Tracing the Ancestry of Operons in Bacteria

Huy Nguyen[1,4], Ashish Jain[2,3], Oliver Eulenstein[3,4], Iddo Friedberg[1,3]*,

1 Dpt. of Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, IA, USA
2 Dpt. of Genetics, Development, and Cell Biology, Iowa State University, Ames, IA, USA
3 Program in Bioinformatics and Computational Biology, Iowa State University, Ames, IA, USA
4 Dpt. of Computer Science, Iowa State University, Ames, IA, USA

* idoerg@iastate.edu

Abstract

Complexity is a fundamental attribute of life. Complex systems are made of parts that together perform functions that a single component, or subsets containing individual components, cannot. Examples of complex molecular systems include protein structures such as the F_1F_o-ATPase, the ribosome, or the flagellar motor: each one of these structures requires most or all of its components to function properly. Given the ubiquity of complex systems in the biosphere, understanding the evolution of complexity is central to biology. At the molecular level, operons are a classic example of a complex system. An operon's genes are co-transcribed under the control of a single promoter to a polycistronic mRNA molecule, and the operon's gene products often form molecular complexes or metabolic pathways. With the large number of complete bacterial genomes available, we now have the opportunity to explore the evolution of these complex entities, by identifying possible intermediate states of operons. In this work, we developed a maximum parsimony algorithm to reconstruct ancestral operon states, and show a simple vertical evolution model of how operons may evolve from the individual component genes. We describe several ancestral states that are plausible functional intermediate forms leading to the full operon. We also offer **R**econstruction **o**f **A**ncestral **G**ene blocks **U**sing **E**vents or ROAGUE as a software tool for those interested in exploring gene block and operon evolution.
The software accompanying this paper is available under GPLv3 license in:
`https://github.com/nguyenngochuy91/Ancestral-Blocks-Reconstruction`.
All figures in this paper are available in enlarged downloadable form from:
`https://github.com/nguyenngochuy91/Ancestral-Blocks-Reconstruction/tree/master/images`

Introduction

The evolution of complex systems is an open problem in biology [1–3], and has recently been studied intensively in genomes [4–6]. To better understand how complex systems evolve, we focus on the problem of the evolution of orthologous gene blocks and operons in bacteria. Orthologous gene blocks or *orthoblocks* are sequences of genes co-located on the chromosomes of several species, whose evolutionary conservation is apparent. Operons can be viewed as a special case of gene blocks where the genes are co-transcribed to polycistronic mRNA and are often associated with a coherent function, such as a metabolic pathway or a protein complex. Several models have been proposed to explain gene block and operon evolution, and it may very well be that the models are not mutually exclusive, and different operons may evolve by different models, or indeed a single operon may be the result of the combination of several models [7–14].

Previously, we proposed a method that explains the evolution of orthoblocks and operons as a combination of events that take place in vertical evolution from common ancestors [15]. In the evolution of an orthoblock, the different gene blocks may gain or lose genes, have genes duplicated, or have them split off (Figure 1 and Table 1). By determining the frequency of the events for any orthoblock in a studied clade, we can determine a cost for each event, and thus create a cost function to determine an optimal vertical path for the evolution of orthoblocks. We used the cost function to determine the conservation of some operons and orthoblocks in proteobacteria, and show that orthoblocks that perform cellular information processing (such as mRNA translation) are more conserved than those that are associated with adaptation to specific environments.

In this study we use the orthoblock evolution cost function to reconstruct ancestral gene blocks. Reconstructing plausible ancestral states of extant complex entities can help us understand how they evolve, identify possible functional intermediate states, and determine which forces might affect their evolution. The rest of this paper is structured as follows: first, we briefly describe our approach. We then present and discuss the results using our algorithms to reconstruct the ancestral states of orthoblocks in a clade of Gram-negative bacteria and a clade of Gram-positive bacteria. This reconstruction involves orthoblocks comprising genes orthologous to those found in operons in *Escherichia coli* and in *Bacillus subtilis*, respectively. Our reconstructions of ancestral states show that: (1) some operons can rapidly evolve independently in several branches in their respective clades, suggesting that positive selection plays a major role in the evolution of gene blocks in bacteria; (2) other operons are highly conserved, their evolution predating the last common ancestor of the clades we chose, (3) some ancestral state can plausibly be described as intermediate functional forms, and (4) some operon conservation is sporadic and cannot be explained solely by vertical transmission suggesting horizontal gene transfer.

Materials and Methods

Definitions

Gene block-based evolutionary events, and event-based distances

A **reference taxon** is a taxon where operons have been identified by experimental means. Here we use *E. coli* K-12 MG1655 and *B. subtilis* as reference taxa. The reference taxon serves as a standard of truth to determine if the genes on a suspected orthoblock do indeed reside, at least in one species, in an operon or a similar co-regulated gene block. **Neighboring genes**: two genes are considered neighboring if they are ≤ 500 nucleotides apart and on the same strand. A **gene block** comprises no less than two neighboring open reading frames (ORFs). **Orthoblocks**, gene blocks that are orthologous, are defined as follows: two organisms have orthoblocks when each organism must have at least two neighboring genes that are homologous to genes in a gene block in the reference taxon's genome. An **event** is a change in the gene block between any two species with homologous gene blocks.

We identify three types of pairwise events between orthoblocks in different taxa: splits, deletions, and duplications. The *event-based distance* between any two orthoblocks is the sum of the minimized count of splits, duplications, and deletions, which is elaborated upon in **Orthoblock Distance Functions**. See Figure 1. The terms reference taxa, neighboring genes, gene blocks, events, and orthoblocks are elaborated upon in [15].

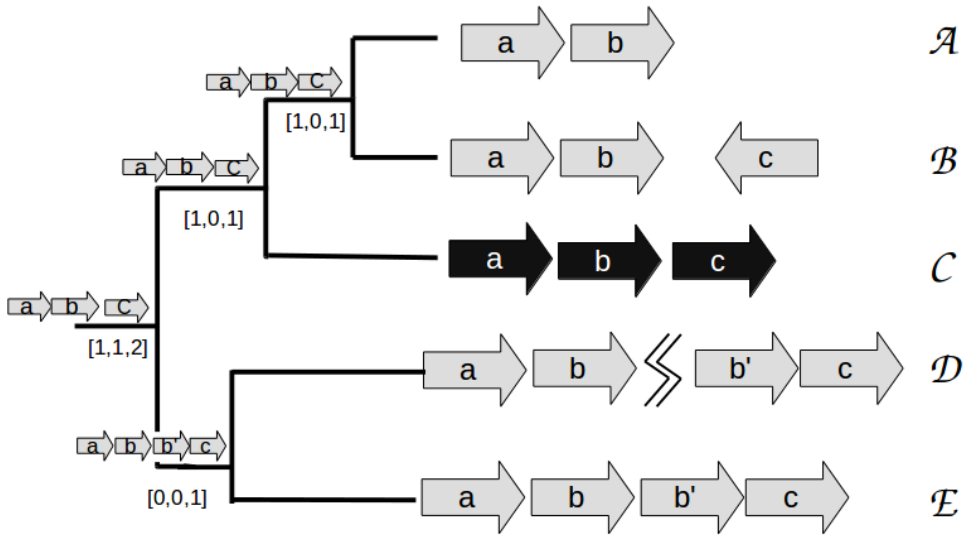

Fig 1. Orthoblocks from species A-E are arranged in a species phylogenetic tree. Species C has an experimentally-determined operon (Black arrows), and serves as the reference taxon. The orthologs in species A,B,D, and E were determined as explained the in the text. The events between C and all other species for this orthoblock are: A-C: deletion (of gene c)
B-C: split (of gene c)
C-D: duplication (of b) and split (jagged line)
C-E: duplication (of b)
The full list of the pairwise events between all species is in Table 1. The tree's inner nodes show proposed intermediate states in the operon's evolution. The numbers in the brackets are a 3-tuple showing the cumulative count of events going from the leaf nodes to the tree root: [deletions, duplications, splits]. The way these ancestral states are determined is elaborated below.

Table 1. All pairwise events for the orthoblocks shown in Figure 1.

	A	**B**	**C**	**D**	**E**
A					
B	split, deletion				
C	deletion	split			
D	duplication, deletion, split	duplication, 2× split,	duplication, split		
E	duplication, deletion	duplication, split	duplication	split	

Choosing species

The species tree for each clade was built using *rpoB* as the species marker. For the study of Gram negatives with *E. coli* as a reference species, we use the group of taxa from [11]. For the study of Gram positives with *B. subtilis* as the reference species, we use the Phylogenetic Diversity Analysis program (PDA) [16, 17] to select 33 equidistant species.

