have been sequenced. TCG clusters are color-coded as in Fig. 1, where full bacterial names matching the corresponding abbreviations can be found. The Cje lineage, marked in aqua, represents a probable TCG grouping. Nodes marked with solid black circles are supported by bootstrap values of 100%. Concatenates of LGT origin within TCG-1 are outlined with a grey pattern.

looser TCG-7 assemblage. The emerging, fuller membership of these seven TCGs is hinted at in Table 3, where additional provisional TCG members that were not included in Figure 2 are listed in regular non-bold type. The latter were assigned (mostly with newly available genomes after completion of our primary analysis) by assessment of the best BLAST (Basic Local Alignment Search Tool) hits obtained after entering each Trp domain of the organisms listed as query sequences. Best-match methods are useful for rapid screening, but are subject to limitations [24]. Tentative TCG assignments were also assisted by other features. For example, *Brucella melitensis*, *Agrobacterium tumefaciens*, *Bradyrhizobium japonicum*, *Sinorhizobium meliloti* and *Rhodopseudomonas palustris* all possess in common the following split-pathway gene organization: *trpAatrpAb* (fused genes), a *trpB*/*trpD* operon, and a *trpC*/*trpEb*/*trpEa* operon [6].

The few exceptions that violate the otherwise good congruency between TCGs and 16S rRNA subtree regions point strongly to instances of whole-pathway *trp*-operon displacement. These include the incongruent presence of Trp-protein concatenates from both *Helicobacter pylori* and the coryneform bacteria in TCG-1. This indicates that inter-genomic transfer of whole-pathway *trp* operons occurred, with the donor identifiable as a member of TCG-1.

Orange-highlighted dashed lines in Figures 1 and 2 mark lineages where closely related genomes have not yet been sequenced, that is, where there is sparse genome representation. There is no congruency of the branching positions of these ten orphan organisms when the 16S rRNA tree (Figure 1) is compared with the Trp-protein concatenate tree (Figure 2). It is generally believed that the position of these lineages on the 16S rRNA tree is reliable because 16S rRNA comparisons can discriminate very well over wide phylogenetic distances. In contrast, support for the position of branching obtained for these lineages with respect to the concatenated Trp-protein tree is not significant, that is, low bootstrap values. One can reasonably expect that, as more genomes are sequenced, new TCGs will emerge that include these current orphan organisms. For example, our preliminary data indicate that the Trp proteins of *Bacteroides thetaiotaomicron* will reside within a new TCG group with *Cytophaga hutchinsonii*.

Individual Trp-protein trees

In the following section, individual protein trees are shown for comparison with the much superior tree of Trp-protein concatenates shown in Figure 2. Each individual tree consists of 47 sequences that comprise segments of each concatenate string used for the tree shown in Figure 2. Included among the concatenates are proteins of LGT origin (from *Helicobacter pylori*, *Corynebacterium glutamicum*, and *Corynebacterium diptheriae*) that function for primary Trp biosynthesis. In addition to the foregoing sequences, Figure 3, Figures 5,6,7, and Figures 9,10,11 also display paralogs and xenologs that were excluded from the concatenate strings because they were not deemed to function for primary biosynthesis. It can be seen that TrpAb (Figure 5) and TrpC (Figure 7), being relatively short and not highly conserved sequences, are the least informative. TCG-2 and TCG-7 are frequently not visualized as entirely cohesive groupings when the individual trees are inspected. In each individual Trp-protein tree, primary-pathway Trp domains are designated by the organism acronym (e.g. Det), whereas unknown xenologs or paralogs are designated with a following number (e.g. Det_2). If functions are known or if names exist in the literature, for example, Sco_CDA or Pae_PhnA, respectively, these designations are used. Remnant proteins are denoted with an 'r', for example, Cgl_r (Figure 9).

Xie *et al.* [6] can be consulted for a detailed overview of the Trp pathway reactions (summarized here in Table 1) and for a perspective on the nomenclature issues. Table 4 is a comprehensive listing [see Additional File 1] that contains gi (gene identification) numbers (in bold) that correspond to each of the seven domains asserted to function in primary Trp biosynthesis. Paralog and xenolog gi numbers are presented in regular type. Each gi number is hyperlinked to the corresponding record at NCBI (National Center for Biotechnology Information).

TrpAa

TrpAa is the aminase subunit of anthranilate synthase. Figure 3 shows an individual protein tree for the TrpAa catalytic domain. Forty-seven of these are domain segments of the 47 concatenates making up the tree of Figure 2. An additional three sequences are putative specialized paralogs or xenologs that have been excluded from Figure 2. These include an apparent TrpAa xenolog from *X. fastidiosa* (Xfa_2) [7]. The origin of Xfa_2 TrpAa is unknown. It is part of an apparent six-gene operon that is divergently oriented next to *trpR* [7]. It is perhaps suggestive of epsilon proteobacteria origin that Xfa_2 TrpAa exhibits a 100% bootstrap score with the TrpAa of *Campylobacter*, although the branch distances are great. Except for Xfa TrpAb (which also exhibits a suggestively close relationship with *Campylobacter* TrpAb; see Figure 5), the *Campylobacter* genome possesses neither the other four genes of the *Xylella* operon, nor *trpR*.

PhnA from *P. aeruginosa* [25], together with PhnB (see Figure 5), encode a partial-pathway operon that originated from the enteric lineage. A TrpAa paralog (Sco_CDA) in *S. coelicolor* functions in the calcium-dependent antibiotic (CDA) pathway [19].

Figure 3 does not show the position of the TrpAa domain from *Thermomonospora fusca*, which possesses a *trpAa•trpAb* gene fusion. Other actinomycetes lack this fusion. The fusion is present in some widely separated organisms (Figure 4) that also use it for primary Trp biosynthesis [7]. These include *Legionella pneumophila* (Lpn), *Azospirillum brasilense*, and the closely related group of organisms: *Rhodopseudomonas palustris*, *Mesorhizobium loti* (Mlo), *Sinorhizobium meliloti* (Sme), and *Brucella melitensis* (Bme). Although *Nostoc punctiforme* (Npu) and *Anabaena* sp. (Asp and Asp_2) also have *trpAa•trpAb* fusions, these are deemed not to be engaged in primary Trp biosynthesis [7]. These TrpAa• domains all cluster closely together to the exclusion of other TrpAa sequences. They do not fall into any of the TCGs that would be expected according to the phylogenetic position of the organism having the fusion. Figure 4 shows a broader phylogenetic sampling of TrpAa sequences than shown in Figure 3. Figure 4 illustrates the distinct cohesion of TrpAa• domains of

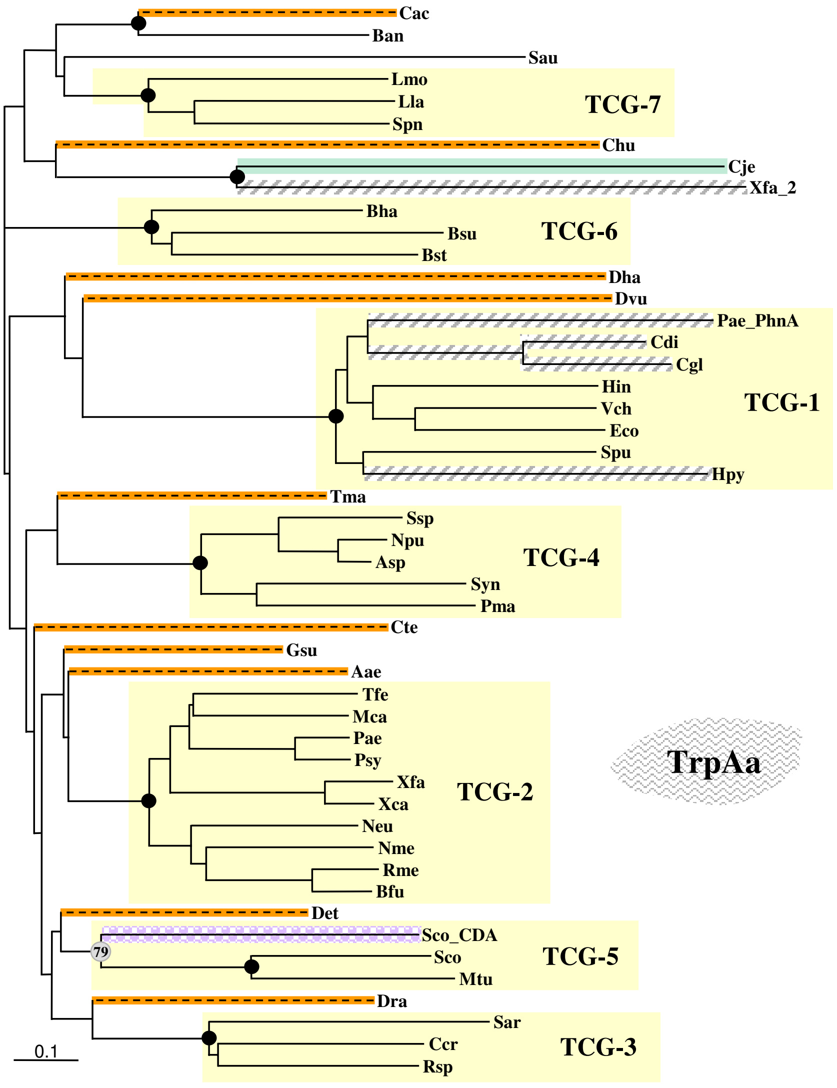

Figure 3

Phylogenetic tree (phylogram view) of TrpAa sequences. Nodes occupied by a solid, black circle are supported by bootstrap values of 100%. Other bootstrap values are given within unfilled circles. Xenologs are shown with grey candy-striped bars, and paralogs are shown with lavender/white patterning. A specialized-pathway paralog from *Streptomyces coelicolor* (Sco_CDA) is shown, as well as a probable specialized-pathway xenolog (Xfa_2) from *Xylella* [7]. Genes in *C. diptheriae*, *C. glutamicum*, and *H. pylori* that belong to whole-pathway operons originating via LGT are shown within TCG-1. TCG-1 also includes PhnA, encoded by a gene of a partial-pathway operon from *P. aeruginosa*. TCG-7 is fragmented, the Ban and Sau TrpAa sequences being too divergent to fall within TCG-7.

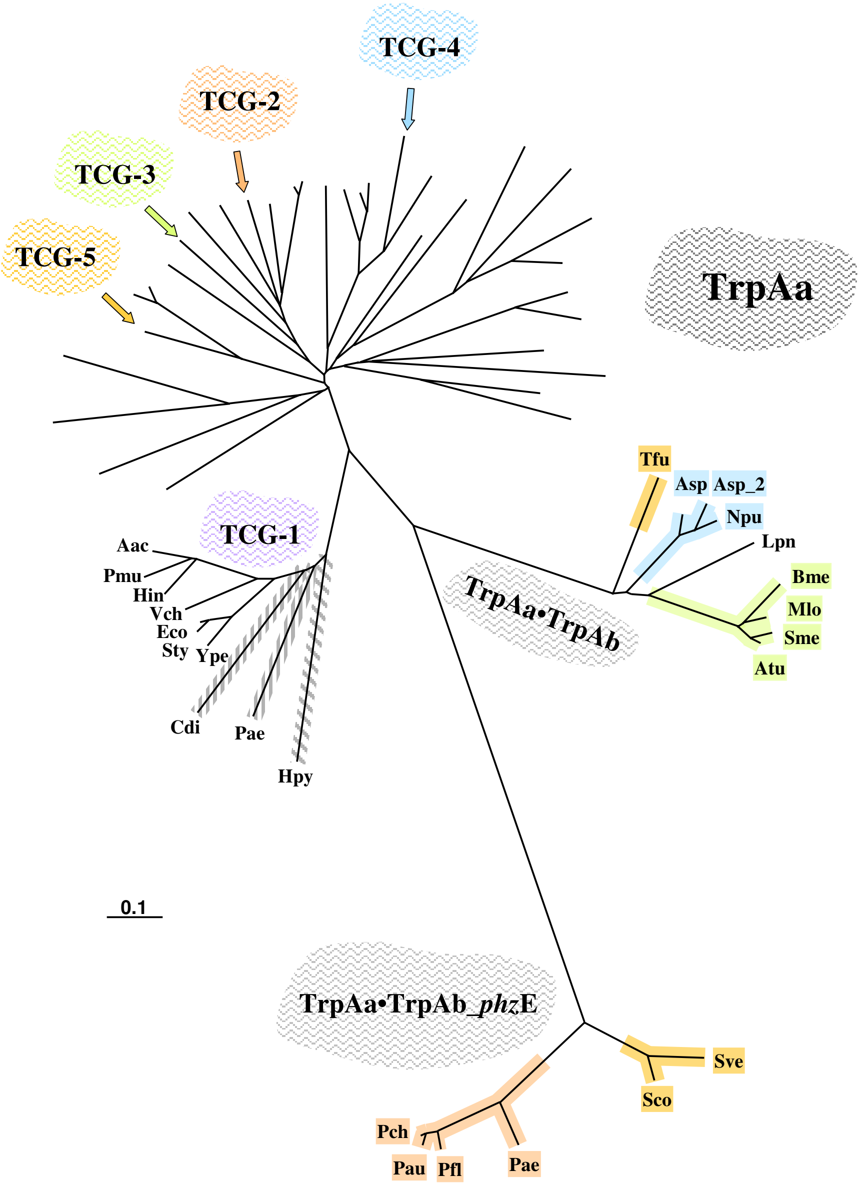

Figure 4
Phylogenetic tree (radial view) showing that free-standing TrpAa domains, TrpAa components of TrpAa•TrpAb fusions, and TrpAa components of TrpAa•TrpAb_phz fusions are all distinct from one another. The position of some TCG groups are marked at the top. TrpA• fusion domains are color-coded to indicate their expected TCG placements if convergent evolution were not a factor. The TCG-1 grouping is distinctly divergent from all of the other free-standing TrpAa sequences.

