Evolution of tRNA Repertoires in *Bacillus* Inferred with OrthoAlign

Olivier Tremblay-Savard,[†,1] Billel Benzaid,[†,1] B. Franz Lang,[2] and Nadia El-Mabrouk[*,1]

[1]Département d'informatique et de recherche opérationnelle (DIRO), Université de Montréal, CP 6128 succursale Centre-Ville, Montreal, QC, Canada

[2]Département de biochimie, Université de Montréal, CP 6128 succursale Centre-Ville, Montreal, QC, Canada

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: mabrouk@iro.umontreal.ca.

Associate editor: Howard Ochman

Abstract

OrthoAlign, an algorithm for the gene order alignment problem (alignment of orthologs), accounting for most genome-wide evolutionary events such as duplications, losses, rearrangements, and substitutions, was presented. OrthoAlign was used in a phylogenetic framework to infer the evolution of transfer RNA repertoires of 50 fully sequenced bacteria in the *Bacillus* genus. A prevalence of gene duplications and losses over rearrangement events was observed. The average rate of duplications inferred in *Bacillus* was 24 times lower than the one reported in *Escherichia coli*, whereas the average rates of losses and inversions were both 12 times lower. These rates were extremely low, suggesting a strong selective pressure acting on tRNA gene repertoires in *Bacillus*. An exhaustive analysis of the type, location, distribution, and length of evolutionary events was provided, together with ancestral configurations. OrthoAlign can be downloaded at: http://www.iro.umontreal.ca/~mabrouk/.

Key words: **alignment, gene order, orthology, rearrangements, duplications, phylogeny, transfer RNA, *Bacillus*.**

Introduction

Transfer RNAs are among the most important and ancient genes. Accordingly, a large number of studies have addressed tRNA gene identification and structure prediction. Yet, little is known about the evolution of tRNA genes in terms of number, organization, and functional specificity, on a genome-wide scale. A first comprehensive analysis of the genomic organization of tRNAs in Eukarya has been conducted in 2010 (Bermudez-Santana et al. 2010), revealing an extensive variability of organization among lineages, which is in striking contrast to the high level of tRNA gene sequence conservation. Other genome-wide studies conducted on specific lineages (Withers et al. 2006; Rogers et al. 2010) also revealed a rapid evolution of tRNA gene families through duplication and loss. Having a clear picture of tRNA repertoire evolution is expected to shed light on many questions such as the link between tRNA copy number and protein synthesis, the evolution of the genetic code, and tRNA functional reassignment (Saks et al. 1998; Lavrov and Lang 2005; Wang and Lavrov 2011; Ling et al. 2014).

The question we address can be formulated as follows. We are given a set of genomes annotated for tRNAs, and a species tree for the corresponding taxa. We want to infer tRNA gene content and order information of ancestral genomes identified with each of the internal nodes of the tree, together with an evolutionary scenario leading to the observed genome organization. This problem is known in the comparative genomics literature as the "small phylogenetic problem," which has been widely studied for various restrictions on genome structure and models of evolution, most of them being difficult and developed heuristics being time-consuming (Pe'er and Shamir 1998; Sankoff and Blanchette 1998; Ma et al. 2007; Alekseyev and Pevzner 2009; Ouangraoua et al. 2011; Zheng and Sankoff 2011; El-Mabrouk and Sankoff 2012; Gagnon et al. 2012; Jones et al. 2012; Zheng and Sankoff 2012). Focusing on a cherry (two neighboring leaves) of the species tree, the problem reduces to the one of comparing two genomes, namely the "two species small phylogenetic problem," which has also been extensively studied (El-Mabrouk 2005; Fertin et al. 2009; El-Mabrouk and Sankoff 2012), and has been proved difficult (NP-hard) for most problem variants. But of much more concern to biologists than details about optimality and efficiency, is the nonuniqueness of solutions. When compared genomes have sufficiently diverged so that corresponding gene orders are almost unrelated, an exponential number of evolutionary scenarios will be inferred for a given optimality criteria, leading to a huge and therefore noninformative number of equally likely ancestral predictions. These observations highlight the need for choosing an appropriate monophyletic group with a sufficient number of available completely sequenced genomes, that have diverged on a long enough timescale to reflect the diversity of evolutionary events. At the same time, genomes must reflect some conservation allowing to eliminate sources of nonuniqueness in the construction. When only few events separate two neighboring genomes, these can be assumed to be nonoverlapping (each gene involved in at most one event) and thus still

Article

"visible" in extant species. Based on these criteria, the bacterium *Bacillus* is a perfect model organism for studying the evolution of tRNA repertoires, as we have access to 50 fully sequenced genomes, and the tRNA gene orders are sufficiently conserved for the hypothesis of nonoverlapping events to apply.

Bacillus species are Gram-positive, rod-shaped bacteria. They can be aerobic or facultative anaerobic and they produce endospores that are normally resistant to heat, radiation, and disinfectants. *Bacillus anthracis* (anthrax) and *B. cereus* (food poisoning) are pathogens for humans, whereas *B. thuringiensis* is a pathogen for insects used as biological control. Some *Bacillus* species are used industrially to produce enzymes and antibiotics. *Bacillus amyloliquefaciens*, for example, is used to produce the well-known BamH1 restriction enzyme.

Recently, considering an evolutionary model restricted to duplications and losses, we reformulated the comparison of two gene orders as an alignment problem: find an alignment minimizing a given cost (Holloway et al. 2013). Interestingly, such an alignment can directly be translated into an evolutionary scenario of visible events leading to a unique ancestral genome. Although alignments are a priori simpler to handle than rearrangements, this problem has been shown NP-hard (Andreotti et al. 2013; Benzaid et al. 2013; Dondi and El-Mabrouk 2013). In this article, we present OrthoAlign, a new polynomial-time heuristic for this problem, and use it in a phylogenetic framework. It extends preliminary versions (Benzaid et al. 2013; Holloway et al. 2013) by including rearrangements and by relaxing some visibility constraints.

Applying OrthoAlign to 50 completely sequenced *Bacillus* strains (see the Materials and Methods section for the complete list of genomes), average rates of duplications, losses, and inversions are found to be extremely low compared with those reported for *Escherichia coli* (Withers et al. 2006). This suggests a striking conservation and a strong selective pressure acting on tRNA gene repertoires in *Bacillus*. Evolutionary rates, rearrangement types, effect on the operon structure and ancestral configurations are thoroughly explored in the Results section. Finally, our results are validated by testing OrthoAlign on simulated data sets.

The Evolutionary Model

The considered genomes are unichromosomal and circular-mapping that can be linearized at the origin of replication. A "genome" G is represented as a string of signed characters, where each character represents a gene family. For example, isoacceptor families can be considered for tRNA genes. Each character α may appear many times in G, all such positions corresponding to genes belonging to the gene family α. The sign of a gene represents the transcriptional orientation of the corresponding gene. Let $X = X_1X_2 \ldots X_n$ be a string. We call the "reverse" of X the string $-X = -X_n \ldots -X_2 -X_1$. We denote by $X[i, i+k]$ the "substring" of X formed by the consecutive genes of the interval $[i, i+k]$.

The Evolutionary Events

The method accounts for most genome-wide evolutionary events acting on a unichromosomal genome X: "duplication" (denoted D), in tandem or transposed, copying a substring $X[i, i+k]$ to a position j outside the interval $[i, i+k]$; 'loss' (L) removing a substring $X[i, i+k]$; "substitution" (S) of a character $X[i]$ to another character; "inversion" (I) transforming a substring $X[i, i+k]$ into its reverse; "inverted duplication" (ID) copying the reverse of a substring $X[i, i+k]$ to a position j outside the interval $[i, i+k]$; "transposition" (T) removing a substring $X[i, i+k]$ and inserting it at another position j, and finally "inverted transposition" (IT) removing a substring $X[i, i+k]$ and inserting its reverse at another position j.

We denote by $\mathcal{O} = \{D, L, S, I, ID, T, IT\}$ the set of all possible operations. As detailed in Appendix A, each event $O \in \mathcal{O}$ can be represented by the "source" $X[i, i+k]$ affected by the event and the "target" $Y[j, j+k]$ which is the result of the event ($Y[j, j+k] = X[i, i+k]$ for a duplication, $\oslash$ for a loss, etc.). Characters of the source and target of O are said to be "covered" by the operation. The "cost" of O is denoted by $c(O(k+1))$, where $k+1$ is the "size" of O, that is, the size of the involved substrings.

The Two Species Small Phylogeny Problem

Given two genomes A and X, an "evolutionary history" $O_{A\to X} = \{O_1(k_1), \ldots, O_l(k_l)\}$ from A to X is a sequence of events from $\mathcal{O}$ (possibly of length 0) transforming A into X. The cost of $O_{A\to X}$ is $C(O_{A\to X}) = \sum_{i=1}^{l} c(O_i(k_i))$.

Now, given two genomes X and Y, the "two species small phylogeny problem" (2-SPP) consists in finding a triplet $(A, O_{A\to X}, O_{A\to Y})$ called a "history" of X and Y, minimizing the cost $C(O_{A\to X}) + C(O_{A\to Y})$, over all possible histories.

Visible Histories

As motivated in the introduction, only "visible histories" are considered, in the sense that the source and target of evolutionary events should be visible in extant genomes.