Orthoblocks in Phylogenetic Trees

For each orthoblock studied, we use a phylogenetic species tree T comprising a set of extant species related to either one of our reference taxa. The topology of T is determined using multiple sequence alignment of gene *rpoB* followed by the Neighbor Joining algorithm as described in [15]. Each leaf node v in T contains the orthologs to the genes in an operon in the reference species (*E. coli* or *B. subtilis*). For any two genes a and b, if the chromosomal distance is less than 500 bp, the genes will be written as ab. If the distance is greater than 500 bp, they are written with the separator character '$|$' thus: $a|b$. For a species tree T, we define the following:

1. $V(T)$: the set of nodes of T.
2. $E(T)$: the set of edges of T.
3. $L(T)$: set of leaf nodes of T.
4. $I(T)$: set of inner nodes of T.
5. For node $v \in V(T)$, let O be the gene block assigned to v, we define:
   (a) $v.gene[g]$: the set that represents the choice of including of gene g in O. There are only 3 possible cases.
      i. $v.gene[g] = \{1\}$: this means that gene g has to be in O.
      ii. $v.gene[g] = \{0\}$: this means that gene g can not be in O.
      iii. $v.gene[g] = \{0, 1\}$: this means that gene g can either be in O or not in v.
   (b) $I_g(v)$: the identity function of gene g in O. It can only takes value of 0 for not appearing in O or 1 otherwise.
   (c) $v.dup[g]$: the set that represents the duplication status of gene g in O. There are only 3 possible cases.
      i. $v.dup[g] = \{1\}$: this means that gene g has to be duplicated in O.
      ii. $v.dup[g] = \{0\}$: this means that gene g can not be duplicated in O.
      iii. $v.dup[g] = \{0, 1\}$: this means that gene g can either be duplicated or not in O.
   (d) $Gene(O)$: the set of gene of O.
   (e) $Dup(O)$: the set of gene that is duplicated in O.
   (f) $HasLeaf(v)$: the set of leaf nodes that can be reached from node v in postorder traversal.
   (g) $FREQ_g(v)$: frequency of gene g in gene blocks assigned to each leaf nodes in $HasLeaf(v)$.
   (h) $DUP_g(v)$: frequency of duplications of gene g in gene blocks assigned to each leaf nodes in $HasLeaf(v)$.

Orthoblock Distance Functions

The distance between any two homologous gene blocks O, O' are defined as follows:

1. *Split distance* (d_s) is the absolute difference in the number of relevant gene blocks between the two taxa. Relevant gene blocks between two taxa can be computed by only including the genes that appear in both taxa. We define $Rel(O, O')$ as relevant gene blocks of O to O' and formalize the split distance as:

$$d_s(O, O') := ||Rel(O, O')| - |Rel(O', O)||$$

   Example: for the reference gene block with genes (abcdefg), genome A has blocks $O := ((ab), (def))$ and genome B has $O' := ((abc), (de), (fg))$. We then compute the relevant gene blocks $Rel(O, O') = ((ab), (def))$ and $Rel(O', O) = ((ab), (de), (f))$ (removing genes c, g). Therefore, $d_s(O, O') = |2 - 3| = 1$.

2. *Duplication distance (d_u)* is the pairwise count of duplications between two gene blocks. We define $Dif(O, O')$ as the set of duplicated genes of gene block O, so that these genes also appear in O' but are not duplicated in O'. We formalize the duplication distance as:

$$d_u(O, O') := |Dif(O, O')| + |Dif(O', O)|$$

Example: For a reference gene block $(abcde)$, genome A has gene block $O = ((abd))$ and genome B has gene block $O' = ((abbcc))$, respectively. The ortholog of gene O_b is duplicated in genome B, creating a duplication distance $d_u(O, O')$ of 1. However, since gene c does not exist in O, it has no bearing on the duplication distance between the homologous gene blocks O and O'. We then compute $Dif(O, O') = \emptyset$ and $Dif(O', O) = \{b\}$. Therefore, $d_u(O, O') = 0 + 1 = 1$

3. *Deletion distance (d_d)* is the difference in the number of orthologs that are in the homologous gene blocks of the genome of one organism, or the other, but not in both. In short, it is the symmetric difference between the set of orthologous genes of the two gene blocks O, O'. We formalize the deletion distance as:

$$d_d(O, O') := |Gene(O) \triangle Gene(O')|$$

In addition, the deletion distance can also be defined using the identity function:

$$d_d(O, O') := |\sum_g (I_g(O) - I_g(O'))|$$

Example: For a reference gene block $(abcde)$, genome A has gene block $O = ((abd))$ and genome B has gene block $O' = ((abce))$, respectively. Since there are only genes a, b that appear in both genomes, $d_d(O, O') = |\{a, b, d\} \triangle \{a, b, c, e\}| = |\{d\}| + |\{c, e\}| = 3$

The duplication distance and split distance depend on the deletion distance. Intuitively, the duplication of a gene g in gene block O requires such gene appearing in O. Split distance depends on the relevant gene blocks from two taxa. Hence, it depends on the genes that appear in both taxa. Therefore, the split distance and the duplication distance depend on the deletion distance. Using the three distance functions above, we define the total distance between any two homologous gene blocks O, O' as:

$$d(O, O') := d_d(O, O') + d_u(O, O') + d_s(O, O')$$

Problem Definition

Let T be a tree, and G be the set of genes in a reference operon. We define Ω as the set of all possible orthoblocks over gene set G. Let $\lambda : L \mapsto \Omega$ be the labeling of L (assign orthoblocks from Ω to the leaf nodes of T, this can include empty orthoblocks). We define the function $\hat{\lambda} : V \mapsto \Omega$ to be an extension of λ on T if it coincides with λ on the leaves of T (assign an orthoblock to each node of T). If $\hat{\lambda}(v) = O$, we say that vertex v is labelled with orthoblock O. Furthermore, given orthoblock O, we define $GeneBlock(O)$ as the set of gene blocks in O. Given a labelling $\hat{\lambda}$ and an edge $(u, v) \in E$, we define the distance between the two labellings of the endpoints u, v as $d(u, v) := d(\hat{\lambda}(u), \hat{\lambda}(v))$ and the total distance function as $d(\hat{\lambda}) := \sum_{(u,v)\in E} d(u, v)$.

The Maximum Parsimony problem is now defined as follows: given a tree T, an operon gene set G, the orthoblock set Ω and a leaf labeling λ, find a labeling $\hat{\lambda}$ that minimizes $d(\hat{\lambda})$

Here we explore two related Maximum Parsimony heuristic approaches, local and global, to reconstruct ancestral gene blocks.

Local Maximum Parsimony

Briefly, the local approach focuses on finding the optimal parent ancestral gene block given its child gene blocks. For each internal node u, let u_1 and u_2 be its 2 direct children. We present a greedy local optimization algorithm. See Figure 2.

Input: T, G, Ω, λ
Result: $\hat{\lambda}$
for *internal node u when traversing T in post-order* **do**
  Let u_1, u_2 to be u's children
  Let $O_1 := \lambda(u_1)$, $O_2 := \lambda(u_2)$
  $initial := GeneBlock(O_1) \cup GeneBlock(O_2)$
  $initial_{gene} := \{g | FREQ_g(u) > .5\}$
  Remove genes in *initial* that is not included in $initial_{gene}$
  Remove gene blocks in *initial* that is a subset of another gene block in *initial*
  Let $U_{1_G} := set()$
  for *gene block $b \in GeneBlock(O_1)$* **do**
    for *gene g in b* **do**
      if $g \notin initial_{gene}$ **then**
        Remove gene g from b
      end
    end
    $U_{1_G} = U_{1_G} \cup b$;
  end
  if $|initial| < |U_{1_G}|$ *or* $|initial| > |U_{1_G}|$ **then**
    $initial := U_{1_G}$
  end
  for *gene block $b \in initial$* **do**
    if *b has a duplication of gene g and $DUP_s(u) \leq .5$* **then**
      Remove the duplicated gene of g from b
    end
  end
  $\hat{\lambda}(u) := initial$
end
Return $\hat{\lambda}$

Algorithm 1: Local cost function minimization for reconstructing ancestral nodes

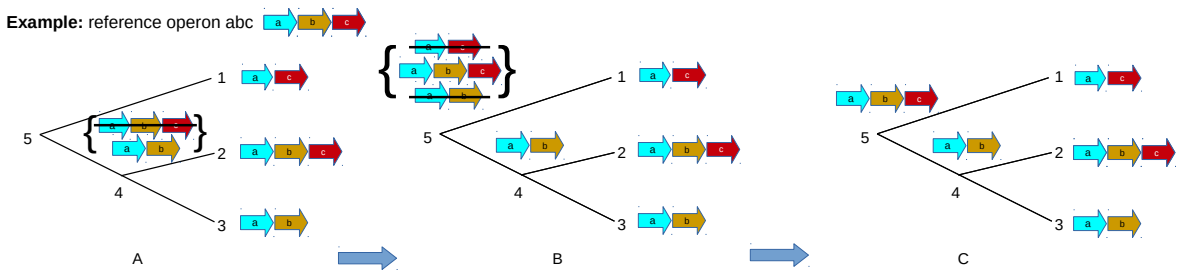

Fig 2. A simplified example of ancestral reconstruction using local maximum parsimony.

Consider a tree with structure as in panels A, B, and C. In each panel, **1,2,3** are the extant nodes that are assigned with gene blocks **ac, abc, ab** respectively, and **4,5** are the inner nodes. The local algorithm traverses the tree bottom-up.

In panel A, node **4**, there are 2 best candidates for the gene block reconstruction. However, we chose to not include gene **c** since $FREQ_c(\mathbf{4}) = .5$. Hence, node **4** is assigned with gene block **ab**

In panel B, at node **5**, there are three best candidates. We chose to assign gene block **abc** to node **5** instead of the other two. The reason is that $FREQ_b(\mathbf{5}) = FREQ_c(\mathbf{5}) = 2/3$ and $FREQ_a(\mathbf{5}) = 1$ which are greater than .5 threshold.