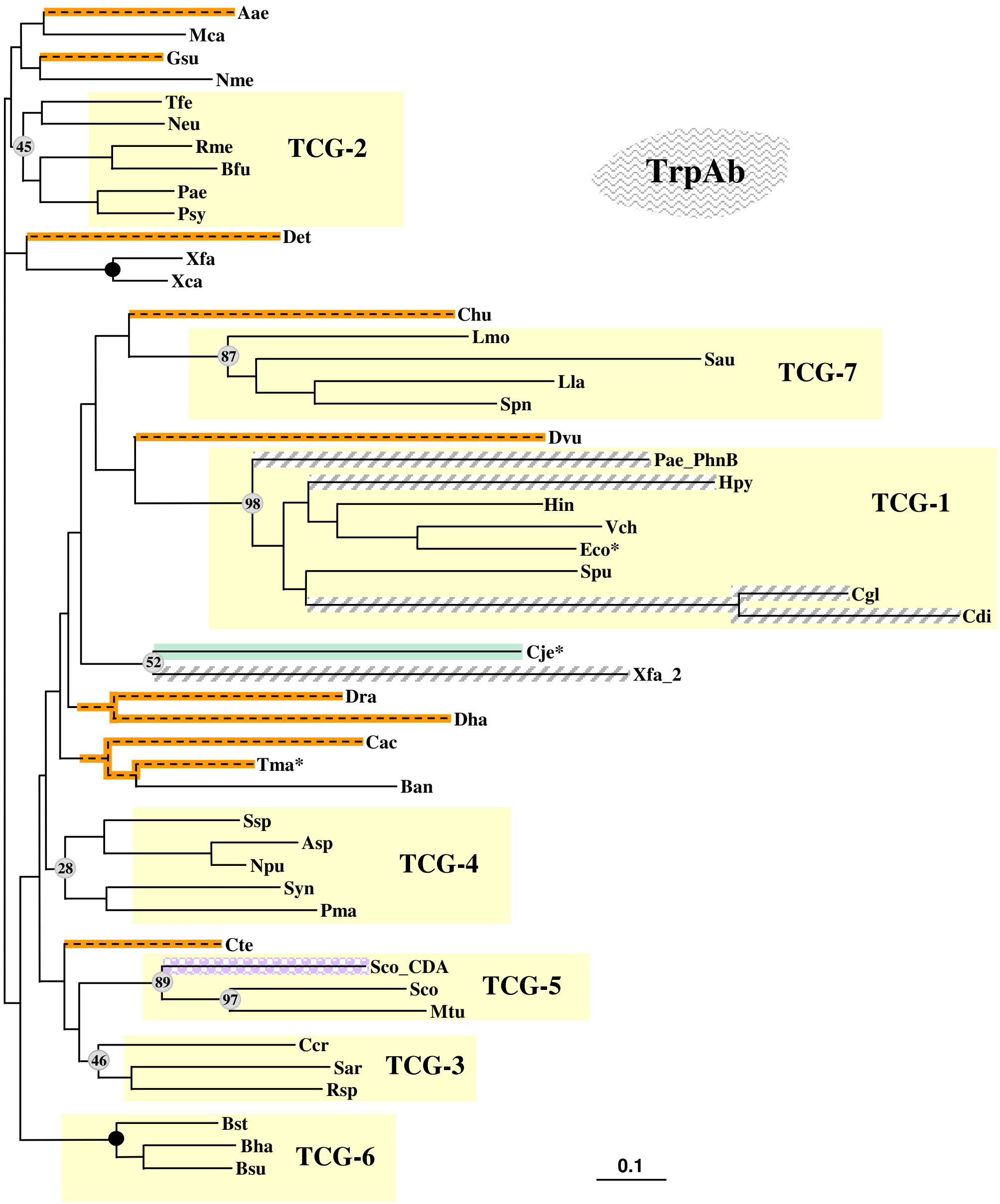

Figure 5

Phylogenetic tree of TrpAb sequences. Xfa_2 is a probable specialized-pathway xenolog from *Xylella*. Sco_CDA is a specialized-pathway paralog within TCG-5. TCG-1 contains a xenolog (PhnB) from *Pseudomonas aeruginosa* that (together with PhnA) is encoded by genes of a partial-pathway operon. Trp Ab genes from *Helicobacter pylori*, *Corynebacterium diptheriae*, and *C. glutamicum* are xenolog members of whole-pathway operons. TCG-2 and TCG-7 are fragmented. (Consult Fig. 2 for intact TCGs.) Asterisks (Tma, Cje, and Eco) indicate domains that exist as part of TrpAb•TrpB fusions.

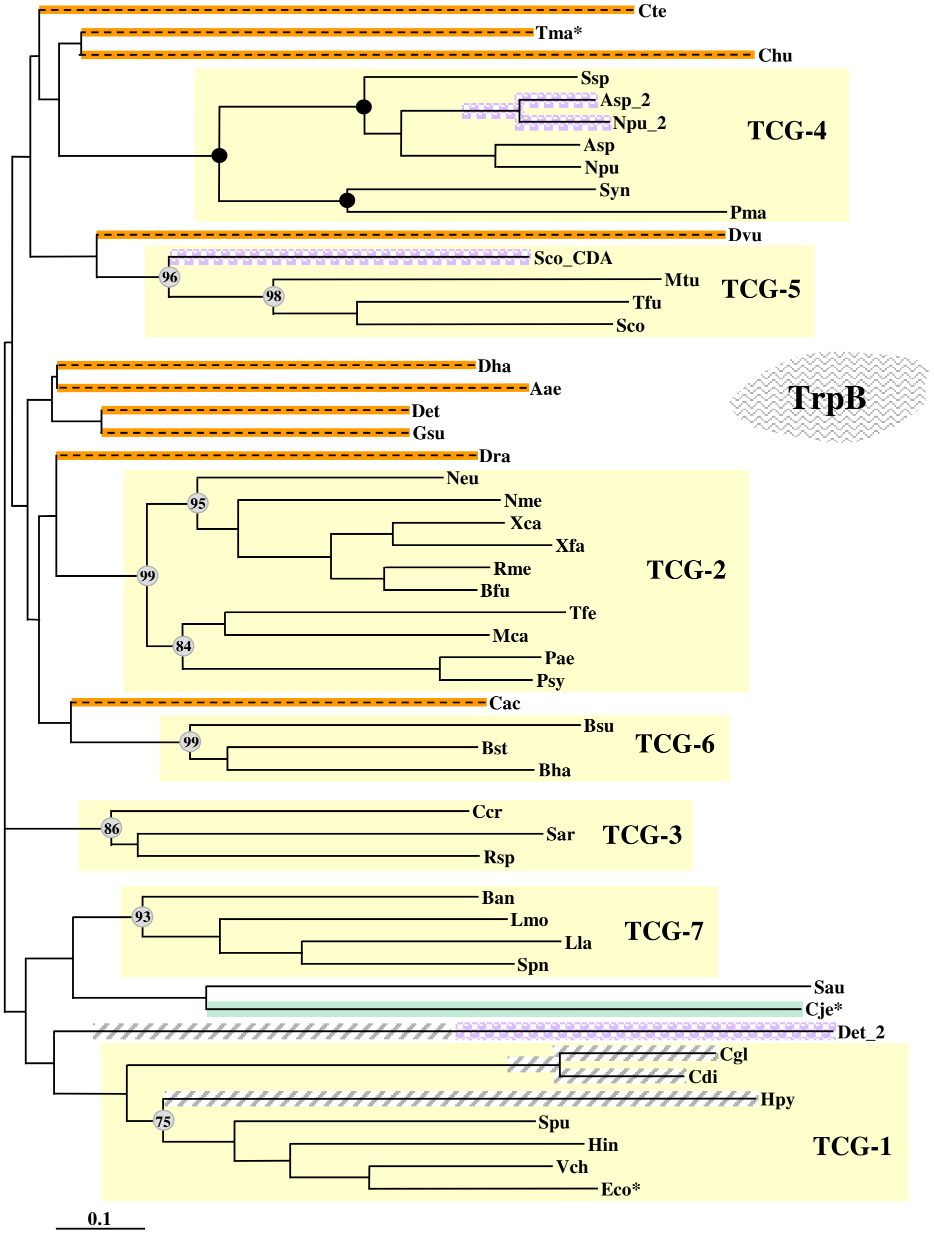

Figure 6

Phylogenetic tree of TrpB sequences. Det_2 is diagrammed with both patterns, indicating it could be either an ancient paralog or a xenolog. Sco_CDA is a specialized-pathway paralog (antibiotic). Asp, Asp_2, Npu, and Npu_2 are paralogs from a gene duplication that preceded speciation of Npu and Asp. The Asp and Npu sequences were arbitrarily used for input into the concatenate tree of Fig. 2. TCG-7 is fragmented. TCG-1 contains xenolog members of whole-pathway operons from *H. pylori*, *C. diptheriae*, and *C. glutamicum*. Asterisks (Tma, Cje, and Eco) indicate domains existing as part of TrpAb•TrpB fusions.

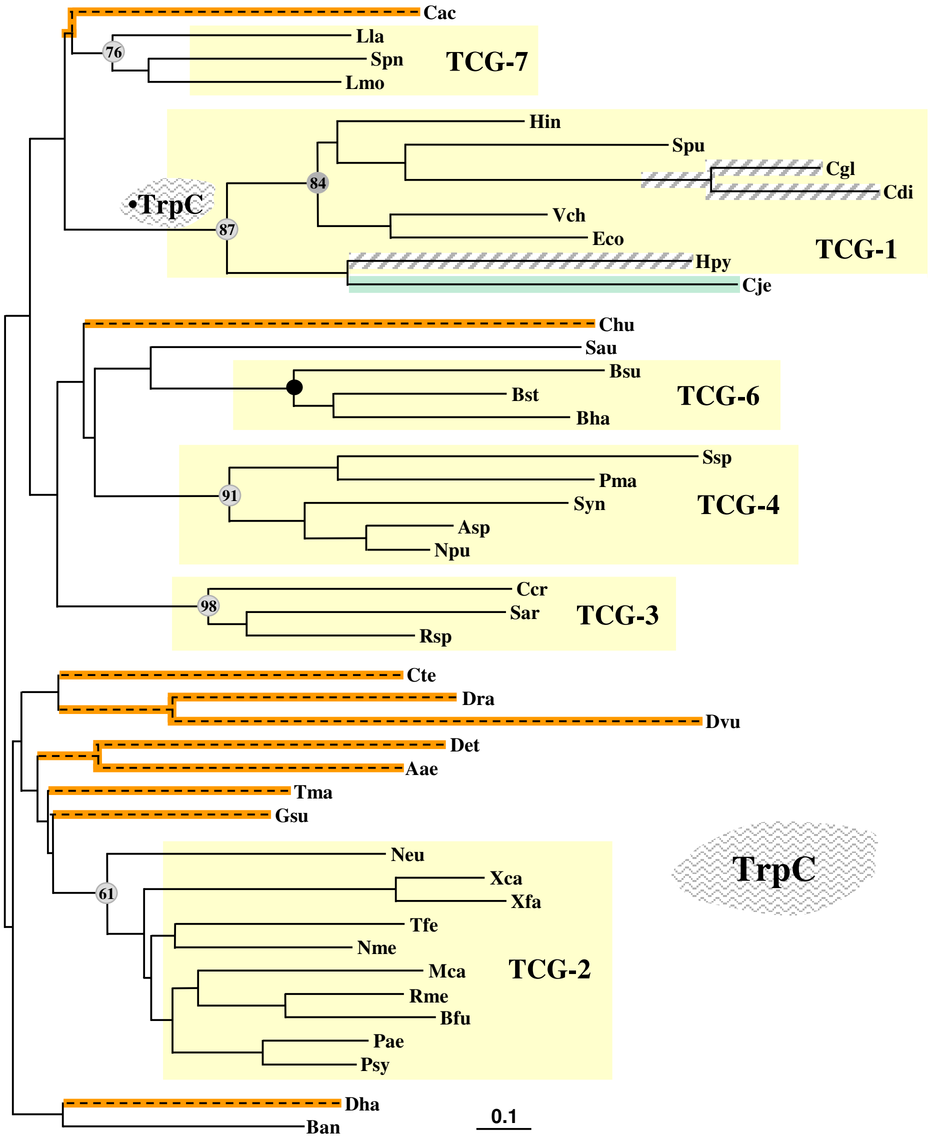

Figure 7

Phylogenetic tree of TrpC sequences. TCG-1 contains xenolog members of whole-pathway operons from *H. pylori*, *C. glutamicum* and *C. diptheriae*. TCG-7 is fragmented. All members of TCG-1 shown (except Cje, whose position is probably coincidental) possess •TrpC as a fusion domain with TrpD•.