Definition 1

Let $(A, O_{A\to X}, O_{A\to Y})$ be a history for X and Y. It is a "visible history" if and only if, for every event O of the history, each of the source and the target of O is a substring of at least one of the two genomes X or Y. A "visible ancestor" of X and Y is a genome A belonging to a visible history $(A, O_{A\to X}, O_{A\to Y})$ of X and Y.

The above definition is a relaxation of the one considered in Holloway et al. (2013), allowing for the source and target of an event to belong to two different genomes. In the case of a duplication, this mimics a duplication in one lineage (say X) followed by the loss of the source, which is still present in the other lineage (Y), or alternatively a transposition.

Stated differently, a visible event is an event that can be seen on an 'alignment' of the two genomes X and Y: an alignment is simply a pair $(\overline{X}, \overline{Y})$ of strings obtained by filling X and Y, respectively, with a special symbol "–" called "gap", such that the resulting "aligned genomes" $\overline{X}$ and $\overline{Y}$ are of

equal length, and for each position i, at most one of $\overline{X}_i$, $\overline{Y}_i$ is a gap. A column $(\overline{X}_i, \overline{Y}_i)$ of the alignment is a "gap" if either of $\overline{X}_i$ or $\overline{Y}_i$ is a gap, is a "match" if $\overline{X}_i = \overline{Y}_i$, and is a "mismatch" otherwise. Now a "labeling" of the alignment is a set of operations such that each character of X and Y is covered by at most one event such that: each matched position (i.e., belonging to a match column) is not covered by any operation, each mismatched position (i.e., belonging to a mismatch column) is covered by a substitution or an inversion, and each gapped position is covered by a loss, a duplication, an inverted duplication (ID), a transposition, or an inverted transposition (IT). This definition does not prevent a labeling to contain operations leading to a cyclic interpretation for a set of intervals. A labeling is "feasible" if it does not contain any subset of events inducing a cycle. A formal definition of a cycle is given in Benzaid et al. (2013).

Duplications and losses are asymmetrical operations that are applied explicitly to one of the two strings, whereas substitutions, inversions, and transpositions may be indiscriminately applied to one of the two sequences. But if we set the source genome for all the substitutions and rearrangements (inversions and transpositions), then a labeling leads to a unique common ancestor A for X and Y. The following theorem states that this and the converse is true.

Theorem 2 (Holloway et al. 2013)
Given two genomes X and Y, there is a one-to-one correspondence between feasible labeled alignments of X and Y and visible ancestors of X and Y.

Consequently, given two genomes X and Y, the two species small phylogeny problem (2-SPP) for visible events can be reformulated as the one of finding a labeled alignment of X and Y of minimum cost, where the cost of a labeled alignment is the sum of costs of the underlying operations.

New Approach: OrthoAlign

We first describe a pairwise alignment approach which is a generalization of that described in Benzaid et al. (2013), augmented with substitutions and rearrangements, and a relaxation of the visibility constraint as explained above. The algorithm is quadratic if inversions are not considered, and cubic otherwise (proof not shown, extension of results from Benzaid B, Dondi R, and El-Mabrouk N, submitted). It is a heuristic, not guaranteed to output a history of minimum cost (various versions of the alignment problem described above have been shown to be NP-hard [Andreotti et al. 2013; Benzaid et al. 2013; Dondi and El-Mabrouk 2013]). We then integrate this algorithm in a phylogenetic framework.

The Pairwise Alignment Algorithm

A transposition (respectively an IT) may alternatively be interpreted as a duplication (respectively an ID) followed by the loss of the duplication source. We therefore restrict the dynamic programming computation to the set of operations $\mathcal{O} = \{D, L, S, I, ID\}$. Let $|X| = n$ and $|Y| = m$. Let $C(i, j)$ be the minimum cost of a labeled alignment of two prefixes $X[1, i]$ and $Y[1, j]$ of X and Y. The problem is to compute $C(m, n)$. For all $1 \leq i \leq n$ and $1 \leq j \leq m$, $C(i, j)$ is the minimum of $M(i, j)$, $S(i, j)$, $I(i, j)$, $L_X(i, j)$, $L_Y(i, j)$, $D_X(i, j)$, $D_Y(i, j)$, $ID_X(i, j)$, $ID_Y(i, j)$ reflecting the minimum cost of an alignment $(\overline{X}_i, \overline{Y}_j)$ of $X[1, i]$ and $Y[1, j]$ satisfying, respectively, the constraint that the last characters of $\overline{X}_i$ and $\overline{Y}_j$ represent a match, a substitution or an inversion, or the last character of $\overline{X}_i$ ($\overline{Y}_j$ respectively) are covered by a loss, a duplication, or an ID. Recurrences are given in Appendix B.

After computing all the values leading to $C(m, n)$, a bottom-up approach allows to output a labeled alignment of minimum cost $C(m, n)$. Unfortunately, such labeled alignment is not necessarily a feasible alignment, as it may lead to a cyclic interpretation of duplications and losses (e.g., one duplication with source $X[i, i + k]$ and target $X[j, j + k]$, and one with source $X[j, j + k]$ and target $X[i, i + k]$).

In order to identify cycles, we use a graph representation for duplications. Namely, let $D = \{D_1, \ldots, D_{k \leq (n+m)}\}$ be the set of duplications of the labeled alignment of $\overline{X}$ and $\overline{Y}$. Construct a directed graph $\mathcal{G} = (V, E)$ (called "duplication graph") as follows: for each duplication D_i, add two vertices corresponding to its source s_i and its target t_i, and one directed edge (s_i, t_i). Moreover, for each pair of overlapping strings (s_i, t_j) with $i \neq j$, add a directed edge (t_j, s_i). Any cycle in this graph is a duplication cycle.

We resolve each cycle $\mathcal{D}$ as follows: 1) a cycle of size 4 for which each duplication has a source and target belonging to different genomes is interpreted as a single transposition or IT; 2) any other cycle is resolved by interpreting the shortest overlapping string of $\mathcal{D}$ as a loss rather than a duplication.

The Small Phylogeny Problem

The standard distance-based approach for the small phylogeny problem, known as the "steinerization" approach, is to start with an initial assignment of ancestral nodes, to traverse the phylogeny in a bottom-up approach and to optimize the median over three nodes of the tree (Sankoff and Blanchette 1998; Kovac et al. 2011). In this context, the pairwise alignment algorithm can be used for the initialization step: traverse the tree in a depth-first manner and compute successive ancestors of pairs of nodes. The issue with this approach is its intractability and the fact that it is not guaranteed to converge to a global optimum. Moreover, due to the nonuniqueness of solutions, the accuracy of a fully automated method is hard to guarantee in general. Finally, this approach requires an algorithm for the median problem, which has been shown NP-hard for most rearrangement events (Pe'er and Shamir 1998).

Therefore, we use a more direct application to a phylogeny. We proceed bottom-up in the phylogenetic tree, taking a cherry (X, Y) at a time. We apply the pairwise alignment algorithm and then correct the obtained labeled alignment $(\overline{X}, \overline{Y})$ by considering a neighboring strain W (any leaf of the subtree rooted at the sibling of the cherry (X, Y)), and aligning X (or Y, but assume X w.l.o.g.)

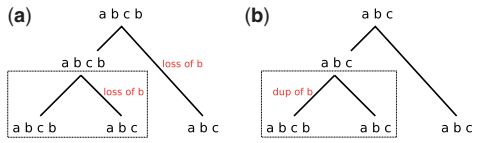

FIG. 1. Two possible evolutionary histories for a phylogeny containing three genomes (leaves). For the first comparison between the two genomes at the bottom of the tree (inside the box), there are two possible scenarios with the same cost: in scenario (*a*), a loss of gene b is inferred whereas in scenario (*b*), a duplication of gene b is inferred. However, when the whole phylogeny is considered, scenario (*b*) is clearly the most parsimonious one (only one event is necessary to explain the phylogeny, instead of two in scenario (*a*)).

with W, leading to $(\overline{X}', \overline{W})$. The correction is based on the following observations: 1) substitutions, inversions, and transpositions are symmetrical operations that can be indiscriminately applied to one or the other of the two sequences in a pairwise alignment, leading to two equally parsimonious scenarios; 2) duplications and deletions are interchangeable in an evolutionary scenario of two genomes. We discriminate between such equally parsimonious scenarios by checking, for the source $X[i, i + k]$ of an event, whether it is a match, or covered by the same event in $(\overline{X}', \overline{W})$ (see fig. 1).

Notice that, if the sibling of a cherry is an internal node, then using different leaves for the correction may lead to different results. Clearly, this practical and simple way of applying the pairwise alignment algorithm in a phylogenic context does not guarantee any optimality result in the case of highly shuffled genomes. However, it appears to perform well in the context of tRNA gene repertoires in *Bacillus* genomes separated by few evolutionary events. In general, our method can be seen as an initialization step for the steinerization approach, and ancestral assignments can then be improved by using a median approach. However, such a median approach remains to be developed for an evolutionary model involving duplications, losses, and rearrangements. Rather, we validate ancestral inferences by exploring the nucleotide sequences of the source and target of the inferred evolutionary events.

Results and Discussion

The Evolutionary History at a Glance

In total, 95 duplications, 141 deletions, 23 inversions, 12 transpositions, and two substitutions were inferred, showing that duplications and losses are prevalent over rearrangement events. Moreover, deletions are more frequent than duplications, probably because duplicates have little if any selective advantage.