Global Maximum Parsimony

In **Orthoblock Distance Functions**, we determined that the split distance and duplication distance depend on the deletion distance. While finding the global minimum for each separate distance is simple, this dependency makes finding the global minimum of the aggregate of the three distances challenging. In the following example, we demonstrate the minimization of the deletion distance, and then of the split distance. After that, we provide an optimal solution that minimizes the aggregate sum of the two distances.

Given an inner node v and its two child nodes v_1 and v_2, let O be the gene block to be assigned to v. Consider the orthoblocks O_1 and O_2 of v_1 and v_2 respectively as:

$$O_1 : ab|cd|ef|g|k$$
$$O_2 : bc|de|fb|f|fo$$

We define the set of genes that appear in both O_1 and O_2 as $S = \{b, c, d, e, f\}$, and the union gene set of O_1 and O_2 as $G = \{a, b, c, d, e, f, g, k, o\}$. Any gene $i \in S$ will contribute a deletion distance of 2 to $d_d(O, O_1) + d_d(O, O_2)$ if O does not contain gene i. Any gene $i \in G$ but $i \notin S$ will contribute a deletion distance of 1 to $d_d(O, O_1) + d_d(O, O_2)$ if O either has it or not. Hence, including all genes from S in O gives us deletion distance: $d_d(O, O_1) + d_d(O, O_2) = 4$, which is the minimum deletion distance. On the other hand, if we just want to minimize the split distance, the most naive way is not including any genes in O. Then, $Rel(O, O_1) = Rel(O, O_2) = \emptyset$, therefore $d_s(O, O_1) + d_s(O, O_2) = 0$. However, if we choose to do it this way, our deletion distance becomes large ($d_d(O, O_1) + d_d(O, O_2) = 10$). Apparently, decreasing split distance might increase deletion distance and vice versa.

If we focus on minimizing the deletion distance, then $Gene(O) = S$, which means that O has to include all genes in S. Then, the relevant gene blocks of O_1, O_2 to O respectively become:

$$Rel(O_1,O) : b|cd|ef$$
$$Rel(O_2,O) : bc|de|fb|f|f$$

Apparently, the split distance of O_1, O_2 is $d_s(O_1, O_2) = |5 - 3| = 2$. If we remove gene f from *Gene(O)*, the relevant gene blocks of the two children to u become:

$$Rel(O_1,O) : b|cd|e$$
$$Rel(O_2,O) : bc|de|b$$

Hence, by setting our gene block O as either $Rel(O_1, O)$ or $Rel(O_2, O)$, the deletion distance increased by 2 because we excluded gene f which is in S; however, the split distance also decreased by 2. Therefore, the new deletion distance is $d_d(O, O_1) + d_d(O, O_2) = 6$, and the new split distance is $d_s(O, O_1) + d_s(O, O_2) = 0$.

Consider another possibility, if we include gene g in $Gene(O)$, this will not increase the deletion distance. The relevant gene blocks of the two children to u become:

$$Rel(O_1,O) : b|cd|ef|g$$
$$Rel(O_2,O) : bc|de|fb|f|f$$

By setting $O := b|cd|ef|g$, the new split distance is $d_s(O, O_1) + d_s(O, O_2) = 0 + 1 = 1$ and the deletion distance is $d_d(O, O_1) + d_d(O, O_2) = 4$. Therefore, we achieve a lower aggregate sum of deletion and split distances (5 compared to 6). We can keep on adding, or removing genes that only appear in one taxon. This process requires iterations through all the subsets of the symmetrical difference set $Gene(O_1) \triangle Gene(O_2)$, which will take exponential time. We therefore provide a heuristic approach that guarantees minimum deletion and duplication distances, but not split distances. See also Figure 3.

```
Input: T, G, Ω, λ
Result: λ̂
for gene g ∈ G do
    for l ∈ Leaf(T) do
        if gene g ∈ Gene(λ(l)) then
            l.gene[g] = {1}
        else
            l.gene[g] = {0}
        end
    end
    if gene g ∈ Dup(l) then
        l.dup[g] = {1}
    else
        l.dup[g] = {0}
    end
end
for internal node u when traversing T in post-order do
    Let u_1, u_2 be children of u
    for gene g ∈ G do
        if u_1.gene[g] == u_2.gene[g] then
            u.gene[g] = u_1.gene[g]
        else
            u.gene[g] = {0, 1}
        end
        if u_1.dup[g] == u_2.dup[g] then
            u.dup[g] = u_1.dup[g]
        else
            u.dup[g] = {0, 1}
        end
    end
end
for inner node u ∈ V(T) do
    for gene g ∈ G do
        if 1 ∈ u.gene[g] then
            Gene(u).add(g)
        end
        if 1 ∈ u.dup[g] then
            Dup(u).add(g)
        end
    end
end
for internal node u when traversing T in post-order do
    Let u_1 be a u's child, O_1 := λ(u_1)
    Let U_{1_G} := set()
    for gene block b ∈ GeneBlock(O_1) do
        for gene g in b do
            if g ∉ initial_gene then
                Remove gene g from b
            end
        end
        U_{1_G} = U_{1_G} ∪ b;
    end
    λ̂(u) := U_{1_G} for gene g ∈ Gene(u) do
        if g ∉ Dup(u) then
            Remove the duplicated of g from λ̂(u)
        end
    end
end
Return λ̂;
```

Algorithm 2: Global approach

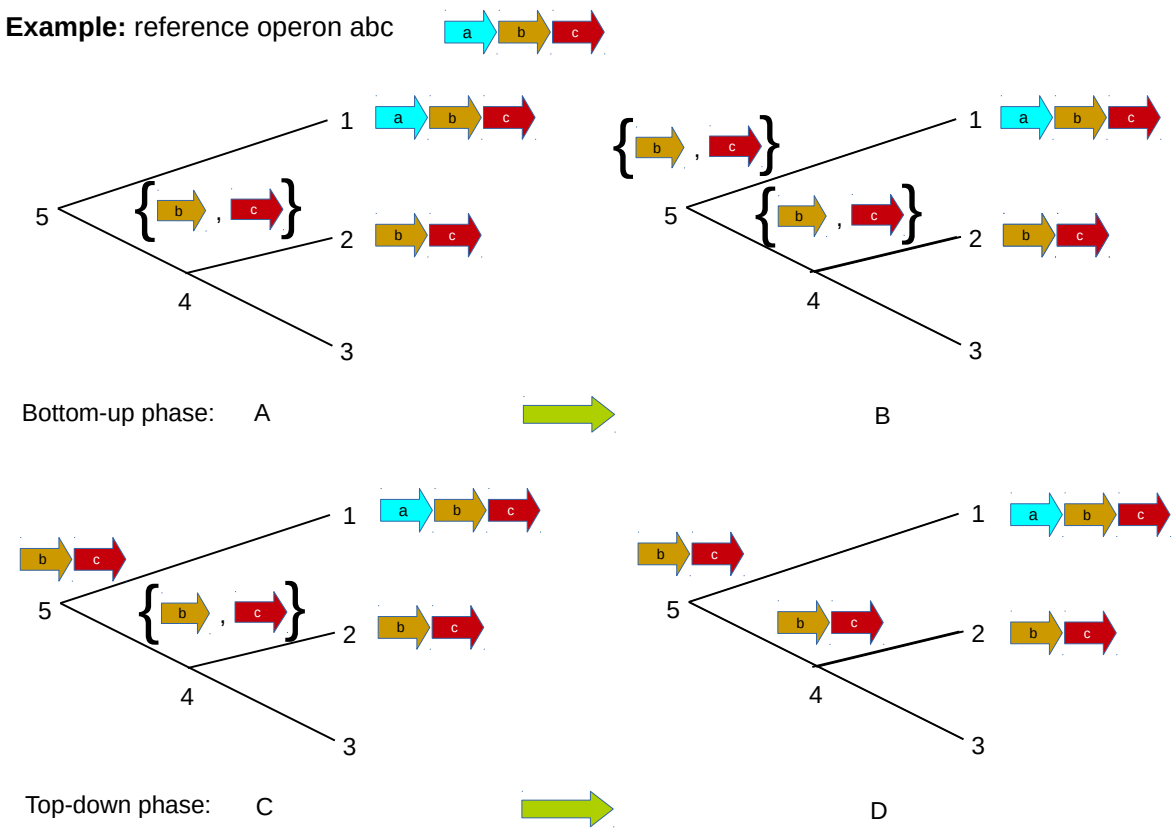

Fig 3. A simplified example of ancestral reconstruction using the global algorithm.

In each panel, **1,2,3** are the extant nodes that are assigned with gene blocks **abc, bc,**$\emptyset$ respectively. The global algorithm traverses the tree bottom-up and top-down.

In bottom up phase, the algorithm constructs the set of genes for the inner nodes (**4,5**). In panel A, at node **4**, the set of genes is {**b,c**} since **2,3** do not share any common genes. In panel B, at node **5**, the set of genes is {**b,c**} because **1,4** share genes **b,c**.