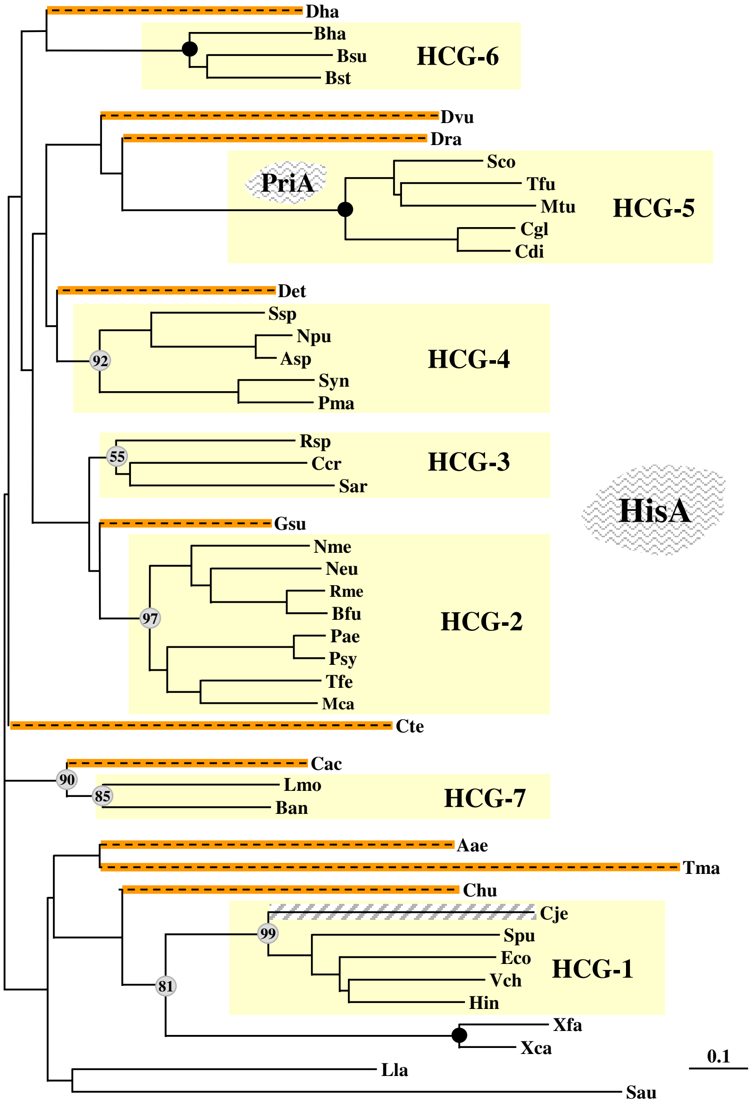

Figure 8

Phylogenetic tree of HisA sequences. Congruency groupings that match TCG clusters are labelled 'HCG' for histidine congruency group. *Helicobacter pylori* and *Streptococcus pneumoniae* are not represented on the tree because they have lost the histidine pathway. *C. jejuni* HisA appears to be a xenologous member of TCG-1. Similar to the relatively loose TCG-2 and TCG-7 groupings, HCG-2 and HCG-8 are fragmented. Actinomycete bacteria possess a tightly clustered section of dual-pathway HisA (PriA) sequences within HCG-5.

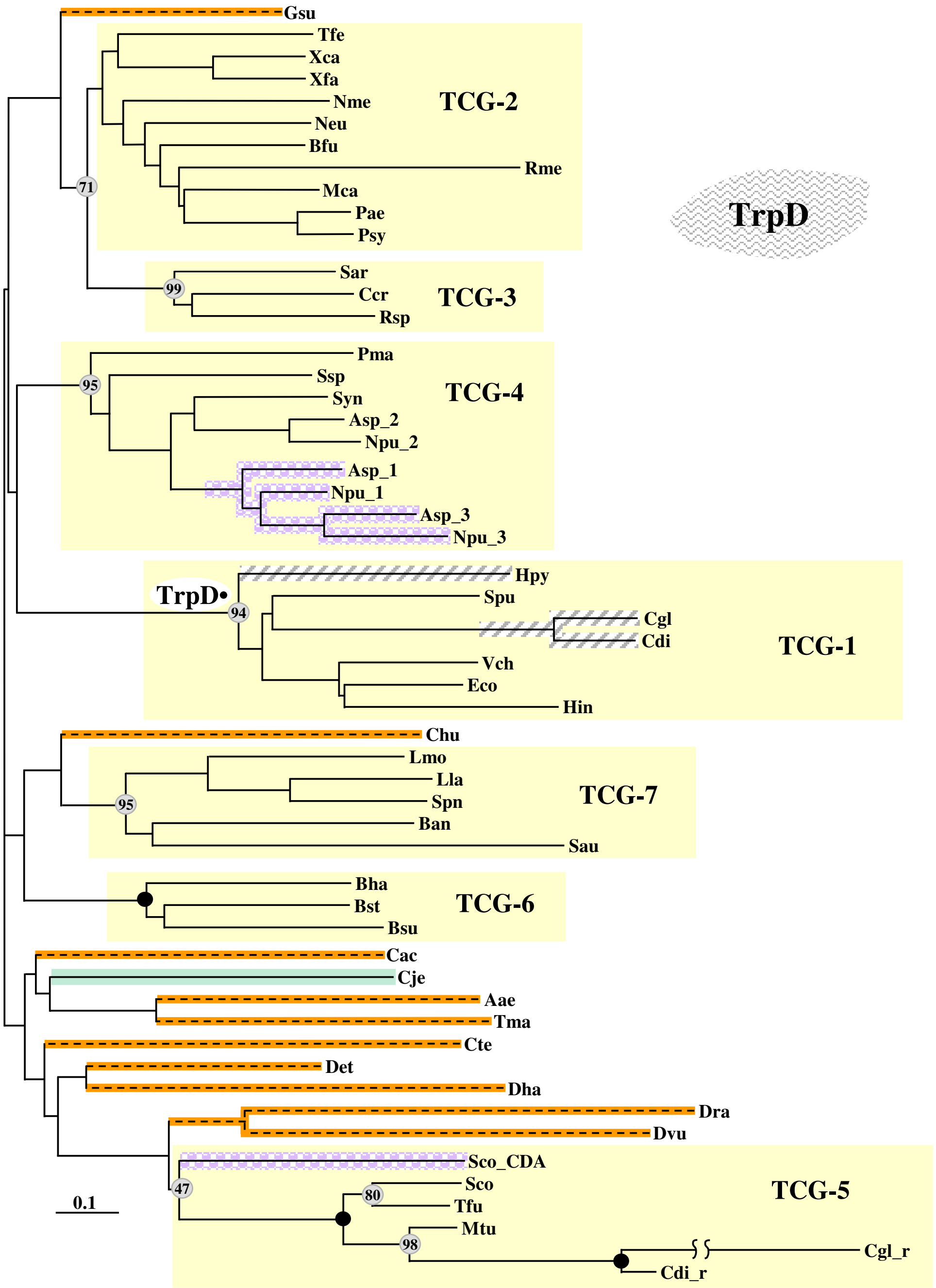

Figure 9

Phylogenetic tree of TrpD sequences. *N. punctiforme* and *Anabaena* sp. each possess a set of three paralogs. The Asp_2 and Npu_2 paralog sequences were used for input into the concatenates of Fig. 2. In addition to the xenolog TrpD• domain of Cgl and Cdi that is present in TCG-1, Cgl and Cdi also possess 'remnant' TrpD proteins denoted Cgl_r and Cdi_r that cluster in TCG-5.

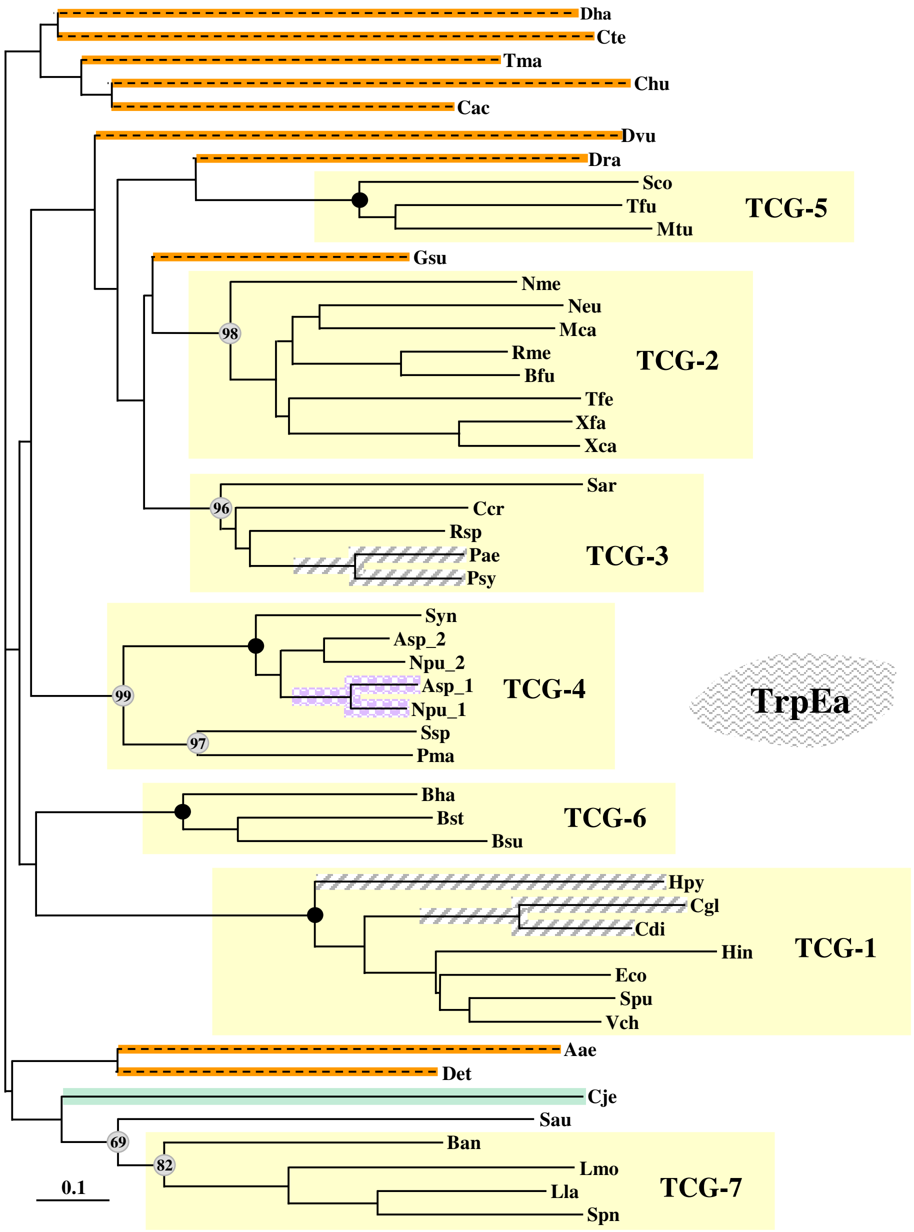

Figure 10
Phylogenetic tree of TrpEa sequences. TrpEa proteins from Pae and Psy fall into TCG-3 instead of into TCG-2. Paralogs Asp_1, Npu_1, Asp_2 and Npu_2 were generated by a gene duplication that preceded speciation of Asp and Npu. Asp_2 and Npu_2 were arbitrarily used for concatenate input (Fig. 2). TCG-7 is not very cohesive.