As illustrated in figure 2, for both duplications and deletions, short events, typically of size one or two, are largely predominant. An exception is an unusual peak at size nine for deletions, due to the red block (containing nine tRNA genes) being deleted many times in the phylogeny of figure 3. This peak is greatly reduced if we consider the corrected phylogeny proposed in Appendix C.

Such exponential decrease of event number according to size does not hold for rearrangement events. Almost

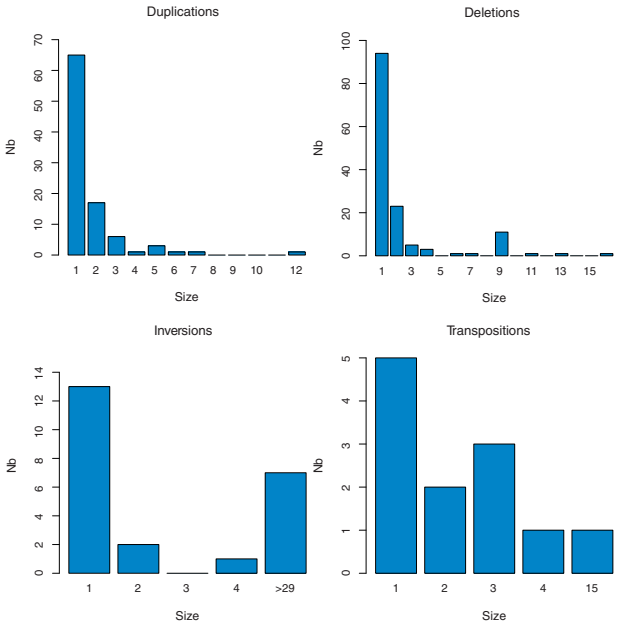

FIG. 2. Size distributions of events inferred on the whole phylogeny.

one-third of the inferred inversions are very large ones affecting 30 tRNA genes or more, and the longest one affects 65 tRNA genes. All these inversions use one of the replication axes as a pivot, which is in agreement with previous studies on rearrangements in bacteria (Tillier and Collins 2000). We represent them as occurring around the terminus of replication. However, due to the circular nature of the bacterial chromosomes, the same inversions (i.e., using the same breakpoints) made around the origin of replication would give the same chromosomes if read in the opposite direction.

As for transpositions, not enough were inferred to be able to see a clear pattern in the size distribution. Only one was relatively large (15 tRNA genes). Note that predicted transpositions could be the result of a series of inversion events.

Evolutionary Rates

In order to compare the average rates of duplications, deletions, and inversions found in *Bacillus* with those in *E. coli* (Withers et al. 2006), the relative divergence time of the most recent common ancestor of the 50 *Bacillus* strains has been measured (details given in the Materials and Methods section). It is found to be roughly 24 times older than that for the *E. coli* ancestor. This result suggesting a very ancient origin for the *Bacillus* strains is quite surprising considering the relatively well-conserved organization of the tRNA genes. Notice that another possible explanation could be that the mutation rate per nucleotide is much higher in *Bacillus* than in *E. coli*. If that was the case, the most recent common ancestor of the 50 *Bacillus* strains would not be as ancient. However, to our knowledge, no evidence for a higher sequence divergence rate in the *Bacillus* genus has been reported in the literature. Consequently, the following results are based on the assumption that the mutation rates are similar for *Bacillus* and *E. coli* genomes.

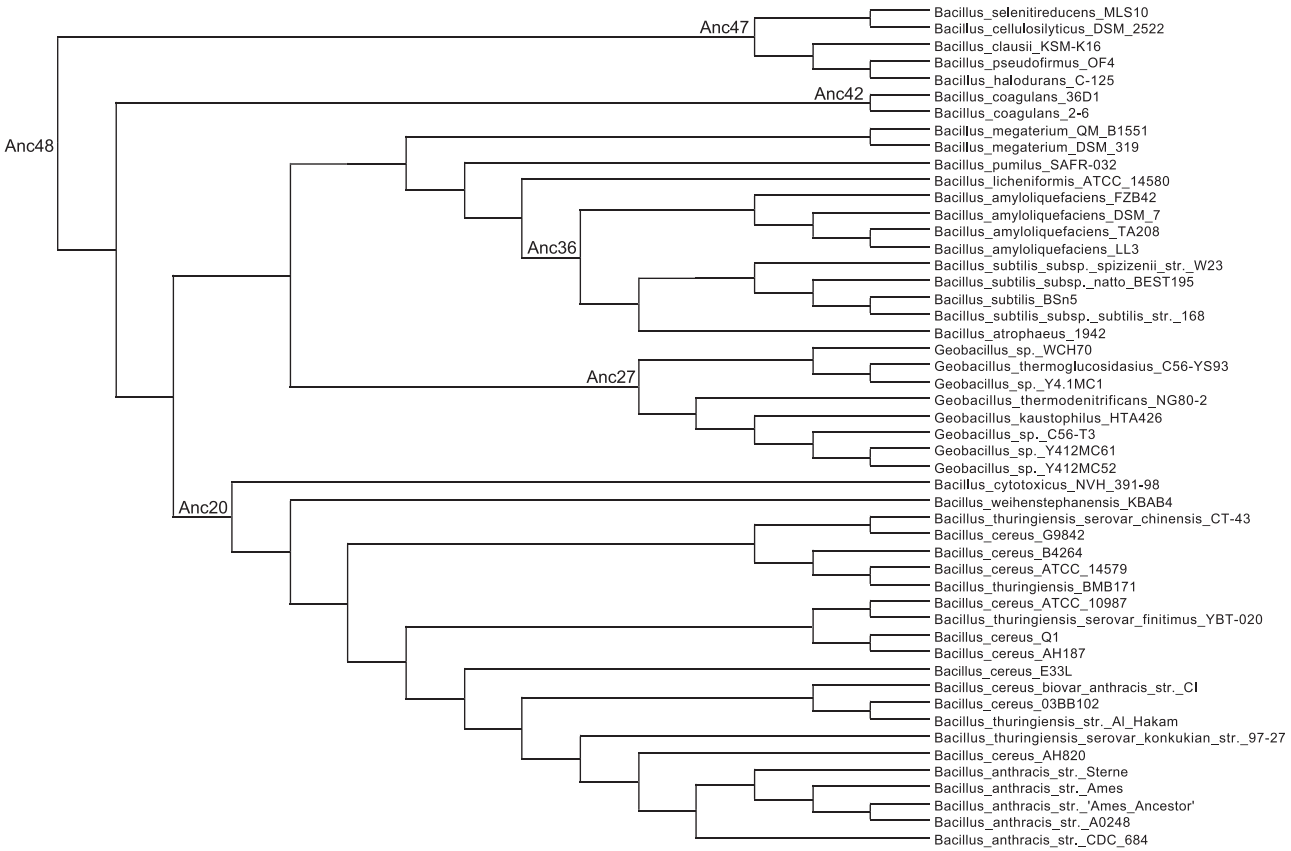

Fig. 3. Phylogenetic tree of the 50 studied *Bacillus* strains redrawn from the Pathosystems Ressource Integration Center (PATRIC) (Gillespie et al. 2011).

We counted the numbers of duplications, losses, and inversions that were inferred on each lineage of the species tree, from the root to the extant genomes. The average number of duplications and deletions on a lineage are, respectively, 12.14 and 11.58, whereas the average number of inversions is only 1.5. Interestingly, even if a larger number of deletions (141) were inferred compared with the number of duplications (95) in the whole species tree, the duplications are predominant on average for each lineage. This is probably due to the fact that deletions were mostly inferred at the bottom of the tree.

Considering the average numbers of duplications, deletions, and inversions on a lineage calculated in the *E. coli* study (Withers et al. 2006) and the relative divergence times, we found that the average rates of the different evolutionary events are much lower in the *Bacillus* genus. The average rate of duplications obtained in *Bacillus* is about 24 times lower than the one reported in *E. coli*, whereas both the average rates of deletions and inversions are about 12 times lower in *Bacillus*. This suggests that the number and organization of the tRNA genes in *Bacillus* genomes is under strong selective pressure and evolve slowly.

Proposing an Absolute Divergence Time

Withers et al. (2006) estimated that the four *E. coli* genomes have diverged 19 Ma. Based on that estimation and on the relative divergence time that we found for the *Bacillus* strains, we obtain an absolute divergence time of 459 Ma for their most recent common ancestor. This gives an average rate of 0.026 per million years (My) for duplications, 0.025 per My for losses and 0.003 per My for inversions. The combined rate of duplications and losses is equal to 0.052 per My. Over an average of 91 genes per species, we find 0.00057 gains and losses per gene per My.

Major Events

In this section we summarize the inferred evolutionary events on the 50 studied *Bacillus* strains. Figure 4 shows a condensed representation of the phylogeny, where we omitted the lineages in which very few events were inferred. The largest inferred evolutionary events are framed by gray boxes. All framed duplications and transpositions were further validated by using whole-genome alignment dot-plots, and verifying that noncoding DNA between the tRNA genes is also highly similar between the source and the target of the event.

Ancestor 20

A large inversion was inferred around the terminus of replication on the branch leading to *B. anthracis* str. CDC 684. A duplication of the green block was also inferred in *B. thuringiensis* serovar chinensis CT-43. In the whole subtree of the *B. anthracis*, *B. cereus*, and *B. thuringiensis* species (not shown entirely in fig. 4), a total of six deletions of one of the red blocks was inferred.