In the top-down phase, the gene block is constructed for each inner node. In panel B, the gene block **bc** is assigned to node **5** using the set of genes of node **5** and the gene block of node **1**. We assign gene block **bc** to node **4** because of its set of genes and gene block in node **2**.

Results and Discussion

We used *E. coli* and *B. subtilis* genomes as gold standards for deriving operons from Gram negative and Gram positive bacteria, respectively. The reason we chose these two species is that they both have well-annotated

genomes, including experimentally verified and functionally annotated operons. We applied our method to groups of Gram-negative and Gram-negative bacteria, using the operons experimentally identified in *E. coli* K-12 and *B. subtilis* str. 168 for the each group, respectively.

Operons from *Escherichia coli*

We chose *E. coli* as a representative of proteobacteria, a major group of Gram-negative bacteria. Our selection resulted in a set of proteobacteria species comprising three ϵ-proteobacteria, six α-proteobacteria, seven β-proteobacteria and 17 γ-proteobacteria, including the reference species *E. coli*. These taxa included two γ-proteobacteria insect endosymbionts: *Buchnera aphidicola* and *Candidatus Blochmania*. These two species have unusually small genomes due to their endosymbiotic nature, and display massive gene loss. We reconstructed ancestors for the following operons from *E. coli* (described below): *atpIBEFHAGD* and *paaABCDEFGHIJK*.

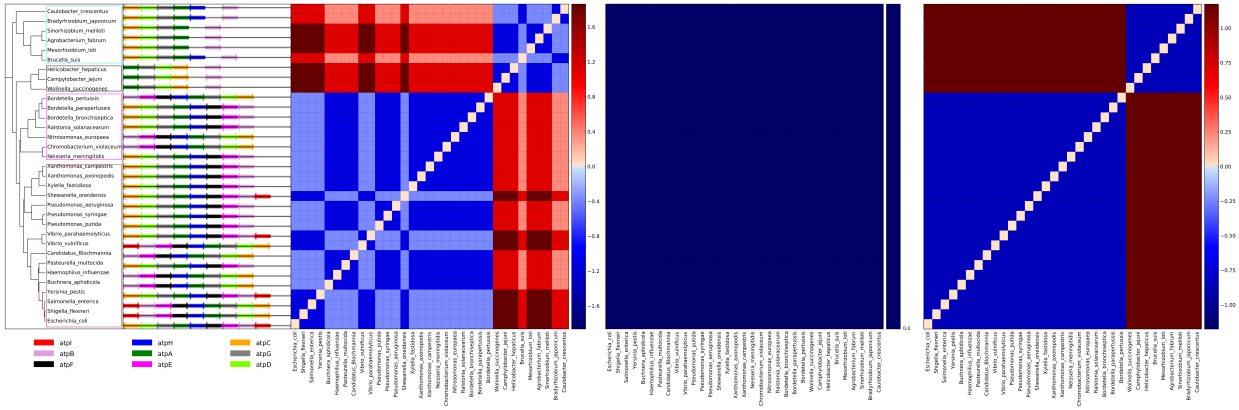

Fig 4. A phylomatrix of gene block *atpIBEFHAGDC*. Each matrix cell depicts the degree of relative conservation of the event between any two species. Blue is more conserved, red is less conserved. Squares, left to right: conservation of deletions, duplications, and splits. The value in the matrix is a z-score, calculated as in [15]. As can be seen, a deletion may have happened only in the common ancestor of the ϵ-proteobacteria), few splits (right square) events, and no duplications events in the pairwise comparison of this gene block, showing a high conservation for all event types. Reproduced from [15] under Creative Common CC-BY-NC 4.0. license. A downloadable higher resolution image is available in `https://github.com/nguyenngochuy91/Ancestral-Blocks-Reconstruction/blob/master/images/Fig4.pdf`

atpIBEFHAGDC. The *atpIBEFHAGDC* operon codes for F_1F_o-ATPase, which catalyzes the synthesis of ATP from ADP and inorganic phosphate [18]. ATP synthase is composed of two fractions: F_1 and F_o [19]. The F_1 fraction contains the catalytic sites and its proteins are coded by five genes (*atpA, atpC, atpD, atpG, atphH*) [19]. The F_o complex constitutes the proton channel and its proteins are coded by three genes *atpF,*

atpE, atpB. atpI is a non-essential regulatory gene. Figure 4 shows the high degree of conservation of this operon.

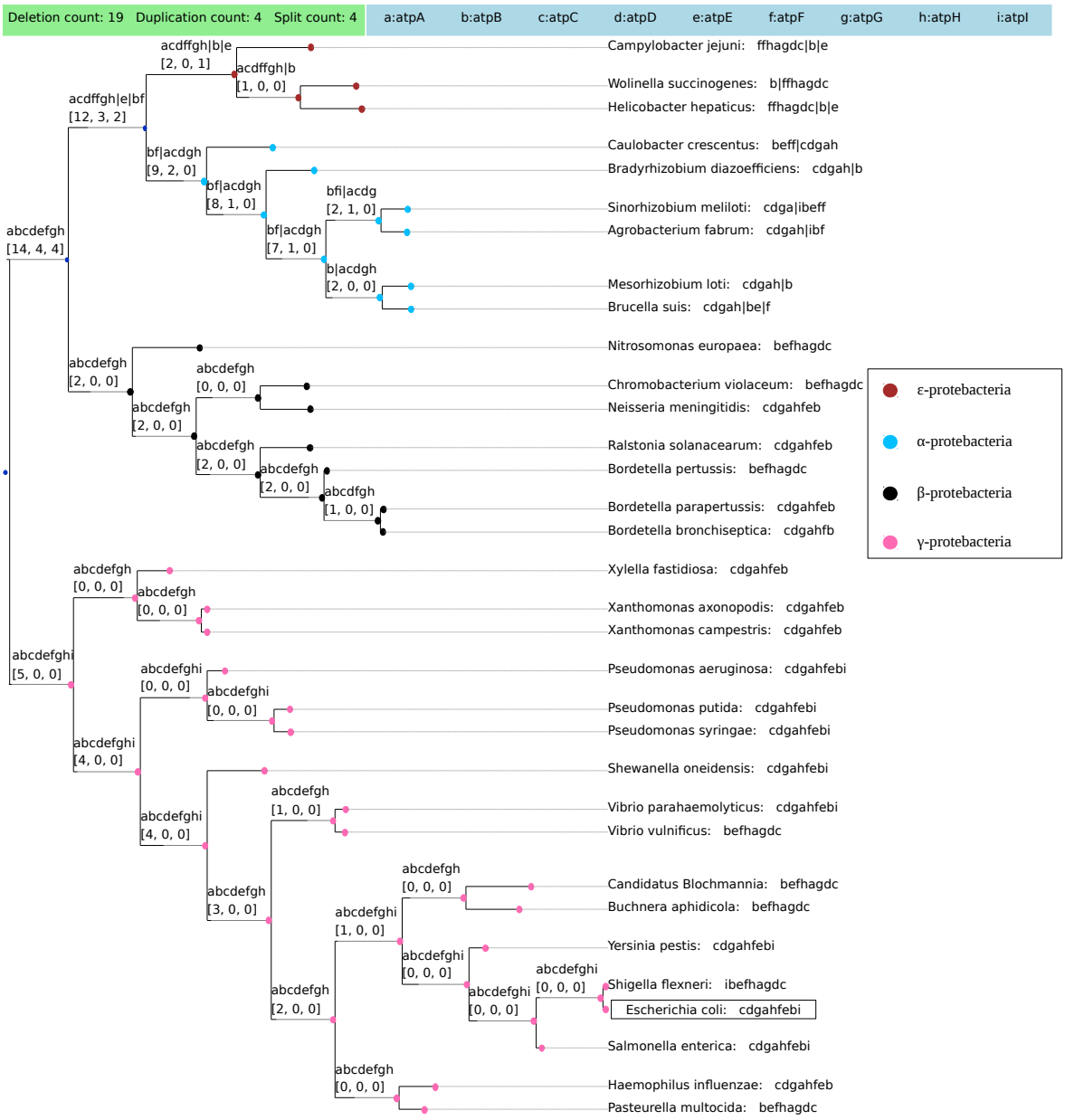

Fig 5. Ancestral reconstruction of operon *atpIBEFHAGDC* using the local optimization approach. The lower-case letters in each tree node represent the genes in the orthoblock (e.g. "a" represents "atpA", see legend in blue bar, top). A '|' designates a split (i.e. a distance ≥500bp between the genes to either side of the '|'). The green bar on top shows the total number of events that took place in this reconstruction. The numbers in the brackets in the inner nodes are a 3-tuple showing the cumulative count of events going from the leafnodes to the tree root in the following order: [deletions, duplications, splits]. These naming and color conventions persist through this paper.