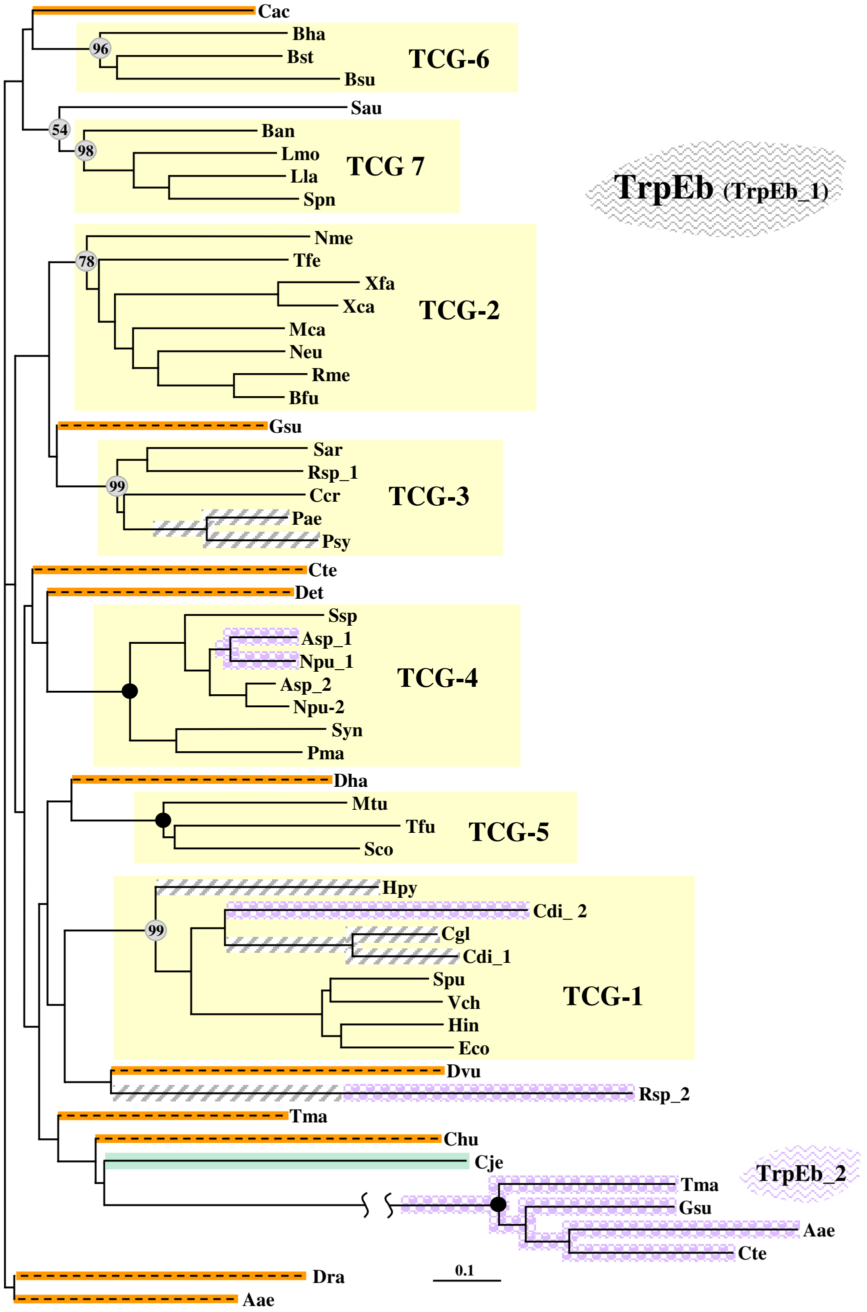

Figure 11
Phylogenetic tree of TrpEb sequences. Cdi_2 is encoded by a paralog copy of *trpEb_1* that has been inserted ahead of the Cdi *trp* operon. TrpEb_2 is encoded by a highly divergent paralog subclass of *trpEb* that probably has a specialized function [28]. No TrpEb_2 sequences were used to construct concatenates (Fig. 2). Rsp_2 is shown with two patterns, indicating uncertainty about whether it is an ancient paralog or a xenolog.

Table 3: Membership composition of Tryptophan Congruency Groups[a]

TCG-1	***[Corynebacterium diptheriae]***[b]
	[Corynebacterium glutamicum][b]
	Escherichia coli
	Haemophilus influenzae
	[Helicobacter pylori][b]
	Shewanella putrefaciens
	Vibrio cholerae
	Actinobacillus actinomycetemcomitans
	Buchnera aphidicola[c]
	Blochmannia floridanus[c]
	[Corynebacterium efficiens][b]
	Erwinia carotovora
	Klebsiella pneumoniae
	Pasteurella multocida
	Photorhabdus luminescens
	Salmonella enterica
	Shigella flexneri
	Vibrio parahaemolyticus
	Vibrio vulnificus
	Yersinia pestis
TCG-2[d]	***Burkholderia fungorum***
	Methylococcus capsulatus
	Neisseria meningitidis
	Nitrosomonas europaea
	Pseudomonas aeruginosa*
	Pseudomonas syringae*
	Ralstonia metallidurans
	Thiobacillus ferrooxidans
	Xanthomonas campestris
	Xylella fastidiosa
	Acinetobacter sp.
	*Azotobacter vinelandii**
	Bordetella bronchisepticum
	Bordetella parapertussis
	Bordetella pertussis
	Burkholderia cepacia
	Burkholderia multivorans
	Chromobacterium violaceum
	Microbulbifer degradans
	Neisseria gonorrhoeae
	*Pseudomonas fluorescens**
	*Pseudomonas putida**
	Psychrobacter sp.
	Ralstonia solanacearum
	Xanthomonas axonopodis
TCG-3	***Caulobacter crescentus***
	Rhodobacter sphaeroides
	Sphingomonas aromaticivorans
	Agrobacterium tumefaciens
	Bradyrhizobium japonicum
	Brucella melitensis
	Brucella suis
	Rhizobium loti
	Rhodopseudomonas palustris
	Sinorhizobium meliloti
TCG-4	***Anabaena*** **(Nostoc) sp. PCC 7120**
	Nostoc punctiforme
	Prochlorococcus marinus **CCMP 1986 (MED4)**
	Synechococcus **sp. WH8102**
	Synechocystis **sp. PCC 6803**
	Anabaena variabilis ATCC 29413
	Crocosphaera watsonii WH 8501

Table 3: Membership composition of Tryptophan Congruency Groups[a] *(Continued)*

	Gloeobacter violaceus PCC 2471
	Prochlorococcus marinus CCMP1375 (SS120)
	Prochlorococcus marinus MIT9313
	Synechococcus elongatus PCC 7942
	Thermosynechococcus elongatus BP-1
	Tricodesmium erythraeum
TCG-5[e]	***Streptomyces coelicolor***
	Mycobacterium tuberculosis
	Thermomonospora fusca
	Bifidobacterium longum
	Mycobacterium avium
	Mycobacterium bovis
	Mycobacterium leprae
	Mycobacterium smegmatis
	Streptomyces avermitilis
TCG-6	***Bacillus halodurans***
	Bacillus stearothermophilus
	Bacillus subtilis
	Oceanobacillus iheyensis
TCG-7	***Bacillus anthracis***
	Lactococcus lactis
	Listeria monocytogenes
	Staphylococcus aureus
	Streptococcus pneumoniae
	Bacillus cereus
	Listeria innocua
	Staphylococcus epidermidis
	Streptococcus gordonii
	Streptococcus mutans

[a] Each tryptophan congruency group (TCG) defined by the concatenated tree for Trp proteins (Fig. 2) is congruent with the color-coded subtree section within the 16S rRNA tree (Fig. 1). Organisms that are included in the concatenated tree of Fig. 2 are indicated in boldface type, whereas additional organisms not included in the concatenated tree but that were qualitatively determined to belong to a given TCG are indicated in regular type. [b] TCG members originating by LGT are indicated within brackets and indented. [c] Insect symbionts. [d] The five organisms marked with asterisks form a distinctive subclade that is, in fact, not a "pure" component of TCG-2 because of the LGT origins of *trpEa* and *trpEb* from TCG-3 (see text). [e] All members of TCG-5 lack *trpC* and presumably utilize a dual-pathway *hisA* (*priA*) for this function.

TrpAa•TrpAb fusions, as well as tight clustering of another variant of TrpAa•TrpAb fusion denoted TrpAa•TrpAb_phz. The latter group has a deleted region, which probably contributes to the dead-end production of 2-amino-2-deoxy-isochorismate (normally an enzyme-bound intermediate in the anthranilate synthase reaction). This compound is a precursor of phenazine pigments [26], hence our acronym TrpAa•TrpAb_phz. Figure 4 shows in the unlabelled (i.e. no organism acronyms) upper group a variety of TrpAa proteins from *Bacteria*, *Archaea*, and one lower eukaryote. It is qualitatively apparent that TrpAa• in a given fusion protein is distinctly separated from TrpAa proteins of relatively close relatives lacking the fusion (as indicated by the color coding in Figure 4). For example, unfused TrpAa proteins from *S. coelicolor* or *Mycobacterium tuberculosis*, very similar to one another as cohesive members of TCG-5, are not even as close to the TrpAa• domain from the fellow actinomycete, *Thermomonospora fusca*, as they are to various archaeal TrpAa proteins or to that from *Saccharomyces cerevisiae*. The •TrpAb domain of TrpAa•TrpAb or TrpAa•TrpAb_phzE fusion proteins showed similar tight clustering when comprehensive TrpAb trees were constructed (data not shown). We have suggested that a number of TrpAa•TrpAb fusions have occurred as independent fusion events, but that the sequences have then converged due to rigid constraints imposed for proper protein-protein interactions of these subunits [7].

TrpAb

TrpAb is the glutamine-binding subunit of anthranilate synthase responsible for function of the TrpAa/TrpAb complex as an amidotransferase. It is the smallest and least conserved of the seven Trp domains. Sequence features of TrpAb do not distinguish it from from PabAb, an amidotransferase subunit of *p*-aminobenzoate (PABA) synthase. Usually, however, the functional role of the two members of this homolog group can be deduced by the presence of at least one homolog in a *trp* operon or a *pab* operon. In some cases, as with *B. subtilis* and other members of TCG-6, a single subunit participates in both anthranilate and PABA synthesis. In this case it has dual-pathway function and has the ability to complex with TrpAa, on the one hand, or with PabAa, on the other hand. Most of the TCGs seen in Figure 2 are recognizable on the TrpAb tree in faint outline (supported with only marginal bootstrap values), with TCG-2 and TCG-7 being especially fragmented (Figure 5). The position of the •TrpAb domain from *T. fusca* is not shown. (The discussion given immediately above for TrpAa• from *T. fusca* applies here.) Note that TrpAb from *Thermotoga*, *Campylobacter*, and a small subclade of TCG-1 represented in Figure 5 by *E. coli* are fused with TrpB (TrpAb•TrpB). These seem to have arisen as independent fusions, and there is no indication here of convergent evolution (in contrast to the TrpAa•TrpAb fusions discussed above).

PhnB from *P. aeruginosa* is encoded by a gene within an operon that also contains *phnA* (Figure 3). The *phnA* and *phnB* xenologs in *P. aeruginosa* are both divergent from the *P. aeruginosa trpAa* and *trpAb* homologs that are engaged in primary biosynthesis. The *phnA/phnB* operon originated by LGT from within TCG-1.

The gene encoding the TrpAb paralog (Sco_CDA) is a member of a large operon (along with TrpAa, TrpB, and TrpD paralogs) that is engaged in "calcium-dependent antibiotic" biosynthesis [19] and presumably arose intragenomically by gene duplication.

TrpB

TrpB catalyzes the anthranilate phosphoribosyl transferase step in which anthranilate and PRPP combine to yield phosphoribosyl-anthranilate. In this paper the usage of "TrpB" refers to TrpB_1, the major and ubiquitous subtype species. The distinctly divergent TrpB_2 subtype is narrowly distributed (being cyanobacteria-specific) and usually co-exists with TrpB_1 in cyanobacteria (Table 4 [see Additional File 1]). TrpB_2 might prove to be an ancient paralog in cyanobacteria that has diverged to perform some other phosphoribosyl-transfer function. However, at present the closest homolog of TrpB_2 proteins are the TrpB_1 enzymes that catalyse the second step of Trp biosynthesis.

Figure 6 includes another *S. coelicolor* paralog dedicated to antibiotic synthesis (Sco_CDA) that can be distinguished from the primary-biosynthesis paralog. *Dehalococcoides ethanogenes* has two highly divergent paralogs, but Det_2 can be reasonably excluded from the primary-biosynthesis pathway (and, therefore, from the Det concatenated sequence of Figure 2) because it alone is not located within the whole-pathway operon *trpAaAbBDC*(*aroA*$_{I\beta}$)*EbEa* [6]. Det_2 could be a very ancient paralog, or it may have come from a LGT donor that happens to be relatively distant from any of the homologs included in Figure 6.

In the case of the unicellular cyanobacteria, a single set of completely dispersed *trp* genes exists in every genome. The *Nostoc/Anabaena* subclade is more complex, having additional operon-organized *trp* genes. Since these species also possess the set of dispersed *trp* genes, it has been concluded that the dispersed *trp* genes constitute a basic set of conserved genes whose functional role is primary Trp biosynthesis in all cyanobacteria [7]. Accordingly, the *trp* genes that are present in operon organization (including *trpB*) have been excluded from the Trp concatenates constructed for *Anabaena* sp. and *Nostoc punctiforme* (Figure 2).

As observed above (Figure 5) with the TrpAb• domain of three phylogentically separated TrpAb•TrpB fusions, the corresponding •TrpB domains (marked with asterisks in Figure 6) also show no indication of evolutionary convergence (suggesting that for this particular domain combination, fusion does not impose a set of rigid constraints that must be met in each independent fusion).

TrpC

TrpC catalyzes the phosphoribosyl-anthranilate isomerase step, which yields 1-(*o*-carboxyphenylamino)-1-deoxyribulose 5-phosphate (CDRP). Figure 7 shows the TrpC-protein tree on which the seven TCG groups are marked, as well as TrpC from the ten lineages (orange highlight) whose branching positions are not supported by significant bootstrap values. The •TrpC domain of *H. pylori* is the most divergent sequence within TCG-1. It clusters with TrpC from *Campylobacter jejuni* with a bootstrap value of 81%. All members of TCG-1 in the TrpC tree are •TrpC domains of TrpD•TrpC fusions, except for *C. jejuni* TrpC (whose location in TCG-1 is probably artifactual).

TCG-5 organisms (actinomycete bacteria) lack a *trpC* gene (except, of course, for the horizontally transferred genes of coryneform bacteria). *hisA* (*priA*) from these organisms presumably encodes a dual-pathway enzyme that performs the isomerase function in both the histidine and Trp biosynthetic pathways [13]. These dual-function *priA* genes are listed in the *hisA* column of Table 4 [see Additional File 1]. Since •*trpC* in the coryneform bacteria is coordinated with the other *trp*-operon genes that originated by LGT, their *hisA* (*priA*) gene can be considered a remnant in the context of Trp biosynthesis. However, an essential role in histidine biosynthesis has undoubtedly selected for maintenance of *hisA*. It would be interesting to know whether the *trpC* function of *hisA* in the coryneform species of bacteria has deteriorated or not.