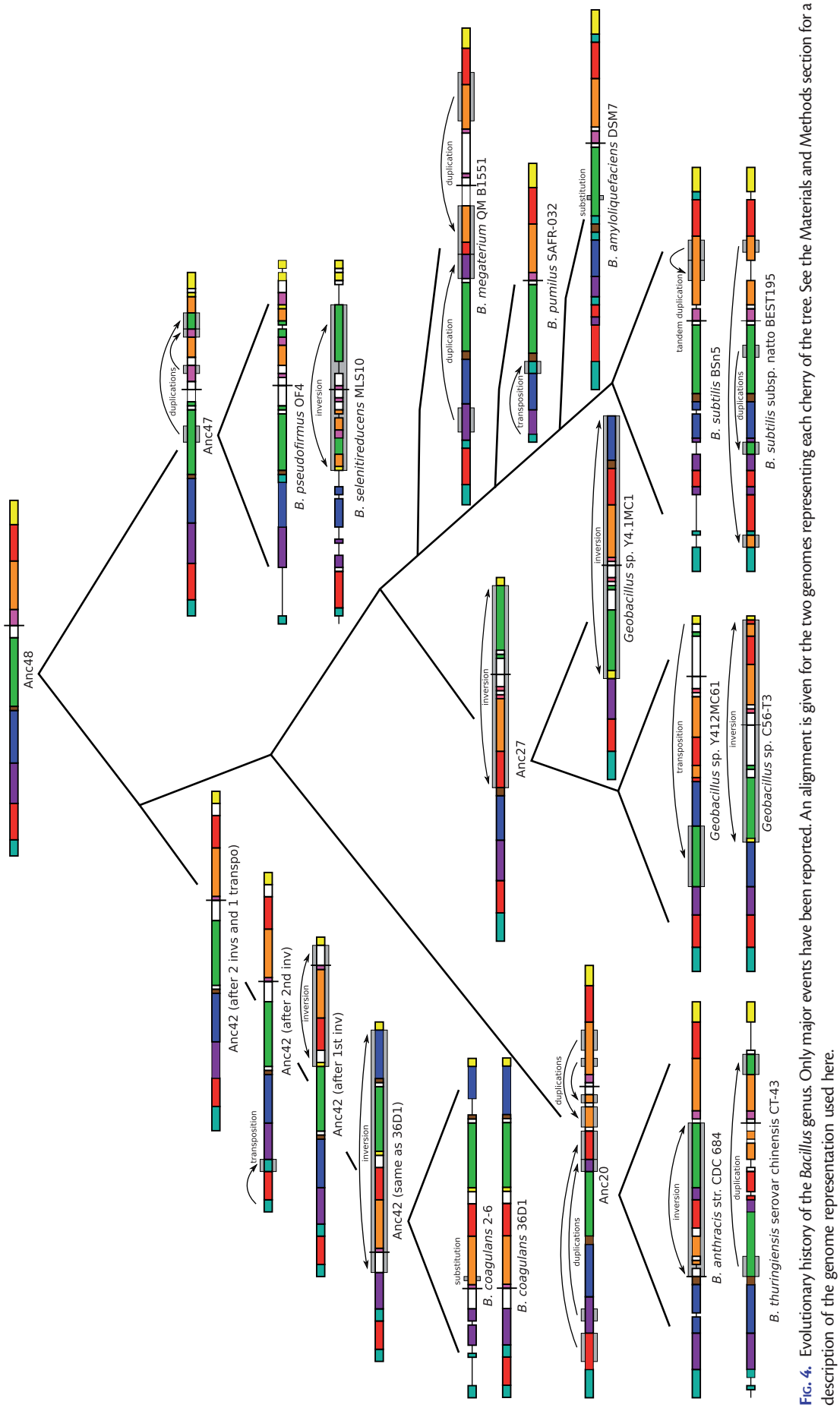

FIG. 4. Evolutionary history of the *Bacillus* genus. Only major events have been reported. An alignment is given for the two genomes representing each cherry of the tree. See the Materials and Methods section for a description of the genome representation used here.

Ancestor 27

Three large rearrangement events have occurred in the *Geobacillus* subtree: a transposition of the green block in *Geobacillus* sp. Y412MC61 and two large inversions around the terminus, one on the branch leading to *Geobacillus* sp. C56-T3 and another on the branch leading to *Geobacillus* sp. Y4.1MC1.

Ancestor 36

In *B. subtilis* BSn5, the orange block has been elongated by a tandem duplication of five genes. Two duplications of size three occurred on the branch leading to *B. subtilis* subsp. natto BEST195, copying part of the orange block inside the turquoise block and part of the green block inside the purple block. Moreover, we observe many duplicated [Ile,Ala] blocks (turquoise) in the *B. amyloliquefaciens* subtree. They have all been inserted inside existing rRNA blocks ([16S, 23S, 5S]), between the 16S and the 23S genes.

Ancestor 47

A large inversion around the terminus of replication and many tRNA gene deletions were inferred on the branch leading to *B. selenitireducens* MLS10. Moreover, this is the only subtree in which we find tRNA-Sec genes. More precisely, *B. selenitireducens* MLS10, *B. cellulosilyticus* DSM 2522 (not shown in fig. 4), and *B. pseudofirmus* OF4 are the only strains of all the ones studied here containing at least one tRNA-Sec gene (note that the tRNA-Sec gene in *B. pseudofirmus* OF4 was not annotated in GenBank but found by a BLAST search and validated with Infernal 1.1 [Nawrocki and Eddy 2013]). We predicted that the tRNA-Sec gene was present in the last common ancestor of the 50 studied strains, which is in line with the hypothesis that selenocysteine utilization is an ancient ability that was lost in many phyla, probably due to limited selenium availability (Copeland 2005; Zhang et al. 2006).

Ancestor 42

As for the *B. coagulans* strains, one transposition and two successive large inversions around the terminus of replication were necessary to explain the observed syntenies.

Ancestor 48

We can see a big duplicated region in Anc20 between the green block and the terminus of replication. This region was created by four separate duplication events: a duplication of part of the purple block (three tRNA genes), a duplication of the end of the red block (seven tRNA genes), and two duplications of parts of the orange block (five and two tRNA genes). A large inversion around the terminus of replication occurred on the branch leading to Anc27. Other interesting events are two duplications that copied part of the pink block and part of the green block inside the orange block of Anc47.

Finally, the most recent common ancestor of the 50 *Bacillus* strains studied (Anc48) is shown at the top of the tree. It contains a total of 88 tRNA genes, which is comparable with the average number of genes found in the strains studied here (91). Sixty-seven of those 88 genes (76%) are present in all the 50 strains. Those 67 "core" tRNA genes represent all the tRNA isoacceptor families except tRNA-Sec.

tRNA Substitutions

We found one gene substitution in the *B. amyloliquefaciens* DSM 7 strain when comparing it with *B. amyloliquefaciens* LL3. This is in fact a tRNA-Met gene that was affected by a few point mutations (only four, as shown at the top of fig. 5), one of them changing the anticodon from CAT to CAC, which is why it was annotated as a tRNA-Val in the DSM 7 strain.

However, it often takes more than a few mutations in a tRNA for it to be recognized and charged by a new tRNA synthetase (Vasileva and Moor 2007). In order to predict if this new tRNA-Val in DSM 7 is really recognized by valyl-tRNA synthetase and charged with valine, we used TFAM (Ardell and Andersson 2006) on its nucleotide sequence and also on the one of the tRNA-Met in LL3 to validate their annotation. Interestingly, both tRNA genes are classified as initiator methionine tRNAs.

In light of this prediction by TFAM, we first hypothesized that the tRNA-Val gene in the DSM 7 strain could in fact be a tRNA-iMet with a CAC anticodon recognizing GTG start codons. Unfortunately, the resequencing of the region containing this new tRNA-Val gene in DSM 7 showed that the four mutations presented in the alignment of figure 5 were in fact the result of sequencing errors in the GenBank sequence of DSM 7.

A second tRNA substitution was detected in *B. coagulans* 2–6 when comparing it with *B. coagulans* 36D1. A tRNA-Ser gene seems to have mutated to a tRNA-Thr gene according to the GenBank annotations. In fact, only three point mutations have occurred and one of them changed the anticodon

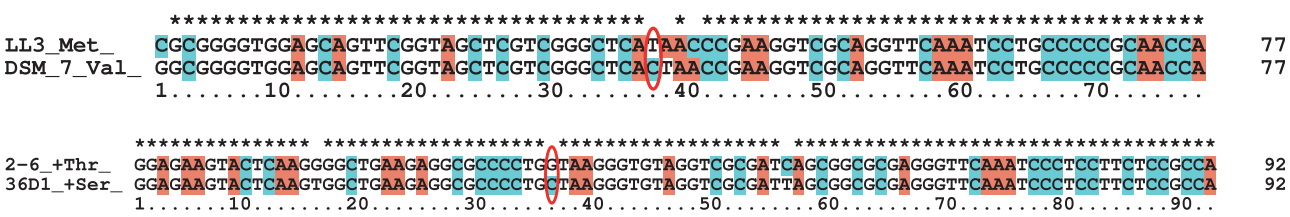

Fig. 5. Top: Sequence alignment of the tRNA-Met gene in *Bacillus amyloliquefaciens* LL3 and the tRNA-Val gene in *B. amyloliquefaciens* DSM 7. The third position of the anticodon is circled in red. Bottom: Sequence alignment of the tRNA-Thr gene in *B. coagulans* 2–6 and the tRNA-Ser gene in *B. coagulans* 36D1. The second position of the anticodon is circled in red.

from GCT to GGT, which explains the new annotation in the 2–6 strain (see bottom of fig. 5).