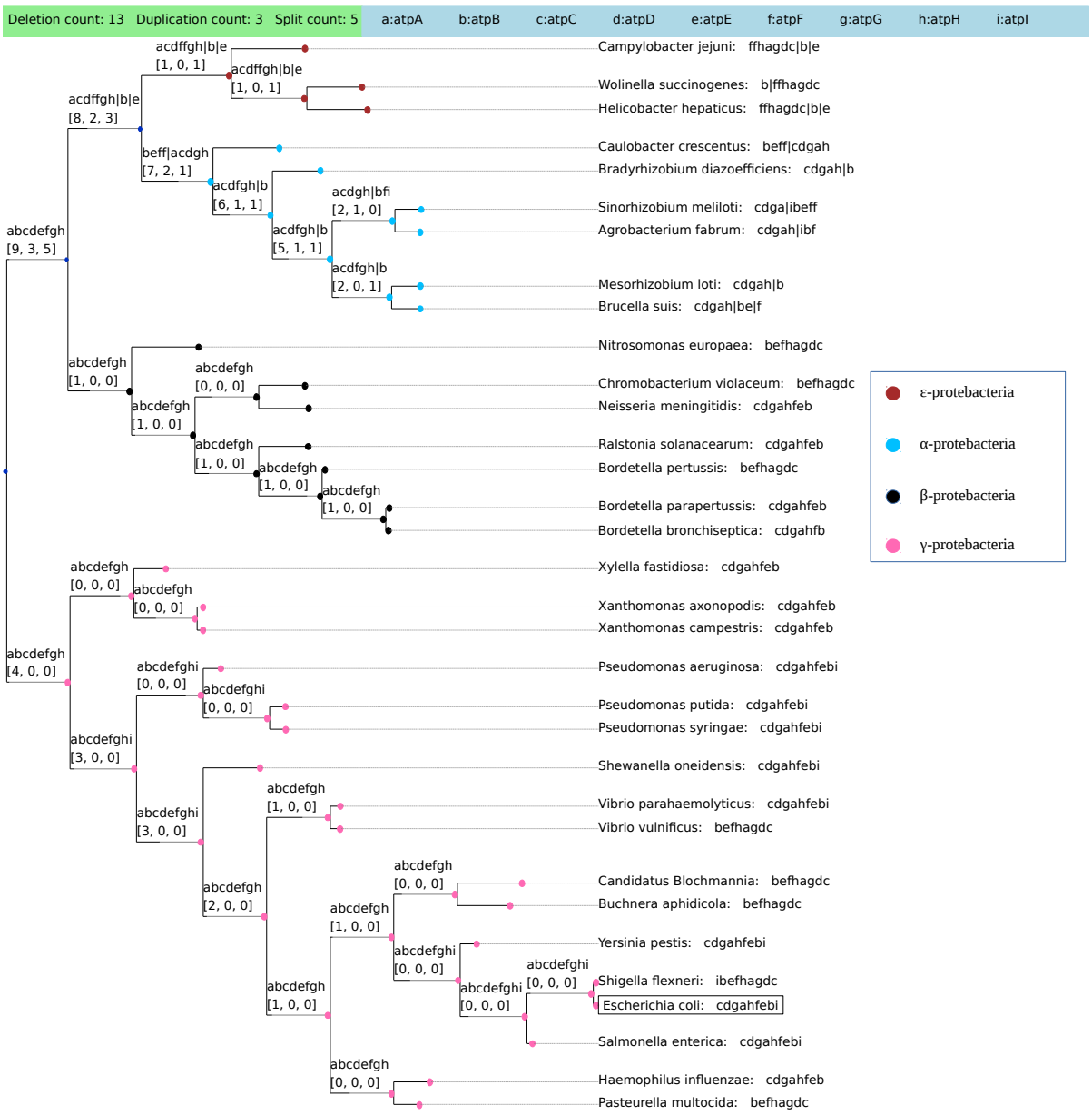

Fig 6. Ancestral reconstruction of operon *atpIBEFHAGDC* using the global optimization approach.

Figures 5 and 6 show ancestral reconstruction using the local and global maximum parsimony algorithms, respectively. Both local and global reconstructions show a consistency of having orthoblocks *atpACDGH* and *atpBF* in the most common ancestors for different Gram negative bacteria. This finding agrees with the long-standing hypothesis that F_o and the F_1 fractions have evolved separately, with the respective fractions having homologs in the hexameric DNA helicases and with flagellar motor complexes. Although we find the

gene *atpI* in several species, the reconstruction predicts that *atpI* is not in the same cluster with other genes. As stated, *atpI* is probably not an essential component of the F_1F_o ATPase [20]. Another interesting finding is the duplication of *atpF* in ϵ-proteobacteria which appears to predate their common ancestor. Note that all genes exist as a gene block even in the endosymbionts *Blochmannia* and *B. aphidicola*.

The ϵ, α, β, and γ -proteobacteria species all have a conserved intact F_1 complex (coded by the *atpACDGH* cluster), which predates their common ancestor. The genes included in the F_o complex in *epsilon*-proteobacteria (gene products *atpB, atpE,atpF*) not in the same cluster as the genes making up F_1. Furthermore, it is unclear whether the gene split that is only found in ϵ-proteobacteria is a split that predates the least common ancestor with the other proteobacteria clades, or whether it is a split introduced in the ϵ-proteobacteria. From the reconstructions provided, the scenario appears to be the latter. Conversely, this observation may also be a result of the small number of species studied here. The species in the ϵ and α-proteobacteria display a known duplication of gene *atpF*. *atpF'* appears as a sister group to *atpF* [21].

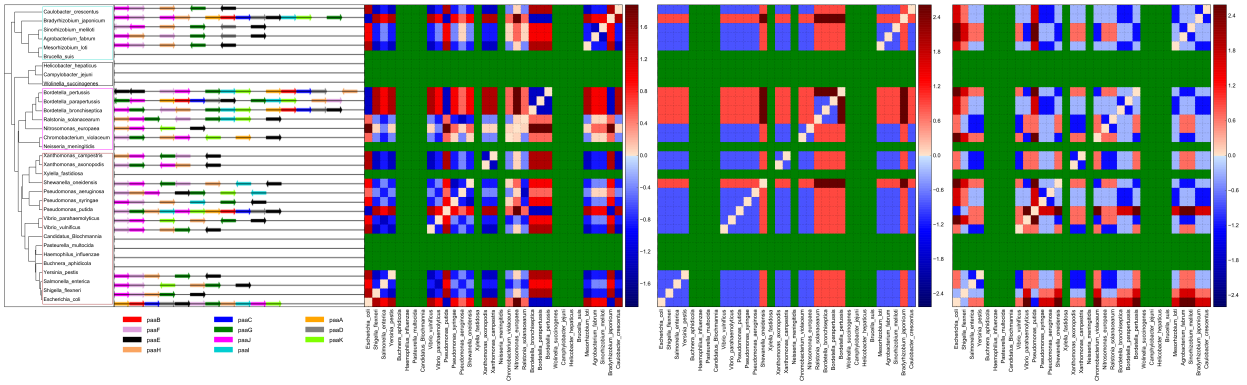

Fig 7. Gene block *paaABCDEFGHIJK* phylomatrices, each show the degree of relative conservation of the event between any two species. Left to right: deletions, duplications, splits. Blue to red scale is high-to-low conservation in z-score [15]. A downloadable higher resolution figure is available in https://github.com/nguyenngochuy91/Ancestral-Blocks-Reconstruction/blob/master/images/Fig7.pdf