HisA

HisA catalyses an Amadori rearrangement that is analogous to the TrpC rearrangement. This HisA isomerase catalyses the rearrangement of N'– [(5'-phospho**ribosyl**) forimino]-5-aminoimidazole-4-carboxamide ribonucleotide to N'– [(5'-phospho**ribulosyl**) forimino]-5-aminoimidazole-4-carboxamide ribonucleotide.

As noted above, the actinomycete group of HisA proteins is dually competent for both the TrpC and HisA isomerase functions, and they have been named PriA to distinguish them from the pathway-specific TrpC and HisA proteins [13]. However, the PriA proteins are clear homologs of HisA proteins, whereas homology with TrpC is uncertain. It is noteworthy that the HisA-protein tree shown in Figure 8 resembles the individual Trp-protein trees in terms of the congruency groupings that can be recognized, and these have been given parallel HCG (Histidine Congruency Group) denotations. HisA clusters are recognizable that parallel TCG-1, TCG-2, TCG-3, TCG-4, TCG-5, and TCG-6. HCG-2 exhibits some fragmentation in that it does not include the sequences from Xca and Xfa. Note that all of the actinomycete HisA (PriA) proteins, including those from *C. diptheriae*, *C. glutamicum* (and *Corynebacterium efficiens*, not shown), form a cohesive cluster. *H. pylori* has no histidine pathway. *C. jejuni* HisA falls within HCG-1. Indeed, *C. jejuni* appears to have acquired the entire histidine pathway from the enteric lineage (tentative observations based upon BLAST queries). Organisms belonging to TCG-7 do not exhibit cohesive clustering of HisA proteins on the HisA tree, but comparable fragmentation of TCG-7 is not unusual for many of the individual Trp-protein trees (i.e. especially with the least conserved proteins).

TrpD

TrpD catalyzes the reaction of indoleglycerol phosphate synthase, whereby CDRP is converted to indole 3-glycerol phosphate with the release of carbon dioxide (CO_2) and water (H_2O). The TCG groupings seen in Figure 2 are well articulated in the TrpD-protein tree of Figure 9. Within the cyanobacterial TCG-4, both *Anabaena* sp. and *N. punctiforme* possess multiple paralogs that are speculated [7] to have arisen by two gene duplications that occurred in a common ancester of *Anabaena* and *Nostoc*. The TrpD• domains of *H. pylori* and the coryneform bacteria cluster within TCG-1. *S. coelicolor* possesses a divergent paralog, Sco_CDA, that is dedicated to antibiotic biosynthesis.

All of the actinomycete bacteria (bottom of Figure 9) possess monofunctional TrpD proteins that cluster within TCG-5. The *trpD* genes from coryneform bacteria in TCG-5 are remnants that have been functionally displaced by the *trpD*• domain (located in TCG-1). The latter exists as part of the *trpD•trpC* fusion within the *trp* operon imported via LGT. The long branches of the TrpD sequences (Cgl_r and Cdi_r) from the coryneform bacteria (especially from *C. glutamicum*) suggest strongly that these remnants may have lost catalytic function.

TrpEa

TrpEa is the alpha subunit of tryptophan synthase. It converts indole 3-glycerol phosphate to indole with the release of glyceraldehyde 3-P. It forms a tight complex with TrpEb [27]. The TCG clusters are well defined, as shown in Figure 10. Within cyanobacterial TCG-4, *trpEa* appears to have been duplicated in a common ancestor of *Anabaena* and *Nostoc*, as discussed previously [7]. The close relatives, *P. aeruginosa* and *Pseudomonas syringae*, possess a single *trpEa* gene, which clearly falls into TCG-3, rather than TCG-2. Thus, the native *trpEa* in these *Pseudomonas* species must have been displaced by a gene originating within the lineage of TCG-3 organisms.

TrpEb (TrpEb_1)

TrpEb is the beta subunit of tryptophan synthase and condenses indole with *L*-serine to produce *L*-tryptophan. Indole is not a free intermediate and passes through a tunnel that is created within the TrpEa/TrpEb complex [27]. A small number of *Archaea* and *Bacteria* possess a homolog of TrpEb that falls into a distinct subcluster termed TrpEb_2 [28]. (In this manuscript the major enzyme species, TrpEb_1, may simply be referred to as TrpEb in any context where there would be no confusion.) It has been suggested that the TrpEb_2 species does not form a complex with TrpEa and might have some other stand-alone function, such as that of serine deaminase [28]. Figure 11 shows four bacterial genomes that possess *trpEb_*2 in addition to *trpEb_*1, namely *Thermotoga*, *Geobacter*, *Aquifex*, and *Chlorobium*. These organisms are widely spaced on the 16S rRNA tree (Figure 1). *Geobacter* has the partial-pathway operon *trpAa/trpAb/trpB/trpEb_2/ trpC*. *trpEa* and *trpEb_1* are located outside the operon and are unlinked to one another. Even though TrpEb_2 is located in the operon, it was excluded from the Gsu concatenate string (Figure 2) because *trpEb_*1 encodes all of the critical contact residues known to be important for physical association with TrpEa [28].

The TrpEb_1 sequences in Figure 11 form TCG clusters that are relatively well defined, judging from the excellent conformation with the tree of concatenated Trp sequences (Figure 2). The paralogs present in *Anabaena* and *Nostoc* within TCG-4 presumably originated following gene duplication in a common ancestor of these genera [7]. The sequences of TrpEb from *Thermomonospora*, *Streptomyces*, and *Mycobacteria* cluster together as expected for actinomycetes (TCG-5). Exceptions are the TrpEb proteins of the coryneform bacteria, which originated by LGT from a source within TCG-1. Interestingly, after LGT, gene duplication appears to have resulted in a second TrpEb species (Cdi_2) in *C. diptheriae*. The branching position shown in Figure 11 suggests that the duplication preceded speciation. If so, *C. glutamicum* has since lost the paralog. Alternatively, rapid divergence associated with probable status of Cdi_2 as a pseudogene might account for the position on the distance tree. It is interesting that it has recently been asserted that duplication occurs more often among laterally transferred genes than for native genes [29].

Aromatic Congruency Groups implicate two deteriorated TCGs

Chlamydophila/Chlamydia

The chlamydiae have an intact common pathway of aromatic biosynthesis, but none have a complete *trp* pathway. In one case, *Chlamydophila pneumoniae*, no *trp* genes are present. At the other extreme, *C. psittaci* possesses all *trp* genes except for *trpAa* and *trpAb*. Other species possess an intermediate number of *trp* genes. Various rationales have been advanced [10,11] in support of novel functions for these incomplete pathways that have some interesting implications for pathogenesis. Those few *trp* proteins remaining in the chlamydiae are always very tightly clustered with one another on individual protein trees. We expect that when sufficient related genomes are sequenced, Trp congruency will be established for the chlamydiae. We mainly base this on the fact that trees corresponding to the seven enzymes of the common aromatic pathway yield a very cohesive Aromatic Congruency Group (Xie and Jensen, in progress).

Epsilon proteobacteria

The epsilon proteobacteria, *Helicobacter pylori* and *C. jejuni*, were among the early complete and published genomes. Recently, the genomes of *H. hepaticus* and *Wolinella succinogenes* have also become available. These four organisms all possess common-aromatic pathway enzymes whose sequences cluster tightly together (Xie and Jensen, in progress). Since the *H. pylori trp* genes had been replaced via LGT after divergence from their closest neighbor (*H. hepaticus*), we expected that the Trp proteins of the three remaining epsilon proteobacteria would define a TCG grouping. However, Trp proteins from *H. hepaticus*, *W. succinogenes* and *C. jejuni* are all divergent from one another. They are also divergent from sequences of other organisms. There is, so far, no indication of an LGT history. *C. jejuni* has a *trpAa/trpAb•trpB/trpC/trpEb/trpEa* operon (*trpD* is isolated in an unlinked position). *H. hepaticus* has spaced the *trp* genes in three locations, where *trpAa/trpAb•trpB*, *trpD•trpC* and *trpEb/trpEa* reside. It is anomalous and quite possibly symptomatic of ongoing reductive evolution that *trpEa* and *trpEb* are divergently oriented. *H. hepaticus* not only has the *trpAb•trpB* fusion, but also has a *trpD•trpC* fusion. The *trpD•trpC* fusion of *H. hepaticus* appears to be of independent origin with respect to all of the others in TCG-1, based upon phylogenetic analysis and inter-domain linker analysis. In *W. succinogenes* the *trp* genes are all dispersed and none of them are fused. In the future, a core of epsilon proteobacteria may become available as a source of sequences to build a Tryptophan Congruency Group. The organisms currently available have experienced dynamic (and perhaps disruptive) evolutionary events that currently prevent definitive conclusions about the common ancestor.

For comparison, we took a preliminary look at the histidine pathway in these four epsilon proteobacteria. *H. pylori* has completely lost the histidine pathway, and *C. jejuni* appears to have displaced its pathway with the histidine operon from the enteric lineage. Genes of histidine biosynthesis are completely dispersed in *H. hepaticus* and *W. succinogenes*, and all yield mutually best BLAST hits. Thus, *H. hepaticus* and *W. succinogenes* may be core genomes that represent a congruency group with respect to histidine biosynthesis.

Some underlying complexities of Trp-protein concatenates

Convergence of TrpAa•TrpAb fusions

Fusions of general aromatic biosynthetic pathway genes, including those of the Trp pathway, have occurred frequently. It is not uncommon for the same genes in different organisms to have undergone fusion independently. In cases where the catalytic domains function separately, there may be great latitude for successful fusion orientations, for example, *aroQ•pheA* fusions [30]. In contrast, TrpAb delivers ammonia to the active site of TrpAa in a way that may impose particular demands upon the spatial orientation of the protein-protein interaction operating for the anthranilate synthase complex. Although we suggest that TrpAa•TrpAb fusions have occurred independently at least five times in widely spaced lineages, they all group tightly together in both TrpAa trees (Figure 4) and TrpAb trees (not shown). We believe that this is due to strong selective pressure for convergence. Thus, the individual TrpAa and TrpAb trees place the *T. fusca* TrpAa and TrpAb proteins in an anomalous group that includes proteins from *Legionella*, *Azospirillum*, *Brucella* (and close relatives) and the *Nostoc/Anabaena* group of cyanobacteria. This phenomenon of convergence disrupts the congruence of the individual TrpAa tree when these organisms are included, but this is not enough to undermine the proper placement of the *T. fusca* Trp-protein concatenate within TCG-5.

The bifunctional priA *gene of TCG-5 actinomycetes*

The enzyme encoded by *priA* is competent for the isomerase step in both the histidine and Trp pathways of actinomycete bacteria [13]. Thus, PriA sequences would be concatenate components defining not only TCG-5, but also a Histidine Congruency Group (HCG-5) (Figure 8). Presumably the acquisition of the *trp* operon in coryneform species leaves *priA* unnecessary for a role in Trp biosynthesis. The dual substrate capabilities of PriA might reflect an ancient state of broad-specificity competence, with *hisA* and *trpC* being derived from specialized offspring of gene duplication [13]. On the other hand, because PriA is clearly a homolog of HisA, but an uncertain homolog of TrpC, it is possible that an ancestral *trpC* was lost and subsequently replaced by a divergent duplicate of *hisA*. The overall effect of PriA being uniquely present in Trp-protein concatenates of actinomycete bacteria (i.e. being forced to align with TrpC in all other concatenates) is to compress slightly the actinomycete TCG grouping.

The dual-pathway pabAb *of TCG-6 bacilli*

Very much like the above situation, TCG-6 *Bacillus* species have lost *trpAb* from an otherwise complete *trp* operon. *pabAb*, located in the *pab* operon, is competent to function for both Trp and folate biosynthesis (see Xie *et al.* [6]). Thus, within the narrow *Bacillus* clade that includes *B. subtilis*, *B. halodurans* and *B. stearothermophilus*, PabAb sequences help to define not only TCG-6, but also a pending Folate Congruency Group.

Discussion

Evolutionary scenario for LGT of whole-pathway Trp operons

Figure 12 depicts the evolutionary events proposed to account for the location in both *H. pylori* and coryneform bacteria of all seven domains that are associated with primary Trp biosynthesis within TCG-1, which otherwise is populated by members of the enteric lineage. Members of TCG-1 all possess whole-pathway operons with a *trpD•trpC* fusion. As an isolated observation, we cannot absolutely rule out independent fusions of *trpD* with *trpC* in *Helicobacter* and coryneform bacteria, followed by convergent evolution with respect to the remaining *trpD•trpC* fusions, to explain their joint locations in TCG-1 for TrpD and TrpC trees. However, the identical overall gene order of the 7-gene operon, the phylogenetic positioning of the remaining five protein domains that are not fused, and the existence of congruent *trpD* remnants in coryneform bacteria are observations that support the conclusion of LGT with conviction. Although the phylogenetic analysis pinpoints the LGT donor of the whole-pathway *trp* operon to a member within the enteric lineage, the subclade that includes *E. coli*, *Salmonella typhimurium*, *Klebsiella pneumoniae*, and *Shigella flexneri* can be excluded. This is because the latter four organisms all possess a recent gene fusion (*trpAb•trpB*) that is absent elsewhere in the enteric lineage (absent also, of course, in *H. pylori* and coryneform bacteria).