Once again, we used TFAM to infer the identity class of both tRNA genes. Even with the anticodon change, TFAM predicts that the tRNA gene of interest in strain 2–6 is still a tRNA-Ser (instead of a tRNA-Thr). Is this tRNA still charged with serine or is it charged with threonine? Resequencing and further experimental validation (including identification of nucleotide modifications) will be necessary to answer this question. Serine and threonine have similar side chains and properties, so the insertion of a serine instead of a threonine into a protein might not be an issue. It is also possible that we are dealing with another sequencing error.

Operons

It has been suggested in Candelon et al. (2004) that some tRNA genes downstream rRNA operons (containing 16S, 23S and 5S genes) could be transcribed independently because of a promotor located between the 23S and the 5S genes. Based on the results of this study, we identified the tRNA operons downstream of the rRNA operons in the strains we are studying (see fig. 6 for the locations of the tRNA operons in the genome of *B. cereus* ATCC 14579).

We analyzed the effect of the largest predicted events on tRNA operons. None of the inferred inversions and transpositions has the effect of breaking an operon into two parts. In other words, every operon is either totally inside or totally outside a segment affected by rearrangements. This is not very surprising as operons are relatively small compared with the genome size. It is also potentially deleterious to have an inversion inside an operon switching some of the genes to the opposite strand. Indeed, in most bacterial genomes, the majority of genes tend to be on the leading strand (the strand that is pointing away from the origin of replication; Rocha 2004). Moreover, it has been shown that essential genes (like tRNA genes) are mostly found on the leading strand (Rocha and Danchin 2003; Mao et al. 2012). This strand bias can be explained by the polymerase avoidance model (Brewer 1988; Rocha 2004): avoiding head-on collisions between the DNA and the RNA polymerases allows to have faster DNA replication and fewer transcript losses. Thus, selective pressure is likely to select strongly against the transfer of tRNA genes from the leading strand to the lagging strand. Inversion events occurring around one of the replication axes are more common because they keep the genes on the leading strand.

We also checked the effect of duplications on operons, to see whether duplications can lead to the creation of new operons. Most inferred duplications are actually inserting new tRNA copies inside an existing tRNA operon (the tandem duplication of five genes inside the orange block in *B. subtilis* BSn5 for example). This is also the case for the observed duplications of tRNA-Ile and tRNA-Ala genes inside rRNA operons (between the 16S and the 23S genes).

Fig. 6. Location of the tRNA operons in the *Bacillus cereus* ATCC 14579 genome. The operons are framed by gray boxes.

However, this is not the only type of duplication that is observed. Both *B. megaterium* strains have a recent large duplicated block of tRNA genes right after the green block. There is an rRNA operon upstream of this block which suggests that it could be a new tRNA operon that was created. The same kind of mechanism occurred on the branch leading to Anc20 and gave rise to a new operon.

Validation on Simulated Data Sets

In order to evaluate the confidence on the obtained results, we tested OrthoAlign on simulated data sets mimicking the bacterial tRNA repertoires. Starting from an ancestral genome of about a hundred genes (typical number of tRNA genes in a *Bacillus* genome) on an alphabet of size 21 (number of tRNA isoacceptor families, counting the tRNA-Sec family), we simulated evolutionary histories according to a given phylogeny. As suggested by the inference in the *Bacillus* genomes (fig. 2), the size of the events was sampled according to a geometric distribution of parameter $p = 0.5$, and the relative numbers of loss (n_1), duplication (n_2), inversion (n_3), transposition (n_4), and substitution (n_5) events were sampled according to a multinomial distribution with respective parameters $p_1 = 0.516$, $p_2 = 0.347$, $p_3 = 0.084$, $p_4 = 0.043$ and $p_5 = 0.01$. All the results shown below were obtained by averaging over a given number of replicates (indicated in the captions of figures).

We focused on testing the ability of OrthoAlign to infer the correct number of evolutionary events, ancestral genomes, and size distributions of the events on small and large phylogenies. More precisely, two measures were used to analyze the performance of OrthoAlign on inferring the evolutionary histories. First, the error rate of a given evolutionary history is defined as Error $= \frac{|\text{NbReal} - \text{NbInfer}|}{\text{NbReal}}$, where "NbReal" is the real (simulated) number of events in the whole tree, and "NbInfer" is the inferred number of events. Second, the distance between the simulated ancestral genome and the inferred ancestor is computed by aligning them using the described pairwise alignment algorithm.

Testing on Small Phylogenies

We first considered triplet phylogenies, that is, cherries with a single sibling, and inferred the ancestral genome representing the root of the cherry. The third genome is used by OrthoAlign for the correction step (as described in the section New Approach: OrthoAlign).

Not surprisingly, as shown in figure 7, the error rate increases with the number l of events performed on each branch of the tree. For an evolutionary model restricted to duplications and losses (blue line), this error rate remains close to 0, and the inferred ancestor is accurate for $l \leq 3$, which is a little bit more than the average number of duplications and losses per branch observed in *Bacillus* (2.4). Adding inversions to the evolutionary model (red line) increases the average error rate, but we still observe the same trend as with the model of duplications and losses only. However, the error rate increases significantly for an evolutionary model involving transpositions (green line). In this

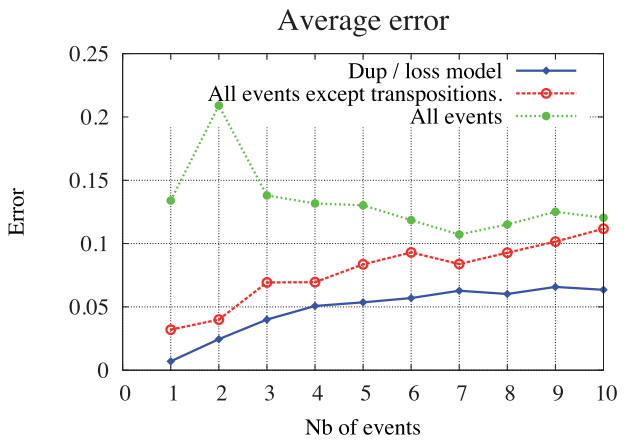

Fig. 7. Error rate for the ancestral inference on cherries. For every given number of evolutionary events per branch (*x* axis of the chart), 500 simulations are performed. Blue refers to the evolutionary model restricted to duplications and losses, red to the model involving all operations but transpositions, and green refers to the model involving all operations including transpositions.

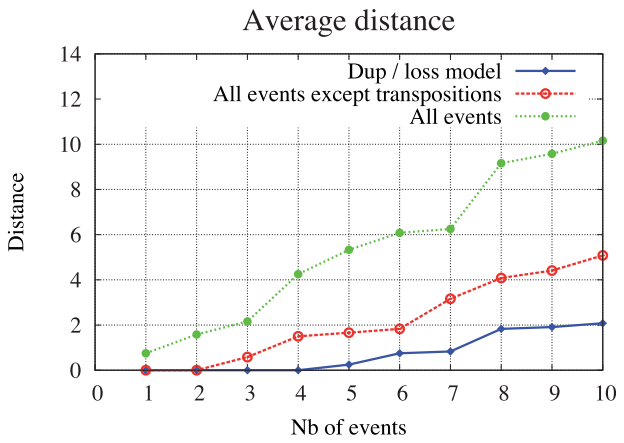

Fig. 8. Ancestral inference using the tree structure of figure 4. The color code is the same as in figure 7. For a given number of evolutionary events per branch (*x* axis), 50 simulations are performed. The *y* axis gives the distance between the true and inferred ancestor. Results are averaged over all simulations, and all internal nodes of the tree.

case, the tendency is to infer fewer events. This is not surprising as a duplication followed by the loss of the source can alternatively be explained by a single transposition.

Testing on Large Phylogenies

In this section, we performed simulations using the tree structure of figure 4. Figure 8 illustrates the performance in inferring correct ancestors. A very good performance is obtained for the duplication/loss model (blue line). Indeed, the algorithm performs well for $l \leq 4$, and for $l \leq 10$ the distance between the true and inferred genomes is close to 2. However, in the case of a general model accounting for rearrangements, due to the nonuniqueness of solutions, the inferred ancestor may be significantly different from the true one, although the distance between the true and inferred ancestors remains low for $l \leq 3$ (average number of duplication and loss events per branch observed in *Bacillus*).

Finally, figure 9 illustrates the accuracy of OrthoAlign to infer the correct duplication and deletion size distributions for $l = 7$ events per branch. Overall the inferred size distributions are similar to the simulated ones, although we observe a slight underestimation of the number of deletions and an overestimation of duplications of size one and two. This shows that OrthoAlign tends to mistake some deletions for duplication events. Notice that a single duplication inferred by mistake instead of a deletion may give rise to many more duplications in the evolutionary history if the error occurred at the bottom of the tree (see fig. 1 for an example).

Overall, OrthoAlign appears to infer ancestral configurations, as well as the number and size distribution of events with good accuracy for data sets that are close to the *Bacillus* tRNA repertoire.