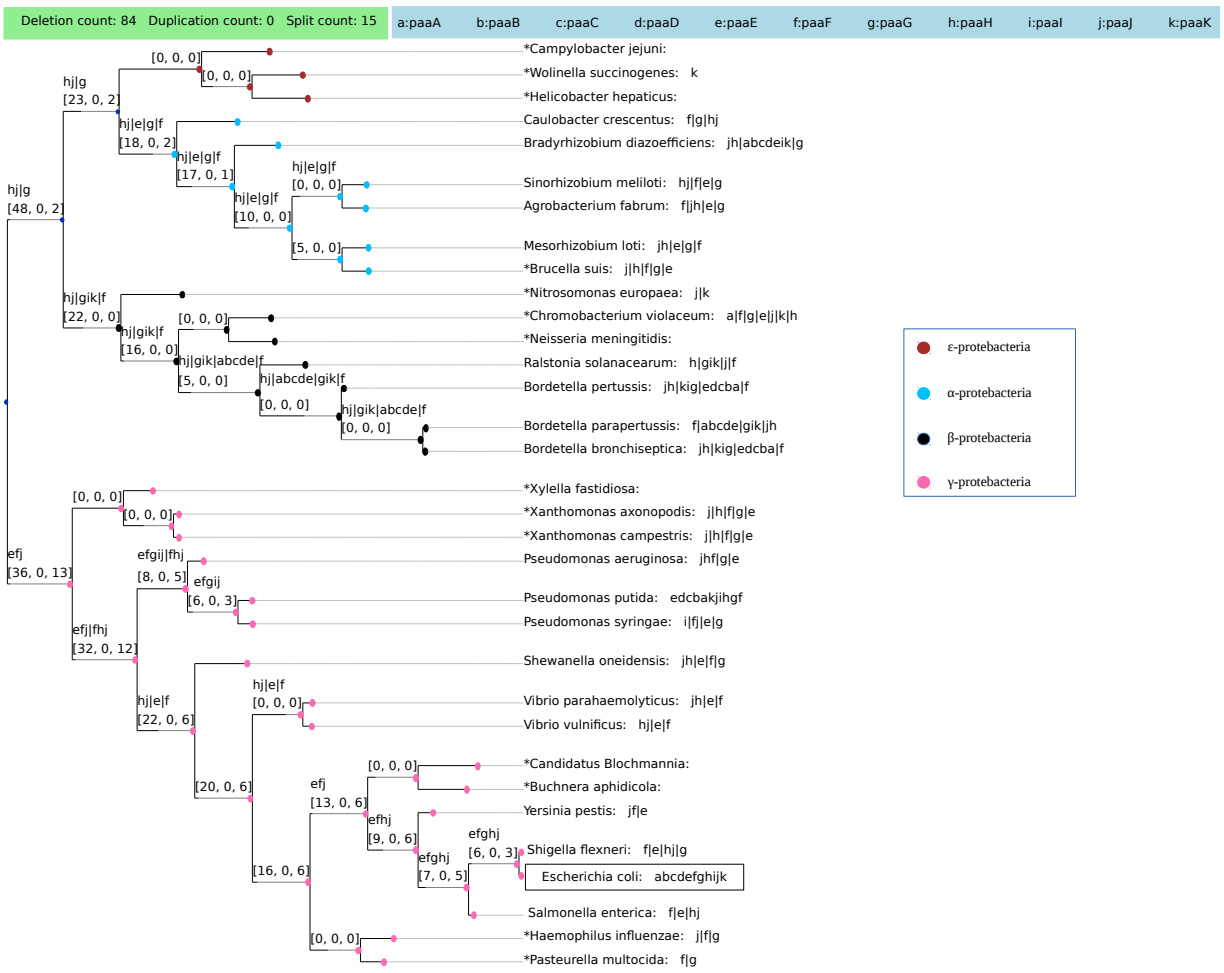

Fig 8. Ancestral gene block reconstruction of *paaABCDEFGHIJK* using the local reconstruction approach. Asterisks in front of species names indicate that a minimal orthoblock (which should consist of two or more proximal genes that are orthologs to genes in the reference operon) was not found.

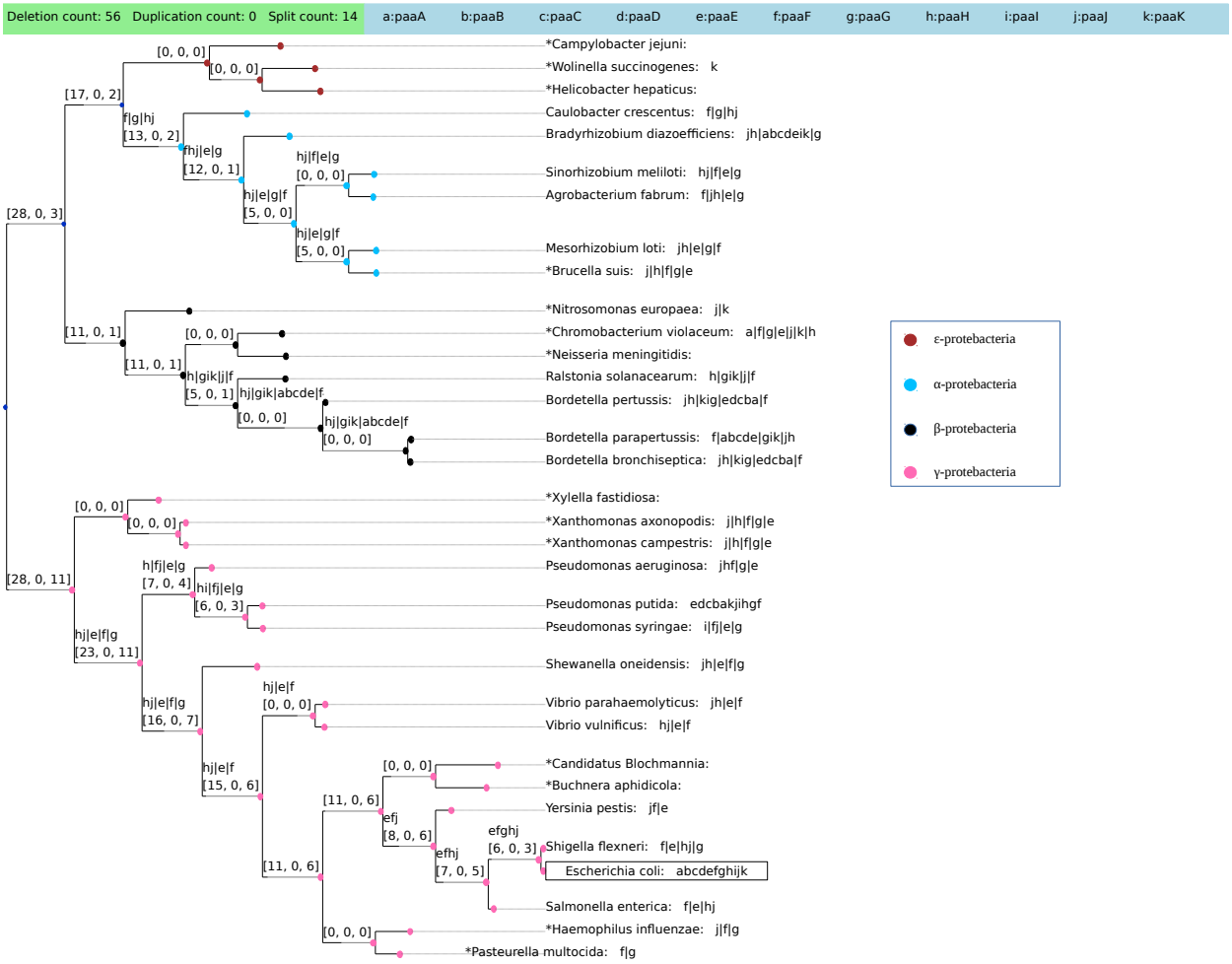

Fig 9. Ancestral gene block reconstruction of operon *paaABCDEFGHIJK* using the global reconstruction approach.

paaABCDEFGHIJK. The operon *paaABCDEFGHIJK* codes for genes involved in the catabolism of phenylacetate [22]. The ability to catabolize phenylacetate varies greatly between proteobacterial species, and even among different *E. coli* K-12 strains. In contrast with *atpABCDEFG* operon which is highly conserved through many species, the operon *paaABCDEFGHIJK* is only found in full complement as an operon in some *E. coli* K-12 strains and some *Pseudomonas putida* strains. It is shown to be much less conserved (Figure 7). However, certain orthoblocks appear to be conserved, providing partial functionality. The orthoblock *paaABCDE* is found in three *Bordetella* species and also in *Bradyrhizobium diazoefficiens*. The products of *paaA*, *paaB*, *paaC* and *paaE* make up the subunits of the 1,2-phenylacetyl-CoA epoxidase, and *paaD* is hypothesized to form an iron-sulfur cluster with the product of *paaE* [23]. We did not find

orthologs in the endosymbionts *B. aphidicola* and *Blochmannia*.

In both the local and global reconstructions (Figure 8 and Figure 9 respectively), only the ancestor of the *Bordetella* species have a combination of *paaABC* complex with *paaE*. It appears that only this combination has full activity [23]. In addition, the global approach only predicts gene blocks for the ancestors of α and most of γ-proteobacteria. Only the common ancestor of the *Bordetella* genus contains the cluster *paaABCE*. It has been confirmed that this cluster of genes is identical to those of *E. coli* [24]. In both approaches, gene *paaF* and *paaG* are not found to be in the same gene blocks, hence the ancestors are most likely missing the hydratase-isomerase complex. The *paaJ* thiolase catalyzes two steps in the phenylacetate catabolism [25–27]. In addition, paaH is the NAD^+-dependent 3-hydroxyadipyl-CoA dehydrogenase involved in phenylacetate catabolism [25]. Therefore, it makes sense that *paaJ* and *paaH* appear in most of the ancestral nodes that have gene blocks.

The results from the study of these operons have provided some interesting and valuable understanding of the evolution of the gene blocks. Specifically, we see the formation of functional intermediate forms both in a highly conserved gene block *atpIBEFHAGDC* and the less conserved gene block based on the operon *paaABCDEFGHIJK*. Also, in both cases, the global approach performs better in term of minimizing events. For brevity, we only provide the global ancestral reconstruction henceforth.

Operons from *Bacillus subtilis*

B. subtilis is a Gram-positive, spore forming bacterium commonly found in soil, and is also a normal gut commensal in humans. It is a model organism for Gram-positive spore forming bacteria, and as such its genome of about 4,450 genes is well annotated. Here we used ROAGUE to reconstruct the ancestors of two *B. subtilis* gene blocks across 33 species. We selected species from the order *Bacillales* using PDA. Species from the following families were selected: *Bacillaceae* (including the reference organism *B. subtilis*), *Staphylococcae*: macrococcus and staphylococcus, *Alicyclobacillaceae*, *Listeriaceae* and *Planococcaceae*.