Helicobacter

The ancestral *trp* genes of *H. pylori* were displaced by a complete *trp* operon, but apparently without a leader region. *H. hepaticus* is the closest published complete genome to *H. pylori*. The LGT event can be pinpointed after divergence of *H. pylori* and *H. hepaticus* since the *trp* genes of *H. hepaticus* do not cluster within TCG-1. *H. hepaticus* possesses a partial-pathway *trpAa/trpAb•trpB* operon and orphan *trpC* and *trpD* genes. There are indications that the pathway is in a state of rapid deterioration, as has been described for some pathogenic enteric bacteria, for example, in *Actinobacillus actinomycetemcomitans* [6]. Most notably, TrpEb and TrpEa, although still linked, have been scrambled to a divergent orientation (thus, not having the typical operon organization). In addition, *H. hepaticus* TrpEa and TrpEb possess alterations in two and four, respectively, of the residues diagrammed as invariant residues by Xie *et al.* [28]. Thus, one or both may be inactive pseudogenes. In view of this rapid deterioration, the contemporary genome of *C. jejuni* might be more similar to the displaced *H. pylori* Trp genes. The *C. jejuni* operon gene order is *trpAaAb•BCEbEa*. In *C. jejuni trpD* is not fused to *trpC* and, in fact, *trpD* has escaped the otherwise intact *trp* operon. *C. jejuni* does not have a leader peptide encoded by a gene upstream of its *trp* operon.

Corynebacterium

The ancestor of coryneform bacteria also imported a whole-pathway *trp* operon, but in this case including a coding region for a leader peptide. It is significant that three species of coryneform bacteria possess an unlinked second copy of *trpD* (denoted TrpD_r), which is a remnant of the native displaced operon that still persists in other actinomycete *Bacteria*. This conclusion is supported (Figure 9) by the clustering of the stand-alone TrpD_r from both *C. glutamicum* and *C. diptheriae* with TrpD proteins from other actinomycete bacteria (species of *Bifidobacteria*, *Mycobacteria*, *Thermomonospora* and *Streptomyces*) in TCG-5. The latter organisms possess *trpD* within a partial-pathway *trp* operon having the order *trpAaDEbEa* [6]. The elongated branches seen for TrpD_r in Cgl and Cdi in Figure 9 are consistent with drift of the remnant genes from their previous functional role, due to the lack of selective pressure that followed functional replacement by TrpD• of the LGT operon.

It is interesting that *C. diptheriae* exhibits evidence of changes to its *trp* operon that are both relatively recent and disruptive (Figure 12). A duplicate of *trpEb* has been inserted at the beginning of the operon between *trpL* (encoding the leader peptide) and *trpAa*. This TrpEb protein may not be functional, considering that invariant residue 162G (*S. typhimurium* numbering) has been changed to 162E and invariant residue 167K has been

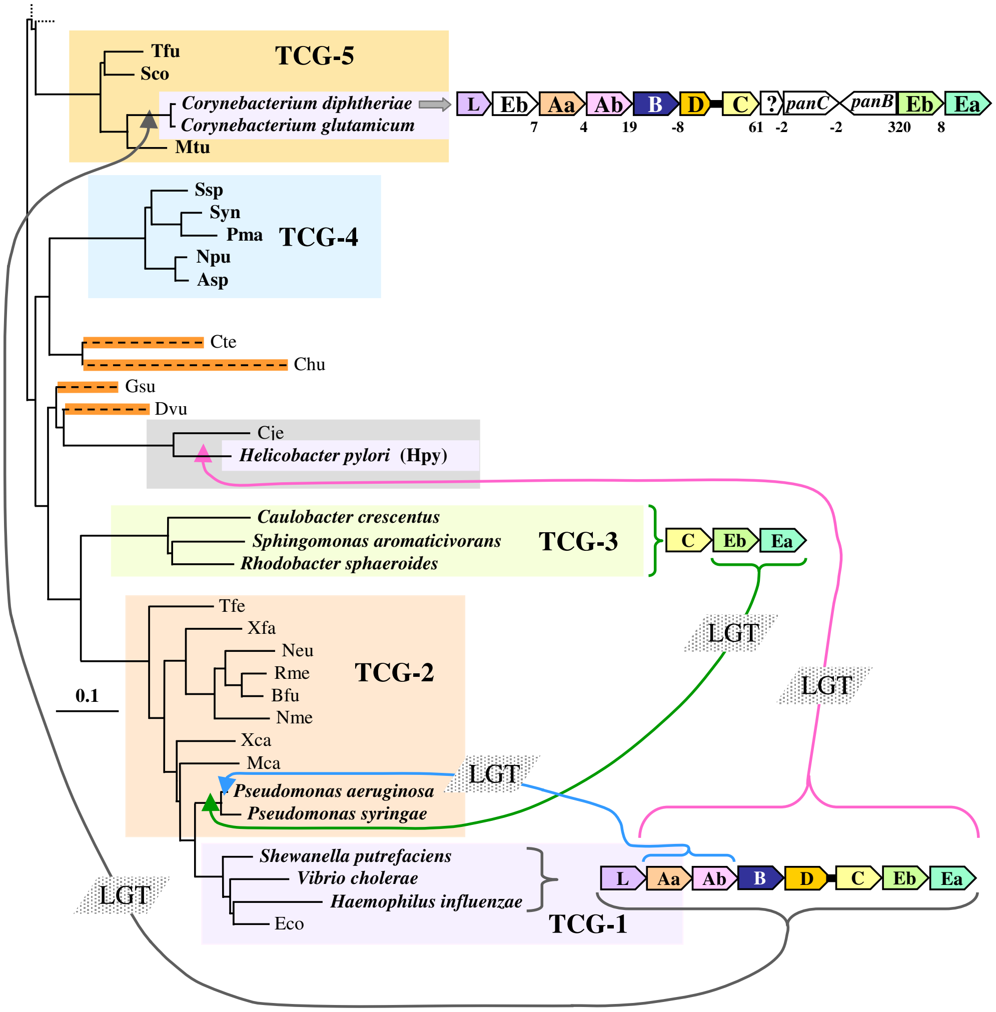

Figure 12
Schematic portrayal of two whole-pathway *trp* operon transfers and two partial-pathway *trp* operon transfers. The partial tree shown is taken from Fig. 1, which identifies the organisms. In *C. diptheriae*, the vertical evolutionary events of gene duplication and insertion that occurred following LGT can be visualized by comparing the gene organization shown to the right (including intergenic spacing) with the gene organization originally received from the enteric donor (bottom right). Gene insertions that followed LGT are shown in white.

changed to 167S, as pointed out earlier [28]. Residue 167K is crucial to the formation of an intersubunit salt bridge with 56D of TrpEa [31].

Another post-LGT event in *C. diptheriae* has been the insertion of two genes of *D*-pantothenate biosynthesis between *trpD•C* and *trpEb*. In *C. glutamicum* the *panB/panC* operon is remote from the *trp* operon and has been well characterized [32]. In *C. diptheriae* the insertion between *trpD•C* and *trpEb* has been in a scrambled orientation such that *panC* and *panB* are transcribed convergently, hence no longer being an operon. The transcription of *panB* in the opposite direction as the flanking genes of *trp* biosynthesis should prevent the formation of a single transcript from the *trp* genes. It would appear that the original xenologous *trp* operon of *C. diptheriae* has been split such that *trpAaAbBD•C* and *trpEbEa* would be separately transcribed. This may illustrate the phenomenon of reductive evolution for a pathogen in the process of discarding un-needed genes.

All elements of operon regulation were not transferred

The *trp* operon of *E. coli* (and presumably other members of the enteric clade) is subject to control at several levels [15,16,27,33]. These include (**i**) allosteric control of anthranilate synthase, whereby Trp acts as a potent feedback inhibitor; and (**ii**) an attenuator mechanism that is mediated by a Trp-rich leader peptide (encoded by *trpL*), and repression control mediated by *trpR*, which binds Trp as a corepressor moiety. Because sensitivity to feedback inhibition is built into the allosteric domain of TrpAa, and because *trpL* is located immediately upstream of the *E. coli trp* operon, it is not surprising that recipient organisms such as the coryneform bacteria possess both of the latter regulatory features. However, *trpR* is distant from the operon and was not co-transferred with the operon.

Perhaps the contemporary *trp* operon acquired by LGT offered selective advantages to the ancestor of coryneform bacteria, but Trp regulation in coryneform bacteria would appear to be relatively unsophisticated without *trpR*. In *E. coli* the impact of attenuation is relatively weak compared with the impact of repression [16,27]. Repression detects free Trp whereas attenuation detects uncharged tRNA^{Trp}. Since free Trp concentrations can be fairly low and still maintain highly charged tRNA^{Trp}, *trpR*-mediated repression acts over a large range of expression. Relief from attenuation ensues after maximal derepression has occurred. In *E. coli*, *trpR* also functions beyond the Trp pathway in that it binds to an operator for an initial gene of aromatic biosynthesis.

LGT of **trpL** *in coryneform bacteria*

The existence of a leader peptide associated with regulation of the *trp* operon (and, indeed, the unexpectedly close relationship with the *E. coli trp* operon) was reported some time ago in *C. glutamicum* [34-36]. The leader-peptide region (encoded by *trpL*) upstream of the *trp* operons of *C. glutamicum*, *C. efficiens*, and *C. diptheriae* are shown (Figure 13) in comparison with the corresponding region of *V. cholerae*, a member of the enteric lineage that represents the LGT donor. In *V. cholerae*, *C. glutamicum*, and *C. efficiens*, *trpL* is immediately upstream of *trpAa*, the initial structural genes of the *trp* operon in these organisms. In *C. diptheriae* a copy of *trpEb* has recently been inserted between *trpL* and *trpAa* (Figure 12). The putative start codons of *trpL* in *Vibrio cholerae*, *C. glutamicum*, and *C. diptheriae* are ATG, TTG, ATG, and GTG, respectively. These yield leader peptides of 30 amino acids in length. The start codon is uncertain, and Heery and Dunican [35] have suggested a start codon (vertical black arrow) for *C. glutamicum* that would yield a 17-amino acid leader peptide.

In the case of *H. pylori*, the *trp* operon was either acquired without *trpL*, or *trpL* was soon lost. Thus, the *trp* pathway of *H. pylori* appears to lack regulation by both attenuation and repression. Assuming that the displaced pathway of *H. pylori* was similar to that of the contemporary *C. jejuni* or other epsilon-proteobacteria, it would be interesting to compare the relative efficiencies of the modern Trp pathways in these fairly close relatives.

LGT of partial-pathway Trp operons

P. aeruginosa possesses a partial-pathway operon consisting of genes encoding the two subunits of tryptophan synthase that are of xenologous origin. These *trpEb/trpEa* genes are engaged in primary biosynthesis. The donor is clearly a member of the TCG-3 clade. Thus, in the scenario considered [6] for gene separation of the ancestral *trpC/trpEb/trpEa* operon to yield a stand-alone *trpC* gene and a *trpEb/trpEa* operon, the native *trpEb/trpEa* must have been displaced by LGT, after the intragenomic translocation. Because the regulatory gene, *trpI*, is adjacent in divergent orientation to *trpEb* in the *P. aeruginosa/P. syringae* clade (and three other close relatives; see asterisks in Table 3), it will be interesting if any genomic members of TCG-3 to be sequenced in the future prove to have *trpI* adjacent to a *trpEb/trpEa* operon. (However, TrpI is a member of the ubiquitous and large family of LysR transcriptional activators, and could easily have emerged recently in the *P. aeruginosa* clade following gene duplication.)

P. aeruginosa also possesses a partial-pathway *trp* operon encoding the two subunits of anthranilate synthase. Phylogenetic analysis indicates that these are *trpAa trpAb* xenologs that originated from within the enteric lineage. As with the whole-pathway operons just discussed, the *E. coli/S. typhimurium/K. pneumoniae/S. flexneri* clade can be excluded as the specific donor because this clade possesses a *trpAb•trpB* fusion. This xenologous two-gene operon of