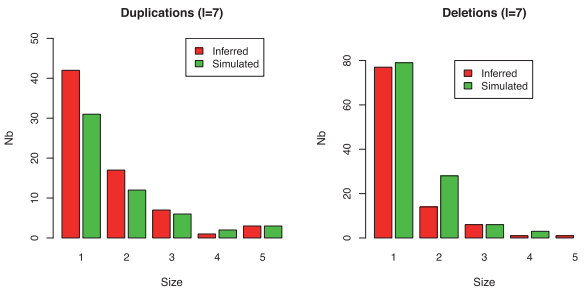

Fig. 9. Simulated (in green) versus inferred (in red) distributions of duplications and deletions sizes, for a number $l = 7$ of events per branch. The results are averaged over 50 simulations.

Conclusion

OrthoAlign is a new tool for inferring evolutionary scenarios and ancestral gene orders. It is based on an alignment approach, and infers nonoverlapping events still visible in the alignment. It is a cubic-time heuristic, which is very fast compared with previously developed exponential approaches (Andreotti et al. 2013; Holloway et al. 2013). The considered model of evolution accounts for both rearrangements (inversions and transpositions) and content-modifying (duplications and losses) operations. Lateral gene transfer is another source of tRNA gene content modification. Although it has not been included in the evolutionary model, a duplication event with source and target in two different genomes may be interpreted as a duplication followed by a loss, a transposition, or alternatively as a lateral gene transfer. Distinction between these events cannot be made without a rigorous exploration of sequence characteristics.

Using OrthoAlign to analyze the tRNA repertoires of 50 sequenced bacteria in the *Bacillus* genus, we have been able to answer many questions regarding the location, distribution, length, and rates of various evolutionary events, and to identify a core set of tRNAs. The high degree of tRNA gene order conservation has largely reduced the issue of nonuniqueness of solutions returned by OrthoAlign. Additional automated and manual curation made using the phylogenetic context further reduced the set of possible solutions. Although the phylogenetic inference was based on a simple grouping of

tRNAs according to their isoacceptor family, the main inferred events were then validated by aligning the nucleotide sequences of the corresponding regions, and thus by considering the full tRNA sequences and also the noncoding DNA inside the regions.

This study stands on the assumption that the genome assemblies and gene sequences are correct, which is not fully satisfied due to sequencing errors. As detailed in the Results section, one of the two inferred tRNA substitutions was actually rather due to sequencing errors affecting the tRNA anticodon. Notice however that sequencing errors occurring elsewhere (not in the tRNA anticodon) does not affect our studies, as such a tRNA is still assigned to the correct family.

Another hidden assumption is that the considered phylogeny is correct. As far as we know, there is no alternative phylogeny in the literature. Yet, due to various reasons such as alignment ambiguities or lateral gene transfer, the phylogeny given in figure 3 may be erroneous. This suspicion is reinforced by the observation that the same event (loss of the red block of size nine) is inferred not less than 11 times in the phylogeny of figure 3. Based uniquely on this observation, we propose an alternative phylogeny in figure 10 (Appendix C).

This study has been conducted under a uniform unit cost model for operations. One may alternatively choose to favor one operation with respect to another. For example, choosing a prohibitive penalty for rearrangement events would lead to an increase of IDs and losses. Similarly, a prohibitive cost for losses would have the effect of favoring duplications. Notice that restricting losses to operations of size one (single gene losses) actually has the effect of favoring duplications. This bias is however largely reduced by using the correction by sibling strategy. Although obtained results are sensitive to parameter setting, the main events inferred in the *Bacillus* phylogenetic tree are large enough to be still visible with alternative costs for rearrangement and content-modifying operations. Moreover, all these events have been verified by aligning the whole DNA sequences encompassing the considered blocks.

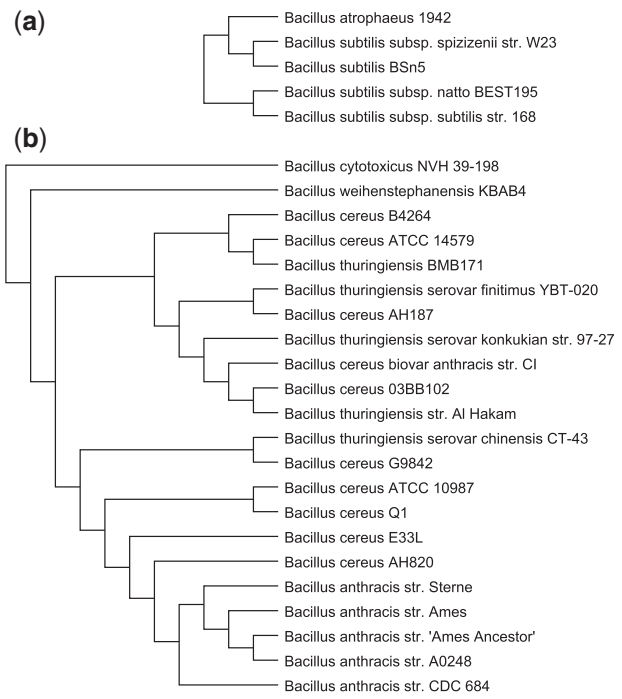

Fig. 10. Corrected phylogeny reducing the number of deletions of the red block. (*a*) The corrected *Bacillus subtilis* monophyletic group. (*b*) The corrected *B. thuringiensis*, *B. cereus*, and *B. anthracis* monophyletic group.

In the near future, we would like to improve OrthoAlign by introducing a more robust algorithm for integrating the pairwise alignment approach in the phylogenetic framework. A natural extension is to consider the median of three genomes rather than the common ancestor of a cherry. As for the evolution of tRNA repertoires, many findings remain to be explored. In particular, the second tRNA substitution revealed by our analysis remains to be validated. On the other hand, what is the effect of a high tRNA-Ile and tRNA-Ala concentration in rRNA operons? Does it correlate with a high co-transcription level? What is the functional implication of the presence/absence of tRNA-Sec? From a phylogenetic point of view, what is the support of the new proposed phylogeny? Indeed, the interpretation of bacterial phylogenies has to be seen in light of potentially massive lateral gene transfer, across large distance but even more easily at short distance.

Materials and Methods

We analyzed the evolutionary history of 50 fully sequenced bacteria in the *Bacillus* genus (including eight *Geobacillus* strains). The tRNA gene content and order were taken from GenBank (Benson et al. 2013) (see table 1 for the complete list of genomes with accession numbers and dates of download). The phylogeny of the studied strains shown in figure 3 was taken from the Pathosystems Ressource Integration Center (PATRIC) (Gillespie et al. 2011).

Tranfer RNA genes were grouped according to their isoacceptor group. In other words, all tRNAs carrying the same amino acid were assigned to the same family. We could instead have separated tRNAs by anticodon, or by whole sequence similarity. Such a grouping is not appropriate for studying large duplications, as singletons would be dominant in the genome representation, and very few duplications would be observed. Notice however that, after applying the automated OrthoAlign software, the major inferred events were subsequently validated by aligning the nucleotide sequences of the corresponding regions, and thus by taking into account the whole tRNA sequences and the noncoding DNA in between.

In order to align the genomes correctly, we needed to know the locations of the origin and the terminus of replication for each strain. We used T-A and G-C skews from comparative genometrics (Roten et al. 2002) and Oriloc (Frank and Lobry 2000) to find those locations.

For the sake of presentation, genomes are subdivided into blocks, using the tRNA operon subdivision available for *B. cereus* ATCC 14579 (Candelon et al. 2004): each tRNA operon is considered to be a block. As shown in figure 11, blocks that are identical or very similar in terms of

Table 1. *Bacillus* Genomes Studied.