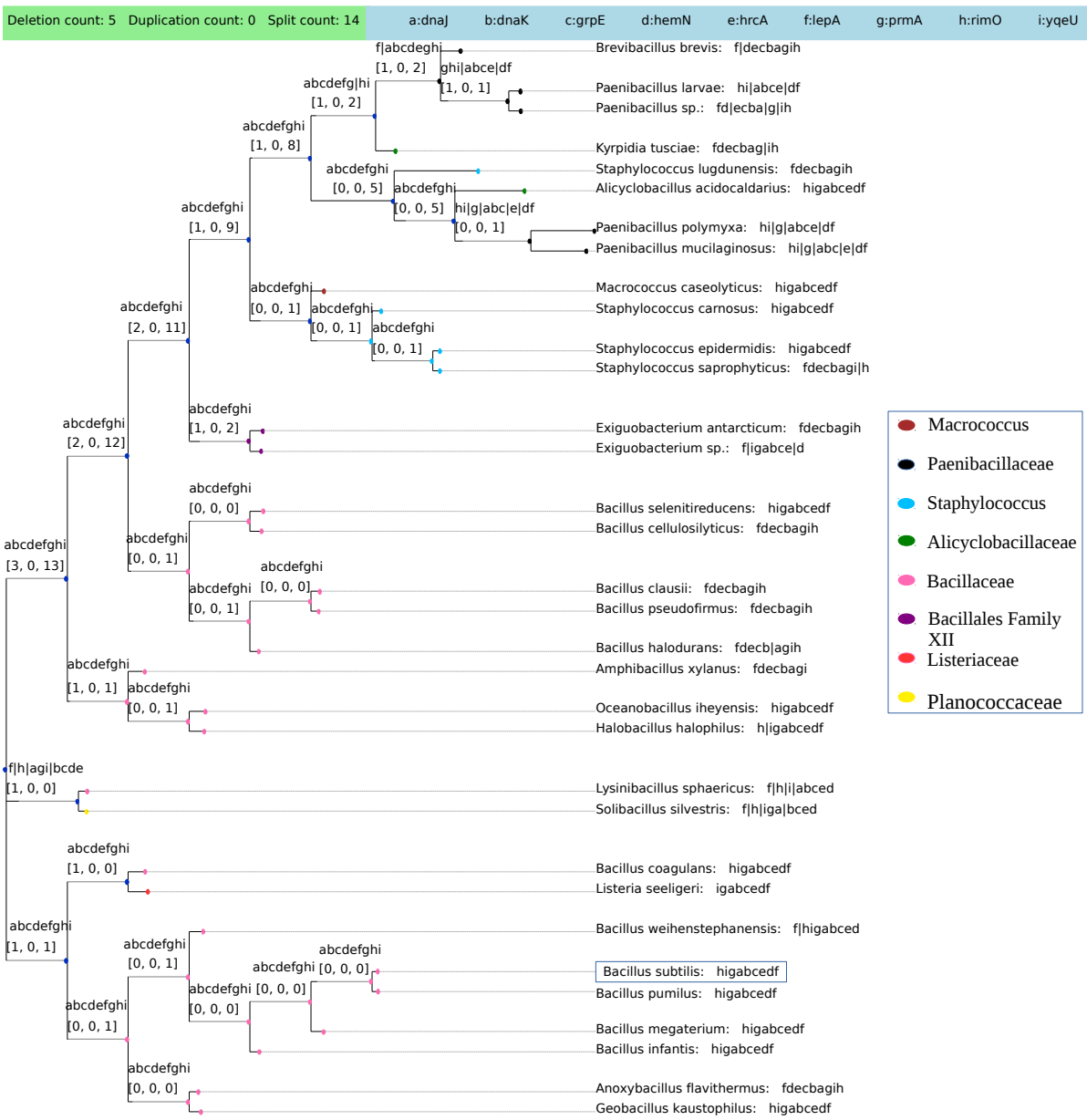

Fig 10. Ancestral reconstruction of *lepA-hemN-hrcA-grpE-dnaKJ-yqeTUV*.

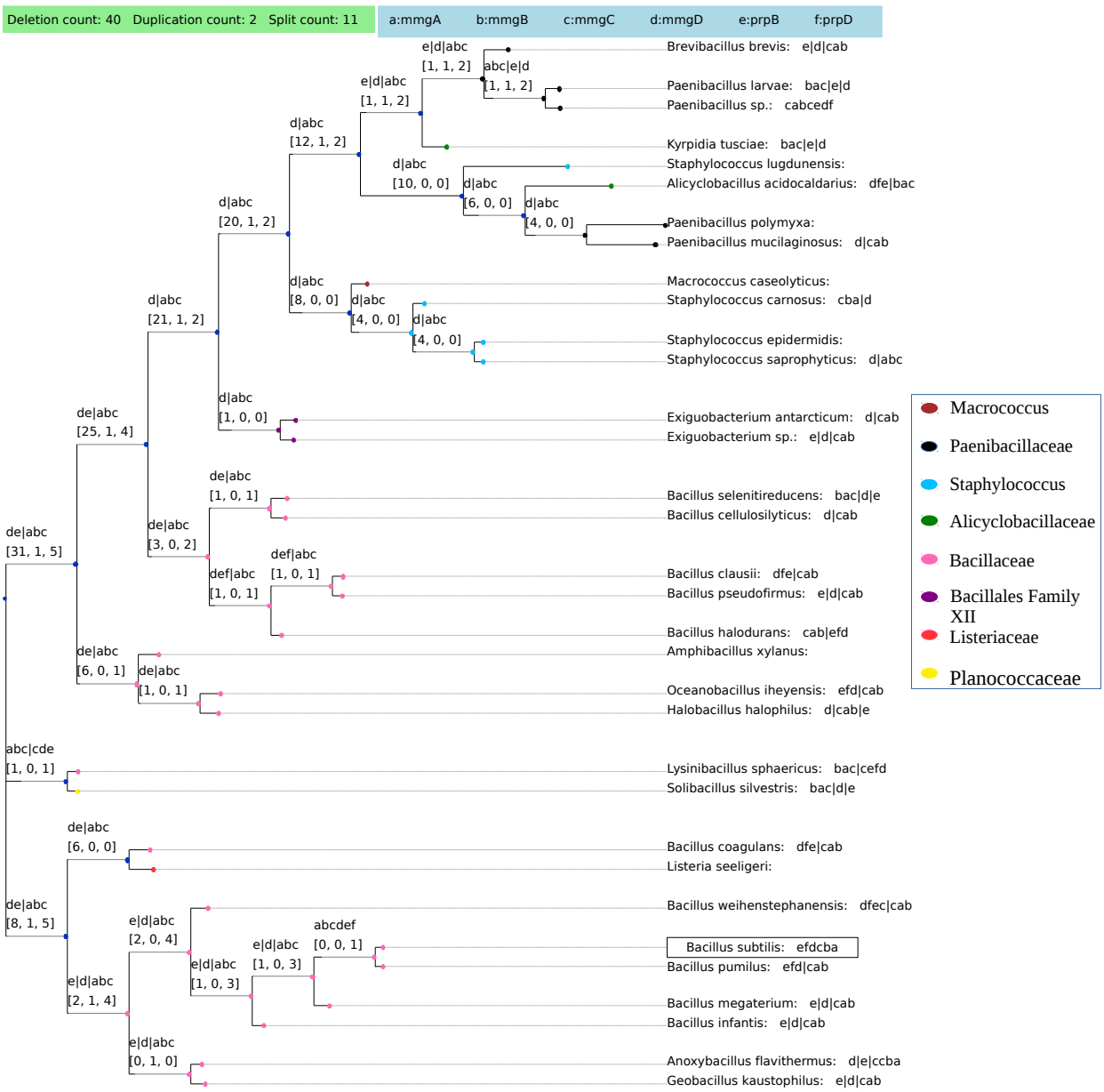

Fig 11. Ancestral reconstruction of *mmgABCDE-yqiQ*.

lepA-hemN-hrcA-grpE-dnaK-dnaJ-prmA-yqeU-rimO. Gene block *lepA-hemN-hrcA-grpE-dnaK-dnaJ-prmA-yqeU-rimO* facilitates the heat shock response in *B. subtilis* and the gene block *hrcA-grpE-dnaK-dnaJ* was the first identified heat shock operon within *Bacillus spp* [28]. The four genes *hrcA, grpE, dnaK, dnaJ* (e,c,b,a in Figure 10) form a tetracistronic structure, which is essential to the heat shock response role [29]. The four genes are proximal in all the species examined, and form the core of the orthoblock.

Overall, this operon is quite conserved, and the ancestral reconstructions are highly similar to the reference operon.

mmgABCDE-prpB. The operon *mmgABCDE-prpB* is expressed during endosporulation [30]. Subunit *mmgABC*'s breakdown of fatty acids is a mean for attaining energy to drive the cell's preparation for dormancy [31]. Hence, it is reasonable to see that the common ancestor has this subunit. In addition, gene *mmgD* and gene *prpB/yqiQ* are predicted to be proximal. Several studies predicted that gene *mmgD*, *prpB*, and *prpD* encode the proteins of the putative methylcitrate shunt [32]. However, they did not specify if deletion mutations might contribute to a defect of the functionality. See Figure 11.