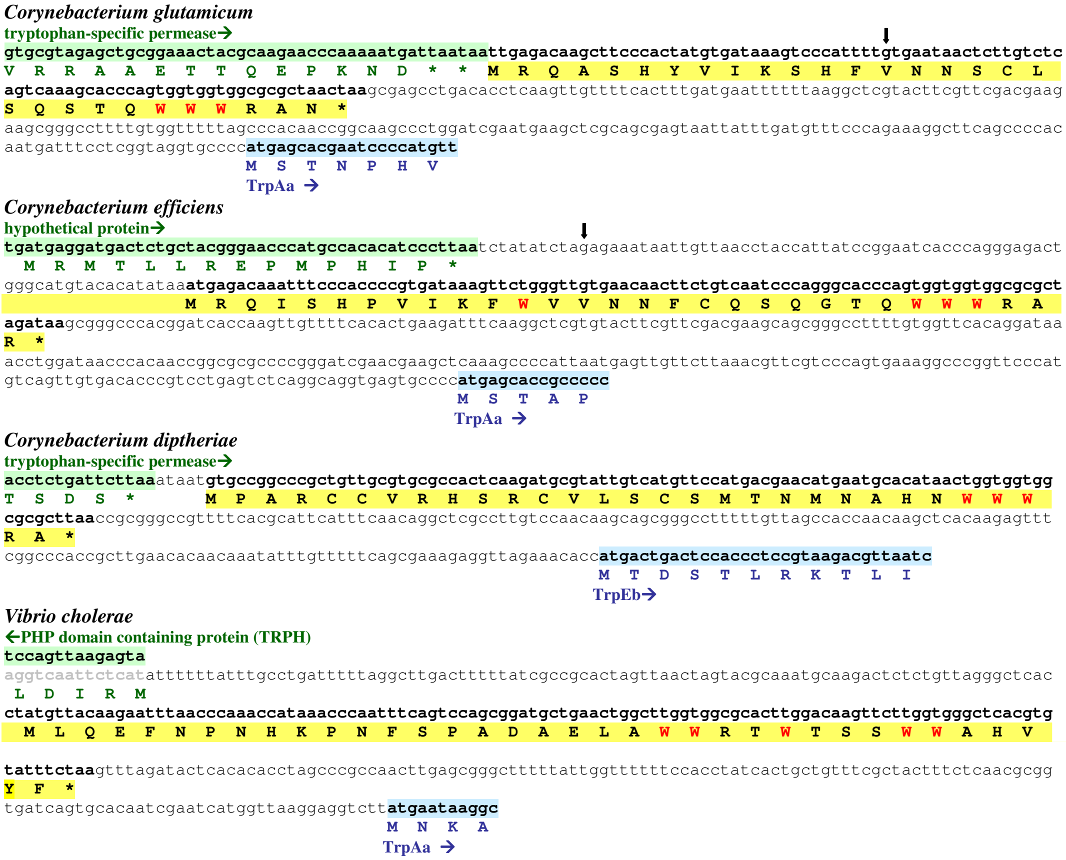
Corynebacterium glutamicum
tryptophan-specific permease→
gtgcgtagagctgcggaaactacgcaagaacccaaaaatgattaataattgagacaagcttcccactatgtgataaagtcccattttgtgaataactcttgtctc
V R R A A E T T Q E P K N D * * M R Q A S H Y V I K S H F V N N S C L
agtcaaagcacccagtggtggtggcgcgctaactaagcgagcctgacacctcaagttgttttcactttgatgaatttttaaggctcgtacttcgttcgacgaag
S Q S T Q W W W R A N *
aagcgggccttttgtggttttagcccacaaccggcaagccctggatcgaatgaagctcgcagcgagtaattatttgatgtttcccagaaaggcttcagccccac
aatgatttcctcggtaggtgccccatgagcacgaatccccatgtt
M S T N P H V
TrpAa →
Corynebacterium efficiens
hypothetical protein→
tgatgaggatgactctgctacgggaacccatgccacacatcccttaatctatatctagagaaataattgttaacctaccattatccggaatcacccagggagact
M R M T L L R E P M P H I P *
gggcatgtacacatataaatgagacaaatttcccaccccgtgataaagttctgggttgtgaacaacttctgtcaatcccagggcacccagtggtggtggcgcgct
M R Q I S H P V I K F W V V N N F C Q S Q G T Q W W W R A
agataagcgggcccacggatcaccaagttgttttcacactgaagatttcaaggctcgtgtacttcgttcgacgaagcagcgggccttttgtggttcacaggataa
R *
acctggataacccacaaccggcgcgccccgggatcgaacgaagctcaaagccccattaatgagttgttcttaaacgttcgtcccagtgaaaggcccggttcccat
gtcagttgtgacacccgtcctgagtctcaggcaggtgagtgccccatgagcaccgccccc
M S T A P
TrpAa →
Corynebacterium diptheriae
tryptophan-specific permease→
acctctgattcttaaataatgtgccggcccgctgttgcgtgcgccactcaagatgcgtattgtcatgttccatgacgaacatgaatgcacataactggtggtgg
T S D S * M P A R C C V R H S R C V L S C S M T N M N A H N W W W
cgcgcttaaccgcgggccgttttcacgcattcatttcaacaggctcgccttgtccaacaagcagcgggccttttttgttagccaccaacaagctcacaagagttt
R A *
cggcccaccgcttgaacacaacaaatatttgtttttcagcgaaagaggttagaaacaccatgactgactccaccctccgtaagacgttaatc
M T D S T L R K T L I
TrpEb→
Vibrio cholerae
←PHP domain containing protein (TRPH)
tccagttaagagta
aggtcaattctcatatttttatttgcctgattttaggcttgactttttatcgccgcactagttaactagtacgcaaatgcaagactctctgttagggctcac
L D I R M
ctatgttacaagaatttaacccaaaccataaacccaatttcagtccagcggatgctgaactggcttggtggcgcacttggacaagttcttggtgggctcacgtg
M L Q E F N P N H K P N F S P A D A E L A W W R T W T S S W W A H V
tatttctaagtttagatactcacacacctagcccgccaacttgagcgggcttttattggtttttccacctatcactgctgtttcgctactttctcaacgcgg
Y F *
tgatcagtgcacaatcgaatcatggttaaggaggtcttatgaataaggc
M N K A
TrpAa →

Figure 13
Comparison of *trpL* leader regions in species of coryneform bacteria and in a representative enteric bacterium. In the coryneform bacteria, the start codon for *trpL* is uncertain, and second start sites at an internal position for *C. glutamicum* and *C. efficiens* are indicated by black, vertical arrows.

P. aeruginosa (originally denoted *phnA* and *phnB* by Crawford *et al.* [25,37]) has not displaced the corresponding *trp* genes that are engaged in primary biosynthesis. Although *phnA* and *phnB* were originally thought to function in phenazine production, as the naming implies, phenazine pigments are not derived from anthranilate [26]. Instead, the xenolog pair appear to have an unknown specialized function that is geared to stationary-phase physiology [38]. The LGT acquisition of the *phnA*/*phnB* xenolog operon by *P. aeruginosa* occurred after its divergence from close relatives such as *P. putida*, *P. fluorescens*, *P. syringae* and *Azotobacter vinelandii* because none of these relatives have the xenologous *phnA*/*phnB* operon. Hence, this acquisition appears to be quite recent.

Species of *Xylella* also possess a xenologous two-gene *trpAa*/*trpAb* partial-pathway *trp* operon that appears to have a specialized function that is distinct from homologs engaged in primary Trp biosynthesis. Xie *et al.* [7] have suggested that the gene denoted *acl* probably encodes an aryl-CoA ligase. Since *acl* appears to exist within the *Xylella trpAa*/*trpAb* operon, its gene product may have the specificity of anthranilate-CoA ligase. Activated anthranilate may then function as a key precursor for production of antibiotic, siderophore, and so on. Note that, if this is correct, the primary *trp*-pathway genes *trpB*/*trpC*/*trpD*/*trpEa*/*trpEb* are irrelevant to the specialized pathway. Thus, reference to the xenologous *trpAa*/*trpAb* pair as a "partial-pathway operon" could be somewhat a misnomer, and perhaps "hybrid operon" might prove to be more apt. The origin of the *Xylella trpAa*/*trpAb* genes might have been from a close relative of *C. jejuni* because the best matches of both genes are with *C. jejuni*. It is also suggestive that these genes have a low-GC ratio (about 38%), compared with a genomic GC ratio of 30.2% for *C. jejuni*. *C. jejuni* cannot have been the direct donor, however, because it possesses a *trpAb•trpB* fusion.

Conclusions

The overall impact of LGT in *Bacteria* and *Archaea* is currently a highly contentious issue. Critical reviews written from completely different viewpoints are recommended for a sense of the status of the ongoing debate, as well as for a substantial listing of key references [39,40]. Gogarten *et al.* [21] have summarized a contemporary rationale to describe the evolutionary process as a phylogenetic "synthesis" that could integrate a traditional tree-like behavior (vertical descent of genes) and web-like, reticulate behavior (LGT). The latter paper follows up on a balanced and highly insightful review produced by Doolittle in 1999 [22]. In an essay by Martin [41] about the extent to which bacterial chromosomes might be mosaic, it was emphasized that "careful gene-by-gene phylogenetic comparisons in addition to genome-by-genome comparisons.....are needed." We consider the Trp system to be a particularly apt choice for gaining detailed insight into the relative contributions of lateral and vertical events in the evolutionary history of a major segment of primary metabolism. On the one hand, genes of Trp biosynthesis are expected to represent a level and type of fundamental metabolism that is least prone to LGT if it is correct that a "core" of genes exists that is relatively recalcitrant to LGT [23,42]. On the other hand, the wide distribution of intact whole-pathway *trp* operons should facilitate the probability of LGT (at least the initial acquisition event). For the Trp pathway, we conclude that events of LGT and paralogy do not obliterate the vertical trace of evolutionary history. This view is reinforced by the contention that whole-genome trees (mean pairwise similarities between shared genomic proteins) have largely converged on the rRNA-sequence tree [43]. The latter work was preceded by the gene-content "trees" introduced by Snel *et al.* [44]. Also, see Eisen [45].

Gogarten *et al.* [21] have discussed the radical possibility that rRNA genes are highly mosaic and are so useful for prokaryotic taxonomy "precisely because they are mosaics and reflect the mosaic character of the genome as a whole". According to this view, a vertical trace of genealogical history will not be found, because it only exists in short jumps. (If so, the two whole-pathway LGT events described here reflect LGT events that stand out against the overall mosaic trends.)

The above caveat aside, 16S rRNA trees do appear to provide a reasonable guide to the vertical trace of evolutionary descent [46]. If so, the mapping of the Trp-pathway system upon the 16S rRNA tree exemplifies a case where the evolutionary history can indeed be tracked as a vertical genealogy that features some intriguing reticulate relationships. Proteins of the Trp pathway generally exhibit a genealogy that is parallel with the 16S rRNA tree. In the case of *Helicobacter* and *Corynebacterium* species, the genomes are a mosaic with respect to Trp genes. This means that the history of the *trp* genes in the latter organisms, instead of being that shared with their closest relatives, is the same as that of enteric bacteria (the donor lineage) up to the time of LGT. Following this time a new vertical genealogical progression has begun. Each organism (lineage) can be envisioned to possess an individualistic repertoire of chimeric features that have been assimilated into a recognizable skeleton of vertical events. It would seem that the evolutionary history of the primary pathway of Trp biosynthesis parallels the organismal phylogeny in most organisms. This presently includes all of the organisms in Figure 2 that belong to TCG groupings, except for *H. pylori* and the coryneform bacteria. Mosaicism can occasionally exist at a level of individual *trp* genes, as illustrated by *P. aeruginosa* and its closest relatives. Here the history of *trpAa*/*trpAb*/*trpB*/*trpC*/*trpD* (primary biosynthesis) parallels the organismal phylogeny, but *trpEb*/*trpEa* exhibit a reticulate genealogy.

It will be important to obtain other analyses of similar detail for other pathways to assess to what extent the evolutionary process that underpins Trp biosynthesis reflects the general process in other pathways. We have already seen in preliminary work that construction of similar congruency groups pertinent to common aromatic-pathway biosynthesis, folate biosynthesis, and histidine biosynthesis has good potential to expand the analysis. It seems probable that there are no prokaryote species whose evolutionary history exclusively involved a vertical line of descent for all of its genes. Lawrence [47] has written a cogent essay in support of the view that the Linnean paradigm of hierarchical descent fails to describe the evolution of prokaryotes because LGT occurs "at all levels of taxonomic inclusiveness". However, Woese *et al.* [48] argue that the dynamic of LGT has become progressively diminished through time as simple, primitive early cells became complex and refined. This is in accord with our thesis [6] that early and simple *trp*-gene assemblages may have been unstable until their progression to complexities of regulation and to establishment of individualistic metabolic ties that conferred increasing operon stability.

Methods

Genomes

Refer to the Gold Genomes Online Database [49] for a list of published complete genomes and links to the corresponding references. Synonymous names include *Rhizobium meliloti = Sinorhizobium meliloti; Corynebacterium glutamicum = Brevibacterium lactofermentum = Brevibacterium flavum = Brevibacterium divaricatum; Rhizobium meliloti = Sinorhizobium meliloti; Rhizobium loti = Mesorhizobium loti; Thermomonospora fusca = Thermobifida fusca; Chlamydophila psittaci = Chlamydophila caviae; Thiobacillus ferrooxidans = Acidithiobacillus ferrooxidans; Shewanella putrefaciens = Shewanella oneidensis; Sphingomonas aromaticivorans = Novosphingobium aromativorans*.

16S rRNA phylogenetic trees

16S rRNA subtrees were derived from the Ribosomal Database site [50,51].

Protein phylogenetic trees

Unrooted phylogenetic trees were derived from ClustalW alignment [52,53]. The neighbor-joining and Fitch [54] programs were employed to obtain distance-based trees. The distance matrix was obtained using Protdist with a Dayhoff Pam matrix. The Seqboot and Consense programs were then used to assess the statistical strength of the tree using bootstrap resampling. Neighbor-joining and Fitch trees yielded similar clusters and arrangement of taxa within them. Bootstrap values indicate the number of times a node was supported in 100 resampling replications.

Concatenated sequences of Trp-pathway proteins

Multiple sequence alignments were derived by input of the indicated homolog amino acid sequences into the ClustalW program (Version 1.4) [52,53]. Manual alignment adjustments were made as needed with the assistance of the BioEdit multiple alignment tool of Hall [55].

After each of the individual Trp-protein domain alignments were generated, both N-terminal and C-terminal unaligned regions were trimmed manually through visual inspection. Then the seven protein domains responsible for the Trp pathway of biosynthesis were concatenated in the order: TrpAa/TrpAb/TrpB/TrpC/TrpD/TrpEa/TrpEb. The resulting concatenated multiple alignment was used as input for generation of a phylogenetic tree using the program package PHYLIP [56].