Genome	GenBank Accession and Version Number	Date of Download
Bacillus amyloliquefaciens DSM 7	NC_014551.1	August 23, 2012
Bacillus amyloliquefaciens FZB42	NC_009725.1	August 23, 2012
Bacillus amyloliquefaciens LL3	NC_017190.1	August 23, 2012
Bacillus amyloliquefaciens TA208	NC_017188.1	August 23, 2012
Bacillus anthracis str. A0248	NC_012659.1	August 23, 2012
Bacillus anthracis str. "Ames Ancestor"	NC_007530.2	August 23, 2012
Bacillus anthracis str. Ames	NC_003997.3	August 23, 2012
Bacillus anthracis str. CDC 684	NC_012581.1	August 23, 2012
Bacillus anthracis str. Sterne	NC_005945.1	August 23, 2012
Bacillus atrophaeus 1942	NC_014639.1	August 23, 2012
Bacillus cellulosilyticus DSM 2522	CP002394.1	August 23, 2012
Bacillus cereus 03BB102	NC_012472.1	August 23, 2012
Bacillus cereus AH187	NC_011658.1	August 23, 2012
Bacillus cereus AH820	NC_011773.1	August 23, 2012
Bacillus cereus ATCC 10987	NC_003909.8	August 23, 2012
Bacillus cereus ATCC 14579	NC_004722.1	August 23, 2012
Bacillus cereus B4264	NC_011725.1	August 23, 2012
Bacillus cereus biovar anthracis str. CI	NC_014335.1	August 23, 2012
Bacillus cereus E33L	NC_006274.1	August 23, 2012
Bacillus cereus G9842	NC_011772.1	August 23, 2012
Bacillus cereus Q1	NC_011969.1	August 23, 2012
Bacillus clausii KSM-K16	NC_006582.1	August 23, 2012
Bacillus coagulans 2-6	NC_015634.1	August 23, 2012
Bacillus coagulans 36D1	NC_016023.1	August 23, 2012
Bacillus cytotoxicus NVH 391-98	NC_009674.1	August 23, 2012
Bacillus halodurans C-125	NC_002570.2	August 23, 2012
Bacillus licheniformis ATCC 14580	NC_006322.1	August 23, 2012
Bacillus megaterium DSM 319	NC_014103.1	August 23, 2012
Bacillus megaterium QM B1551	NC_014019.1	August 23, 2012
Bacillus pseudofirmus OF4	CP001878.2	August 23, 2012
Bacillus pumilus SAFR-032	NC_009848.1	August 23, 2012
Bacillus selenitireducens MLS10	NC_014219.1	August 23, 2012
Bacillus subtilis BSn5	NC_014976.1	August 23, 2012
Bacillus subtilis subsp. natto BEST195	AP011541.1	August 23, 2012
Bacillus subtilis subsp. spizizenii str. W23	NC_014479.1	August 23, 2012
Bacillus subtilis subsp. subtilis str. 168	NC_000964.3	August 23, 2012
Bacillus thuringiensis BMB171	NC_014171.1	August 23, 2012
Bacillus thuringiensis serovar chinensis CT-43	NC_017208.1	August 23, 2012
Bacillus thuringiensis serovar finitimus YBT-020	NC_017200.1	August 23, 2012
Bacillus thuringiensis serovar konkukian str. 97-27	NC_005957.1	August 23, 2012
Bacillus thuringiensis str. Al Hakam	NC_008600.1	August 23, 2012
Bacillus weihenstephanensis KBAB4	NC_010184.1	August 23, 2012
Geobacillus kaustophilus HTA426	NC_006510.1	August 30, 2012
Geobacillus sp. C56-T3	NC_014206.1	August 30, 2012
Geobacillus sp. WCH70	NC_012793.1	August 30, 2012
Geobacillus sp. Y412MC52	NC_014915.1	August 30, 2012
Geobacillus sp. Y412MC61	NC_013411.1	August 30, 2012
Geobacillus sp. Y4.1MC1	NC_014650.1	August 30, 2012
Geobacillus thermodenitrificans NG80-2	NC_009328.1	August 30, 2012
Geobacillus thermoglucosidasius C56-YS93	NC_015660.1	August 30, 2012

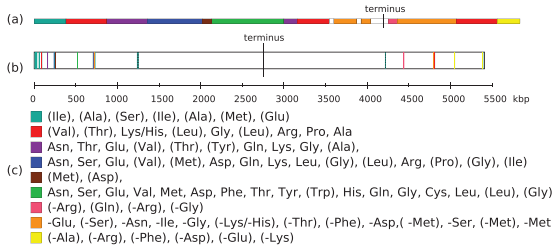

Fig. 11. (*a*) Representation of tRNA syntenic blocks on the *Bacillus cereus* ATCC 14579 genome. (*b*) Locations of the tRNA blocks on the whole genome of *B. cereus* ATCC 14579. Dashed lines represent the white variable blocks. (*c*) Ordered sequences of tRNA isoacceptor families, representing the signature of colored blocks. A slash (/) between two tRNA genes indicates that one or the other can be found at that position. The tRNA genes inside parentheses are absent from the block in some strains.

tRNA gene content and order are assigned the same color. A vertical line shows the location of the terminus of replication. Colored blocks represent syntenic blocks that are well conserved across all the strains. A signature (consensus representation) of blocks is given in figure 11*c*. On the opposite, variable regions are drawn in white. The block length is proportional to the number of tRNA genes inside it. Figure 11*b* illustrates where the tRNA gene blocks are located on the chromosome of *B. cereus* ATCC 14579. In all the studied genomes, the majority of tRNA genes are found near the origin of replication and variable regions tend to be located far from it.

Implementation

OrthoAlign has been implemented in Java. It has been run with a default cost of one per event. In other words, the edit distance is considered, which is the minimum number of operations required to transform one genome into the other. Operations can be of any size, except losses and substitutions that are of size one. Observed substitutions may point out to a tRNA change of identity due to gene conversion, or may simply be due to sequencing errors. Considering substitutions independent among sites is therefore sound. As for losses, restriction to single gene losses is a methodological requirement to avoid biases toward very long deletions (if a loss of any size costs one, then any alignment costs at most two: simply delete the entire first genome and then delete the entire second genome). To cope with losses of size greater than one, we perform a postprocessing by simply grouping all consecutive gene losses into a single event.

Relative Dating of Divergence Time

To estimate the relative divergence time of the most recent common ancestor of the 50 *Bacillus* strains studied, we proceeded as follows. We used progressiveMauve (Darling et al. 2010) to make multiple whole genome alignments for 13 of the 50 *Bacillus* strains (*B. anthracis* str. "Ames Ancestor," *B. cellulosilyticus* DSM 2522, *B. cereus* 03BB102, *B. thuringiensis* serovar chinensis CT-43, *B. coagulans* 36D1, *B. cytotoxicus* NVH 391-98, *Geobacillus* sp. WCH70, *Geobacillus* sp. Y412MC52, *B. licheniformis* ATCC 14580, *B. megaterium* DSM 319, *B. pseudofirmus* OF4, *B. subtilis* subsp. natto BEST195, and *B. thuringiensis* serovar finitimus YBT-020) and the four genomes analyzed in the *E. coli* study (Withers et al. 2006) (*E. coli* CFT073, *E. coli* O157:H7 str. EDL933, *E. coli* str. K-12 substr. MG1655, and *Shigella flexneri* 2a str. 301, hereafter referred to as the *E. coli* genomes for short). The 13 *Bacillus* strains were selected as to represent each subgroup in the species tree. Then, the stripSubsetLCBs script, available on the Mauve snapshots webpage (http://darlinglab.org/mauve/snapshots/, last accessed February 23, 2015), was used to extract the core alignments containing sequences from all the genomes.

The software Beast (Drummond et al. 2012) was used with an arbitrary root height of one on the core alignments to estimate the relative divergence time of the *Bacillus* strains based on the *E. coli* genomes. Indeed, since indisputable reference points are not available for microbial genomes, comparing the *Bacillus* genomes with the *E. coli* genomes allows us to get a relative measure of the divergence time and then to compare the rates of the different evolutionary events.

More precisely, in order to force the root height to keep a value of one, a uniform prior on the treeModel.rootHeight variable was set with bounds of 0.99 and 1.01. Two monophyletic taxa were created: one including the four *E. coli* genomes and one including the 13 *Bacillus* genomes. In order to estimate the divergence time of the most recent common ancestor for both taxa, Beast requires an initial value for the date of divergence: 0.5 was chosen for both taxa. The Hasegawa-Kishino-Yano substitution model and a strict clock model were used. Ten million was used for the length of the Markov chain Monte Carlo chain.

Acknowledgments

The authors thank Jiqiang Ling and his lab at the University of Texas-Houston Medical School for the resequencing experiments of the region containing the tRNA-Val gene in DSM 7. The authors also thank Krister M. Swenson for his help with the initial version of the program, and for fruitful discussions. This work was supported by "Fonds de recherche du Québec-Nature et technologies" (FRQNT) and the Natural Sciences and Engineering Research Council of Canada (NSERC).

Appendix A: Evolutionary events

- A duplication $D = (X[i, i+k], Y[j, j+k])$, where $Y[i, i+k] = X[i, i+k]$, is an operation that copies the substring $X[i, i+k]$ to a location j outside the interval $[i, i+k]$ (i.e., preceding i or following $i+k$).
- A loss $L = (X[i, i+k], \oslash)$ ($\oslash$ for empty string) is an operation that removes the substring $X[i, i+k]$ from genome X.

- A substitution $S = (X[i, i+k], Y[j, j+k])$ is an operation that replaces the substring $X[i, i+k]$ by a string $Y[i, i+k]$ of the same length.
- An inversion $I = (X[i, i+k], Y[i, i+k])$, where $Y[i, i+k] = -X[i, i+k]$, is an operation that transforms the substring $X[i, i+k]$ into its reverse.
- An inverted duplication $\mathrm{ID} = (X[i, i+k], Y[j, j+k])$, where $Y[j, j+k] = -X[i, i+k]$, is an operation that copies the reverse of the substring $X[i, i+k]$ to a location j outside the interval $[i, i+k]$.
- A transposition $T = (X[i, i+k], Y[j, j+k])$, where $Y[j, j+k] = X[i, i+k]$, is an operation that moves the substring $X[i, i+k]$ to another position j outside the interval $[i, i+k]$.
- An inverted transposition $\mathrm{IT} = (X[i, i+k], Y[j, j+k])$, where $Y[j, j+k] = -X[i, i+k]$ is an operation that removes the substring $X[i, i+k-1]$ and places its reverse substring somewhere else in the genome.