Conclusions

The evolution of operons offer us a tractable model for the evolution of complexity. Understanding how simple units of genes may converge into an operon can lead us to a better understanding of complexity in general. We developed a method for the reconstruction of ancestral gene blocks using maximum parsimony. Using this method we provide several examples of ancestral gene block reconstructions based on reference operons in *E. coli* and *B. subtilis*. A few interesting observations emerge regarding conservation and ancestry of operons. It appears that essentiality (the trait of being essential to life) and the formation of a protein complex are the main drivers for gene block conservation. This is most apparent in the *atp* operon coding for F_1F_o-atpase in proteobacteria. There are few evolutionary events identified in the *atp* operon ancestry. The ribose transporter block also seems to preserve the core ribose transporter (*rbsABC*), while not the ribose phosphorylation genes *rbsD* and *rbsK*. ROAGUE also highlights intermediate functional forms of the orthoblocks, as we see in the pattern of conservation in *paaABCDEFGHIJK*.

Our study does not account for horizontal gene transfer, which is a major driver in operon evolution [9, 14]. Detecting horizontal gene transfer is typically done by looking for conservation of genes and gene structures between distant OTUs, and for anomalous codon usage [33]. Our method opens up a new way of HGT detection, by reconciling a species tree with an operon tree, in the same way that phylogenomic analyses do for gene trees and species trees [34], which would be an interesting future development for this study. In addition, the gene order in a gene block is ignored. While the relationship between gene organization and expression in operons is not well understood, it is clear from several studies that gene order does have an effect on expression and on the functionality of the operon in general (e.g. [35–37]). Adding the parameters of horizontal gene transfer, gene order preservation, or both to ROAGUE would be highly valuable. We invite the community to contribute to ROAGUE, as well as use the tool for identifying orthologous gene blocks,

and reconstructing their ancestry.

Acknowledgments

This work was funded, in part, by a National Science Foundation ABI-1551363 grant awarded to IF, and National Science Foundation grant ECC-1617626 awarded to OE.

References

1. Wagner GP, Altenberg L. Perspective: Complex Adaptations and the Evolution of Evolvability. Evolution. 1996;50(3):967–976. doi:10.1111/j.1558-5646.1996.tb02339.x.

2. Bonner JT. The Evolution of Complexity by Means of Natural Selection. Princeton University Press; 1988. Available from: `https://press.princeton.edu/titles/4307.html`.

3. Pál C, Papp B. Evolution of complex adaptations in molecular systems. Nature ecology & evolution. 2017;1(8):1084–1092.

4. Adami C, Ofria C, Collier TC. Evolution of biological complexity. Proceedings of the National Academy of Sciences. 2000;97(9):4463–4468. doi:10.1073/pnas.97.9.4463.

5. Lynch M, Conery JS. The Origins of Genome Complexity. Science. 2003;302(5649):1401–1404. doi:10.1126/science.1089370.

6. Koonin EV, Dolja VV. Evolution of complexity in the viral world: the dawn of a new vision. Virus research. 2006;117(1):1–4. doi:10.1016/j.virusres.2006.01.018.

7. Horowitz NH. On the Evolution of Biochemical Syntheses. Proceedings of the National Academy of Sciences of the United States of America. 1945;31(6):153–157.

8. Stahl FW, Murray NE. The evolution of gene clusters and genetic circularity in microorganisms. Genetics. 1966;53(3):569–576.

9. Omelchenko M, Makarova K, Wolf Y, Rogozin I, Koonin E. Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. Genome Biology. 2003;4(9):R55+. doi:10.1186/gb-2003-4-9-r55.

10. Lawrence JG, Roth JR. Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics. 1996;143(4):1843–1860.

11. Fani R, Brilli M, Liò P. The origin and evolution of operons: the piecewise building of the proteobacterial histidine operon. Journal of molecular evolution. 2005;60(3):378–390. doi:10.1007/s00239-004-0198-1.

12. Price MN, Arkin AP, Alm EJ. The Life-Cycle of Operons. PLOS Genetics. 2006;2(6):e96+. doi:10.1371/journal.pgen.0020096.

13. Alm E, Huang K, Arkin A. The Evolution of Two-Component Systems in Bacteria Reveals Different Strategies for Niche Adaptation. PLOS Computational Biology. 2006;2(11):e143+. doi:10.1371/journal.pcbi.0020143.

14. Koonin EV. Evolution of genome architecture. The international journal of biochemistry & cell biology. 2009;41(2):298–306. doi:10.1016/j.biocel.2008.09.015.

15. Ream DC, Bankapur AR, Friedberg I. An event-driven approach for studying gene block evolution in bacteria. Bioinformatics. 2015;31(13):2075–2083.

16. Chernomor O, Minh BQ, Forest F, Klaere S, Ingram T, Henzinger M, et al. Split diversity in constrained conservation prioritization using integer linear programming. Methods in Ecology and Evolution. 2015;6(1):83–91.

17. Faith DP. Conservation evaluation and phylogenetic diversity. Biological conservation. 1992;61(1):1–10.

18. Kasimoglu E, Park SJ, Malek J, Tseng CP, Gunsalus RP. Transcriptional regulation of the proton-translocating ATPase (atpIBEFHAGDC) operon of Escherichia coli: control by cell growth rate. Journal of bacteriology. 1996;178(19):5563–5567.

19. Senior AE. The Proton-Translocating ATPase of Escherichia Coli. Annual Review of Biophysics and Biophysical Chemistry. 1990;19(1):7–41. doi:10.1146/annurev.bb.19.060190.000255.

20. NJ G. Construction and characterization of an Escherichia coli strain with a uncI mutation. Journal of Bacteriology. 1984;158:820–825.

21. Koumandou VL, Kossida S. Evolution of the F 0 F 1 ATP synthase complex in light of the patchy distribution of different bioenergetic pathways across prokaryotes. PLoS Comput Biol. 2014;10(9):e1003821.

22. Martin FJ, McInerney JO. Recurring cluster and operon assembly for phenylacetate degradation genes. BMC evolutionary biology. 2009;9(1):36.

23. Grishin AM, Ajamian E, Tao L, Zhang L, Menard R, Cygler M. Structural and Functional Studies of the Escherichia coli Phenylacetyl-CoA Monooxygenase Complex. Journal of Biological Chemistry. 2011;286(12):10735–10743. doi:10.1074/jbc.M110.194423.

24. Luengo JM, Garcia JL, Olivera ER. The phenylacetyl-CoA catabolon: a complex catabolic unit with broad biotechnological applications. Molecular microbiology. 2001;39(6):1434–1442.

25. Ismail W, El-Said Mohamed M, Wanner BL, Datsenko KA, Eisenreich W, Rohdich F, et al. Functional genomics by NMR spectroscopy. European Journal of Biochemistry. 2003;270(14):3047–3054. doi:10.1046/j.1432-1033.2003.03683.x.

26. Teufel R, Mascaraque V, Ismail W, Voss M, Perera J, Eisenreich W, et al. Bacterial phenylalanine and phenylacetate catabolic pathway revealed. Proceedings of the National Academy of Sciences. 2010;107(32):14390–14395. doi:10.1073/pnas.1005399107.

27. Nogales J, Macchi R, Franchi F, Barzaghi D, Fernandez C, Garcia JL, et al. Characterization of the last step of the aerobic phenylacetic acid degradation pathway. Microbiology. 2007;153(2):357–365.

28. Wetzstein M, Völker U, Dedio J, Löbau S, Zuber U, Schiesswohl M, et al. Cloning, sequencing, and molecular analysis of the dnaK locus from Bacillus subtilis. Journal of Bacteriology. 1992;174(10):3300–3310.

29. Homuth G, Masuda S, Mogk A, Kobayashi Y, Schumann W. The dnaK operon of Bacillus subtilis is heptacistronic. Journal of bacteriology. 1997;179(4):1153–1164.

30. Acharya R. Overexpression, purification, and characterization of MmgD from Bacillus subtilis strain 168. University of North Carolina at Greensboro; 2009.

31. Quattlebaum AL. Characterization of Biosynthetic and Catabolic Pathways of Bacillus Subtilis Strain 168. University of North Carolina at Greensboro; 2009.

32. Voigt B, Hoi LT, Jürgen B, Albrecht D, Ehrenreich A, Veith B, et al. The glucose and nitrogen starvation response of Bacillus licheniformis. Proteomics. 2007;7(3):413–423.

33. Koonin EV, Makarova KS, Aravind L. Horizontal gene transfer in prokaryotes: quantification and classification. Annual review of microbiology. 2001;55(1):709–742. doi:10.1146/annurev.micro.55.1.709.

34. Eisen JA. A phylogenomic study of the MutS family of proteins. Nucleic acids research. 1998;26(18):4291–4300.

35. Hiroe A, Tsuge K, Nomura CT, Itaya M, Tsuge T. Rearrangement of Gene Order in the phaCAB Operon Leads to Effective Production of Ultrahigh-Molecular-Weight Poly[(R)-3-Hydroxybutyrate] in Genetically Engineered Escherichia coli. Applied and Environmental Microbiology. 2012;78(9):3177–3184. doi:10.1128/aem.07715-11.

36. Wells JN, Bergendahl LT, Marsh JA. Operon Gene Order Is Optimized for Ordered Protein Complex Assembly. Cell reports. 2016;14(4):679–685.

37. Lim HN, Lee Y, Hussein R. Fundamental relationship between operon organization and gene expression. Proceedings of the National Academy of Sciences. 2011;108(26):10626–10631. doi:10.1073/pnas.1105692108.