Analysis of raw DNA sequence data

Raw DNA contig sequences available from NCBI [57] and the TIGR Unfinished Microbial Genomes database [58] were screened using the built-in BLAST service. The protein sequences from GenBank were used as query entries. The BLAST 2.0 and the ORF Finder (Open Reading Frame Finder) offered by NCBI [59] were used to locate open reading frames and to confirm the similarity search result of the raw sequence.

Analysis of fusion proteins

Fusion protein sequences from GenBank and NCBI Microbial Genomes Blast Databases [57] were screened using the BLAST [60] program. Multiple alignments were obtained by input of single-domain and fusion-protein sequences into the ClustalW [52,53] program (version 1.4).

Authors' contributions

The major roles of the authors were as follows. GX obtained the sequences, did the alignments, and obtained the phylogenetic trees. CB drew the figures. JS undertook the systematic management and organization of sequence data with the objective of achieving a dynamic and progressively updateable website. NK managed and coordinated the general laboratory operations. The initial guiding concepts and the initial draft came from RJ. All of the authors participated in the data analyses and in the formulation of conclusions. All of the authors read and approved the final version of the manuscript.

Additional material

Additional File 1

Table 4, entitled "Keys to sequence identifiers", is provided as supplementary material in an html document. This table contains a full collection of sequence data and annotations contained in this paper, and gi (gene identification) numbers are included and hyperlinked to facilitate access to the corresponding GenBank records. For future reference to a progressively updated table, refer to http://biosphere.lanl.gov/aropath/table4.html.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1741-7007-2-15-S1.html]

Acknowledgements

This is Florida Agricultural Experiment Station Journal Series No. R-09802. R Jensen thanks the National Library of Medicine (Grant G13 LM008297) for partial support. Some of the preliminary sequence data were obtained from the Institute for Genomic Research [58], National Center for Biotechnology Information [57], and Integrated Genomics, Inc. [61]. We thank Charles Yanofsky and Ross Overbeek for their critiques of an early version of this manuscript.

References

1. Lawrence JG, Ochman H: **Molecular archaeology of the *Escherichia coli* genome.** *Proc Natl Acad Sci USA* 1998, **95:**9413-9417.
2. Lawrence JG, Ochman H: **Amelioration of bacterial genomes: rate of change and exchange.** *J Mol Evol* 1997, **44:**383-397.
3. Song J, Ware A, Liu S-L: **Wavelet to predict bacterial *ori* and *ter*: a tendency towards a physical balance.** *BMC Genomics* 2003, **4:**17.

4. Lawrence J: **Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes.** *Curr Opin Genet Dev* 1999, **9:**642-648.
5. Lawrence JG: **Selfish operons and speciation.** *Trends Microbiol* 1997, **5:**355-359.
6. Xie G, Keyhani NO, Bonner CA, Jensen RA: **The ancient origin of the tryptophan operon and tracking the subsequent dynamics of evolutionary change.** *Microbiol Mol Biol Rev* 2003, **67:**303-342.
7. Xie G, Bonner CA, Brettin T, Gottardo R, Keyhani NO, Jensen RA: **Lateral gene transfer and ancient paralogy of operons containing redundant copies of tryptophan-pathway genes in *Xylella* species and in heterocystous cyanobacteria.** *Genome Biol* 2003, **4:**R-14.
8. Lawrence JG: **Gene transfer, speciation, and the evolution of bacterial genomes.** *Curr Opin Microbiol* 1999, **2:**519-523.
9. Boucher Y, Huber H, L'Haridon S, Stetter KO, Doolittle WF: **Bacterial origin for the isoprenoid biosynthesis enzyme HMG-CoA reductase of the archaeal orders *Thermoplasmatales and Archaeoglobales*.** *Mol Biol Evol* 2001, **18:**1378-1388.
10. Xie G, Bonner CA, Jensen RA: **Dynamic diversity of the tryptophan pathway in chlamydiae: reductive evolution and a novel operon for tryptophan recapture.** *Genome Biol* 2002, **3:**research0051.1-0051.17.
11. Fehlner-Gardniner C, Roshick C, Carlson JH, Hughes S, Belland RJ, Caldwell HD, McClarty G: **Molecular basis defining human *Chlamydia trachomatis* tissue tropism: a possible role for tryptophan synthase.** *J Biol Chem* 2002, **277:**26893-26903.
12. Wood H, Roshick C, McClarty G: **Tryptophan recycling is responsible for the interferon-gamma resistance of *Chlamydia psittaci* GPIC in indoleamine dioxygenase-expressing host cells.** *Mol Microbiol* 2004, **52:**903-916.
13. Barona-Gómez F, Hodgson DA: **Occurrence of a putative ancient-like isomerase involved in histidine and tryptophan biosynthesis.** *EMBO Reports* 2003, **4:**296-300.
14. Osterman A, Overbeek R: **Missing genes in metabolic pathways: a comparative genomics approach.** *Curr Opin Chem Biol* 2003, **7:**238-251.
15. Yanofsky C: **Advancing our knowledge in biochemistry, genetics, and microbiology through studies on tryptophan metabolism.** *Ann Rev Biochem* 2001, **70:**1-37.
16. Yanofsky C: **Reflections: using studies on tryptophan metabolism to answer basic biological questions.** *J Biol Chem* 2003, **278:**10859-10878.
17. Auerbach S, Gao J, Gussin GN: **Nucleotide sequences of the *trpI*, *trpB*, and *trpA* genes of *Pseudomonas syringae*: positive control unique to fluorescent pseudomonads.** *Gene* 1993, **123:**25-32.
18. Hu DS-J, Hood DW, Heidstra R, Hodgson D: **The expression of the *trpD*, *trpC* and *trpBA* genes of *Streptomyces coelicolor* A3(2) is regulated by growth rate and growth phase but not by feedback repression.** *Mol Microbiol* 1999, **32:**869-880.
19. Ryding NJ, Anderson TB, Champness WC: **Regulation of the *Streptomyces coelicolor* calcium-dependent antibiotic by *absA*, encoding a cluster-linked two-component system.** *J Bacteriol* 2002, **184:**794-805.
20. Brown JR, Doolittle WF: ***Archaea* and the prokaryote-to-eukaryote transition.** *Microbiol Mol Biol Rev* 1997, **61:**456-502.
21. Gogarten JP, Doolittle WF, Lawrence JG: **Prokaryotic evolution in light of gene transfer.** *Mol Biol Evol* 2002, **19:**2226-2238.
22. Doolittle WF: **Phylogenetic classification and the universal tree.** *Science* 1999, **284:**2124-2128.
23. Eisen JA: **Horizontal gene transfer among microbial genomes: new insights from complete genome analysis.** *Curr Opin Genet Dev* 2000, **10:**606-611.
24. Eisen JA: **Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis.** *Genome Res* 1998, **8:**163-167.
25. Crawford IP, Milkman R: **Orthologous and paralogous divergence, reticulate evolution, and lateral gene transfer in bacterial *trp* genes.** In: *Evolution at the Molecular Level* Edited by: Selander RK, Clark AG, Whittam TS. Sunderland, Mass.: Sinauer & Assoc; 1990:77-95.
26. McDonald MD, Mavrodi DV, Thomashow LS, Floss HG: **Phenazine biosynthesis in *Pseudomonas fluorescens* : branchpoint from the primary shikimate biosynthetic pathway and role of phenazine-1,6-dicarboxylic acid.** *J Am Chem Soc* 2001, **123:**9459-9460.
27. Yanofsky C, Miles E, Bauerle R, Kirschner K: **Trp Operon.** *Volume 4*. New York: John Wiley & Sons, Inc; 1999.
28. Xie G, Forst C, Bonner CA, Jensen RA: **Significance of two distinct types of tryptophan synthase beta chain in *Bacteria*, *Archaea* and higher plants.** *Genome Biology* 2001, **3:**research005.1-005.13.
29. Hooper SD, Berg OG: **Duplication is more common among laterally transferred genes than among indigenous genes.** *Genome Biol* 2003, **4:**R48.
30. Calhoun DH, Bonner CA, Gu W, Xie G, Jensen RA: **The emerging periplasm-localized subclass of AroQ chorismate mutases, exemplified by those from *Salmonella typhimurium* and *Pseudomonas aeruginosa*.** *Genome Biol* 2001, **2:**research0030.1-0030.16.
31. Miles EW, Bauerle R, Ahmed SA: **Tryptophan synthase from *Escherichia coli* and *Salmonella typhimurium*.** *Methods Enzymol* 1987, **142:**398-414.
32. Sahm H, Eggeling L: ***D*-Pantothenate synthesis in *Corynebacterium glutamicum* and use of *panBC* and genes encoding *L*-valine synthesis for *D*-pantothenate overproduction.** *Appl Environ Microbiol* 1999, **65:**1973-1979.
33. Xiu ZI, Chang ZY, Zeng AP: **Nonlinear dynamics of regulation of bacterial *trp* operon: model analysis of integrated effects of repression, feedback inhibition, and attenuation.** *Biotech* 2002, **18:**686-693.
34. Matsui K, Sano K, Ohtsubo E: **Sequence analysis of the *Brevibacterium lactofermentum trp* operon.** *Mol Gen Genet* 1987, **209:**299-305.
35. Heery DM, Dunican LK: **Cloning of the *trp* gene cluster from a tryptophan-hyperproducing strain of *Corynebacterium glutamicum*: identification of a mutation in the *trp* leader sequence.** *Appl Envir Microbiol* 1993, **59:**791-799.
36. Su YC, Chen SL: **Cloning of the tryptophan operon of *Brevibacterium divaricatum* and its expression in *E. coli*.** *Proc Natl Acad Sci USA* 1996, **20:**87-91.
37. Essar DW, Eberly L, Hadero A, Crawford IP: **Identification and characterization of genes for a second anthranilate synthase in *Pseudomonas aeruginosa* : interchangeability of the two anthranilate synthases and evolutionary implications.** *J Bacteriol* 1990, **172:**884-900.
38. Mavrodi DV, Bonsall RF, Delaney SM, Soule MJ, Phillips G, Thomashow LS: **Functional analysis of genes for biosynthesis of pyocyamin and phenazine-1-carboxamide from Pseudomonas aeruginosa PAO1.** *J Bacteriol* 2001, **183:**6454-6465.
39. Kurland CG, Canback B, Berg OG: **Horizontal gene transfer: A critical view.** *Proc Natl Acad Sci USA* 2003, **100:**9658-9662.
40. Lawrence JG, Hendrickson H: **Lateral gene transfer: when will adolescence end?** *Mol Microbiol* 2003, **50:**739-749.
41. Martin W: **Mosaic bacterial chromosomes: a challenge en route to a tree of genomes.** *Bioessays* 1999, **21:**99-104.
42. Lerat E, Daubin V, Moran NA: **From gene trees to organismal phylogeny in prokaryotes: the case of the alpha-proteobacteria.** *PLoS Biology* 2003, **1:**E1.
43. Clarke AR, Beiko RG, Ragan MA, Charlebois RL: **Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLAST scores.** *J Bacteriol* 2002, **184:**2072-2080.
44. Snel B, Bork P, Huynen MA: **Genome phylogeny based on gene content.** *Nat Genet* 1999, **21:**108-110.
45. Eisen JA: **Assessing evolutionary relationships among microbes from whole-genome analysis.** *Curr Opin Microbiol* 2000, **3:**475-480.
46. Woese CR: **Interpreting the universal phylogenetic tree.** *Proc Natl Acad Sci USA* 2000, **97:**8392-8396.
47. Lawrence J: **Gene transfer in *Bacteria* : speciation without species?** *Theoret Populat Biol* 2002, **61:**449-460.
48. Woese CR, Olsen GJ, Ibba M, Soll D: **Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process.** *Microbiol Mol Biol Rev* 2000, **64:**202-236.
49. **Gold Genomes OnLine Database** [http://www.genomesonline.org/]
50. Maidak BL, Cole JR, Lilburn TG, Parker CT Jr, Saxman PR, Farris RJ, Garrity GM, Olsen GJ, Schmidt TM, Tiedje JM: **The RDP-II (Ribosomal Database Project).** *Nucl Acids Res* 2001, **29:**173-17.
51. **Ribosomal Database Project II** [http://rdp.cme.msu.edu/html]

52. Thompson JD, Higgins DG, Gibson TJ: **Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucl Acids Res* 1994, **22:**4673-4680.
53. **ClustalW, version 1.4** [http://www.ebi.ac.uk/clustalw]
54. Fitch WM: **Toward defining the course of evolution: minimum change for a specific tree topology.** *Syst Zool* 1971, **20:**406-416.
55. Hall TA: **Biological sequence alignment editor for Windows 95/98/NT, 4.8.6.** 2000 [http://http//:www.mbio.ncsu.edu/BioEdit/bioedit.html].
56. Felsenstein J: **PHYLIP-Phylogeny Inference Package (version 3.2).** *Cladistics* 1989, **5:**164-166.
57. **National Center for Biotechnology Information** [http://www.ncbi.nlm.nih.gov/Microb_blast/unfinishedgemome.html.old/]
58. **The Institute for Genomic Research** [http://www.tigr.org/tdb/ufmg/]
59. **National Center for Biotechnology Information** [http://www.ncbi.nlm.nih.gov/]
60. **BLAST** [http://www.ncbi.nlm.nih.gov/blast/]
61. **Integrated Genomics, Inc** [http://wit.integratedgenomics.com/ERGO]