Appendix B: Recurrences of the dynamic programming algorithm for 2-SPP

- $$M(i,j) = \begin{cases} C(i-1, j-1) & \text{if } X[i] = Y[j] \\ +\infty & \text{otherwise} \end{cases}$$
- $$S(i,j) = \begin{cases} C(i-1, j-1) + c(S(1)) & \text{if } X[i] \neq Y[j] \\ +\infty & \text{otherwise} \end{cases}$$
- $$I(i,j) = \begin{cases} min_{k \in E}[C(k, j-(i-k)) + c(I(i-k))] \\ \qquad \text{if } E \neq \varnothing \\ +\infty \qquad \text{otherwise} \end{cases}$$

  where E is the set $\{s_1, s_2, \ldots, s_{|<i}\}$ of maximum cardinality such that $X[i - s_p, i]$ is the reverse of $Y[j - (i - s_p), j]$ for all $1 \leq p \leq l$.
- $L_X(i,j) = min_{0 \leq k \leq i-1}[C(k,j) + c(L(i-k))]$
  (the corresponding formula holds for $L_Y(i,j)$)
- $$D_X(i,j) = \begin{cases} +\infty & \text{if } X[i] \text{ is a singleton} \\ \min_{l \leq k \leq i-1}[C(k,j) + c(D(i-k))] \\ \qquad \text{otherwise} \end{cases}$$

  where $X[l, i]$ is the longest suffix of $X[1, i]$ that has an occurrence elsewhere in X or Y.
  (the corresponding formula holds for $D_Y(i,j)$)
- $$ID_X(i,j) = \begin{cases} +\infty & \text{if } X[i] \text{ is a singleton} \\ min_{l \leq k \leq i-1}[C(k,j) + c(ID(i-k))] \\ \qquad \text{otherwise} \end{cases}$$

  where $X[l, i]$ is the longest suffix of $X[1, i]$ such that $-X[l, i]$ is present elsewhere (nonovrelapping $X[l, i]$) in X or Y.
  (the corresponding formula holds for $ID_Y(i,j)$)

Appendix C: Corrected phylogeny

We propose a correction to the phylogeny of figure 3 based on the repetitive deletion of the red blocks described in figure 11. Indeed, the red blocks (there can be up to three red blocks in the genomes) are particularly well-conserved throughout all the strains studied here in terms of their relative positions in the genomes and their composition. Thus, it is very likely that they were created early in the ancestral species and then lost in some strains. However, the phylogeny provided by the PATRIC website made it impossible to infer fewer than 11 deletions of the red block which, as mentioned previously, seems unlikely because long duplications and deletions tend to occur less often.

By rearranging two monophyletic groups (the *B. subtilis* group and the *B. thuringiensis*, *Bacillus cereus*, and *B. anthracis* group) in the phylogeny of figure 3 while preserving as much of the original topology as possible, we were able to eliminate six deletions of the red block from our evolutionary history. Moreover, our proposed modifications did not change the number of the other events inferred in those two subtrees. The corrected monophyletic groups are shown in figure 10.

References

Alekseyev M, Pevzner PA. 2009. Breakpoint graphs and ancestral genome reconstructions. *Genome Res.* 19:943–957.

Andreotti S, Reinert K, Canzar S. 2013. The duplication-loss small phylogeny problem: from cherries to trees. *J Comput Biol.* 20(9):643–659.

Ardell D, Andersson S. 2006. TFAM detects co-evolution of tRNA identity rules with lateral transfer of histidyl-tRNA synthetase. *Nucleic Acids Res.* 34:893–904.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. Genbank. *Nucleic Acids Res.* 41(D1):D36–D42.

Benzaid B, Dondi R, El-Mabrouk N. 2013. Duplication-loss genome alignment: complexity and algorithm LNCS. In: Dediu A-H, Martín-Vide C, Truthe B, editors. Vol. 7810 (Language and Automata Theory and Applications (LATA)). New-York: Springer Berlin Heidelberg. p. 116–127.

Bermudez-Santana C, Attolini CS, Kirsten T, Engelhardt J, Prohaska S, Steigele S, Stadler P. 2010. Genomic organization of eukaryotic tRNAs. *BMC Genomics* 11(270):1–14.

Brewer BJ. 1988. When polymerases collide: replication and the transcriptional organization of the *E. coli* chromosome. *Cell* 53(5):679–686.

Candelon B, Guilloux K, Ehrlich SD, Sorokin A. 2004. Two distinct types of rRNA operons in the *Bacillus cereus* group. *Microbiology* 150(3):601–611.

Copeland P. 2005. Making sense of nonsense: the evolution of selenocysteine usage in proteins. *Genome Biol.* 6(6):221.

Darling AE, Mau B, Perna NT. 2010. Progressivemauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5(6):e11147.

Dondi R, El-Mabrouk N. 2013. Aligning and labeling genomes under the duplication-loss model. In: Bonizzoni, Paola and Brattka, Vasco and Löwe, Benedikt, editors. The Nature of Computation. Logic, Algorithms, Applications, (Lecture Notes in Computer Science (LNCS)), Vol. 7921. New-York: Springer Berlin Heidelberg. p. 97–107.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with beauti and the beast 1.7. *Mol Biol Evol.* 29(8):1969–1973.

El-Mabrouk N. 2005. Genome rearrangement with gene families. In: Olivier G, editor. Mathematics of evolution and phylogeny. New-York: Oxford University Press. p. 291–320.

El-Mabrouk N, Sankoff D. 2012. Part II: Evolutionary genomics: statistical and computational methods Analysis of gene order evolution beyond single-copy genes. In: Anisimova M, editor. Vol. 855 (Methods in Molecular Biology). New-York: Humana Press. p. 397–429.

Fertin G, Labarre A, Rusu I, Tannier E, Vialette S. 2009. Combinatorics of genome rearrangements. Cambridge (MA): The MIT Press.

Frank AC, Lobry JR. 2000. Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics* 16(6):560–561.

Gagnon Y, Blanchette M, El-Mabrouk N. 2012. A flexible ancestral genome reconstruction method based on gapped adjacencies. *BMC Bioinformatics* 13(Suppl 19), S4.

Gillespie JJ, Wattam AR, Cammer SA, Gabbard JL, Shukla MP, Dalay O, Driscoll T, Hix D, Mane SP, Mao C, et al. 2011. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect Immun.* 79(11):4286–4298.

Holloway P, Swenson K, Ardell D, El-Mabrouk N. 2013. Ancestral genome organization: an alignment approach. *J Comput Biol.* 20(4):280–295.

Jones BR, Rajaraman A, Tannier E, Chauve C. 2012. Anges: reconstructing ancestral genomes maps. *Bioinformatics* 28(18):2388–2390.

Kovac J, Brejova B, Vinar T. A practical algorithm for ancestral rearrangement reconstruction. In: Przytycka TM, Sagot M-F, editors. LNBI. Vol. 6833 (WABI). New-York: Springer Berlin Heidelberg. p. 163–174.

Lavrov D, Lang B. 2005. Transfer RNA gene recruitment in mitochondrial DNA. *Trends Genet.* 21:129–133.

Ling J, Daoud R, Lajoie M, Church G, Soll D, Lang B. 2014. Natural reassignment of CUU and CUA sense codons to alanine in *Ashbya*. *Nucleic Acids Res.* 42(1):499–508.

Ma J, Zhang L, Suh B, Raney B, Burhans R, Kent W, Blanchette M, Haussler D, Miller W. 2007. Reconstructing contiguous regions of an ancestral genome. *Genome Res.* 16:1557–1565.

Mao X, Zhang H, Yin Y, Xu Y. 2012. The percentage of bacterial genes on leading versus lagging strands is influenced by multiple balancing forces. *Nucleic Acids Res.* 40(17):8210–8218.

Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29(22):2933–2935.

Ouangraoua A, Tannier E, Chauve C. 2011. Reconstructing the architecture of the ancestral amniote genome. *Bioinformatics* 27(19):2664–2671.

Pe'er I, Shamir R. 1998. The median problems for breakpoints are NP-complete. *Elec Colloq Comput Complexity*. 71.

Rocha E, Danchin A. 2003. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat Genet.* 34(4):377–378.

Rocha EPC. 2004. The replication-related organization of bacterial genomes. *Microbiology* 150(6):1609–1627.

Rogers H, Bergman C, Griffiths-Jones S. 2010. The evolution of tRNA genes in *Drosophila*. *Genome Biol Evol.* 2:467–477.

Roten CAH, Gamba P, Barblan JL, Karamata D. 2002. Comparative genometrics (CG): a database dedicated to biometric comparisons of whole genomes. *Nucleic Acids Res.* 30(1):142–144.

Saks M, Sampson J, Abelson J. 1998. Evolution of a transfer RNA gene through a point mutation in the anticodon. *Science* 279: 1665–1670.

Sankoff D, Blanchette M. 1998. Multiple genome rearrangement and breakpoint phylogeny. *J Comput Biol.* 5:555–570.

Tillier E, Collins R. 2000. Genome rearrangement by replication-directed translocation. *Nat Genet.* 26:195–197.

Vasileva I, Moor N. 2007. Interaction of aminoacyl-tRNA synthetases with tRNA: general principles and distinguishing characteristics of the high-molecular-weight substrate recognition. *Biochemistry* 72(3):247–263.

Wang X, Lavrov D. 2011. Gene recruitment—a common mechanism in the evolution of transfer RNA gene families. *Gene* 475(1):22–29.

Withers M, Wernisch L, Reis MD. 2006. Archaeology and evolution of transfer RNA genes in the *Escherichia coli* genome. *Bioinformatics* 12: 933–942.

Zhang Y, Romero H, Salinas G, Gladyshev V. 2006. Dynamic evolution of selenocysteine utilization in bacteria: a balance between selenoprotein loss and evolution of selenocysteine from redox active cysteine residues. *Genome Biol.* 7(10):R94.

Zheng C, Sankoff D. 2011. On the pathgroups approach to rapid small phylogeny. *BMC Bioinformatics* 12:S4.

Zheng C, Sankoff D. 2012. Gene order in rosid phylogeny, inferred from pairwise syntenies among extant genomes. *BMC Bioinformatics* 13(Suppl 10), S9.