# On the Ability to Reconstruct Ancestral Genomes from *Mycobacterium* Genus

C. Guyeux[1], B. Al-Nuaimi[1,2], B. AlKindy[3], J.-F. Couchot[1], and M. Salomon[1]

[1] FEMTO-ST Institute, UMR 6174 CNRS, DISC Computer Science Department,
Univ. Bourgogne Franche-Comté (UBFC), France
[2]Department of Computer Science, University of Diyala, Iraq
[3] Department of Computer Science, University of Mustansiriyah, Baghdad, Iraq
christophe.guyeux@univ-fcomte.fr

**Abstract.** Technical signs of progress during the last decades has led to a situation in which the accumulation of genome sequence data is increasingly fast and cheap. The huge amount of molecular data available nowadays can help addressing new and essential questions in Evolution. However, reconstructing evolution of DNA sequences requires models, algorithms, statistical and computational methods of ever increasing complexity. Since most dramatic genomic changes are caused by genome rearrangements (gene duplications, gain/loss events), it becomes crucial to understand their mechanisms and reconstruct ancestors of the given genomes. This problem is known to be NP-complete even in the "simplest" case of three genomes. Heuristic algorithms are usually executed to provide approximations of the exact solution. We state that, even if the ancestral reconstruction problem is NP-hard in theory, its exact resolution is feasible in various situations, encompassing organelles and some bacteria. Such accurate reconstruction, which identifies too some highly homoplasic mutations whose ancestral status is undecidable, will be initiated in this work-in-progress, to reconstruct ancestral genomes of two *Mycobacterium* pathogenetic bacterias. By mixing automatic reconstruction of obvious situations with human interventions on signaled problematic cases, we will indicate that it should be possible to achieve a concrete, complete, and really accurate reconstruction of lineages of the *Mycobacterium tuberculosis* complex. Thus, it is possible to investigate how these genomes have evolved from their last common ancestors.

**Keywords:** Mycobacterium tuberculosis, genome rearrangements, ancestral reconstruction.

## 1 Introduction

*Mycobacterium tuberculosis* is presently still one of the principal causes of death worldwide. Approximately one-third of the world population is infected by the *Mycobacterium tuberculosis complex* (MTBC), with about 9 million event cases annually, leading to estimated a million deaths each year. Due to their different host tropism and phenotypes, members of MTB complex display various pathogenicities ranging from particularly human (*M. tuberculosis*, *M. africanum*, and *M. canetti*) or rodent pathogens (*M. microti*) to *Mycobacteria* with a broad host spectrum (*M. bovis*) [1–3]. *Mycobacterium tuberculosis* has been in the human population around for thousands of years, as fragments of the spinal column of Egyptian mummies from 2300 BCE show definite pathological signs of tubercular decay. It has been recognized as the leading cause of mortality by 1650, while using a new staining technique, Robert Koch identified the bacterium responsible for causing consumption in 1882.

The MTB complex belongs to the slow-growing sublineage of *Mycobacteria*. Based on topographical characteristics, MTBC can be categorized into six clusters, including species such as *M. tuberculosis*, *M. africanum*, *M. bovis*, *M. microti*, and *M. canettii*. Members in MTBC share 99.95% of their genomic sequences and a rigorously clonal population structure [4]. Compared to more ancient species (*e.g.*, *M. marinum*), MTBC has shorter but more virulent chromosomes [5, 6]. Considering that they all are derived from a common ancestor, it is interesting that some are human or rodent pathogens, whereas others have a wide host spectrum [7]. The genome of *M. tuberculosis* was studied using the strain *M. tuberculosis H37Rv*. It has a circular chromosome of about 4,200,000 nucleotides long, while containing about 4,000 genes [8]. The different species of the *Mycobacterium tuberculosis* complex show a $95-100\%$ DNA relatedness based on studies of DNA homology, and the sequences of the 16S rRNA gene are the same for all the species.

MTBC genomes have been modified during the evolution by mutation, insertion-deletion of nucleotides, by large-scale changes (inversion, duplication or deletion of large DNA strands), or by other modifications specific to repetition (insertion sequences, etc.). Being able predict both its past or its future evolution may have multiple applications: to reconstruct the past history and the ancestors of bacteria, or to better understand their mechanism of virulence and resistance acquisition. The relatively short timescale (tuberculosis disease is relatively recent, as its most recent common ancestor evolved $\approx$ 40,000 years ago [9]), the relatively reasonable sizes of considered genomes, the relative rarity of recombination events, and the recent possibility to have access to old and present bacterial DNA sequences, may lead to the possibility to model the evolution of these genomes, in order to reconstruct and to understand their ancient history and to predict their future evolution.

To do so, new algorithms of detection and of evolution regarding genomic modifications must be written. People working on this problematic mainly focus on predicting the evolution of nucleotide mutations, and by assuming specific forms for matrix mutations which seem incompatible with recent experimental measures [10]. These models for evolution must be designed differently, in order to better reflect the reality. Additionally, the serious impact of other modifications operating on the genomes (as insertions and deletions of nucleotides (indels), inter and intra chromosomic recombinations, or modifications specific to repetition), must be taken into account more deeply, while a concrete ancestral reconstruction of bacterial lineage must be finally achieved.

The objective of this work-in-progress is to prove that, given a set of close bacterial genomes, it is possible to reconstruct in practice their recent sequence evolution history, by mixing state-of-the-art tools with a pragmatic manual completion and cross-validation. We will illustrate that, in practice, it should be possible to reconstruct ancestral genomes for some lineages of the *Mycobacterium* genus, using all available complete genomes of such a lineage (for instance, 65 complete genomes of the MTB complex are currently available, and we have more than 1,000 archives of reads).

The remainder of this article is organized as follows. In Section 2, we start by giving reviews of computational approaches and tools for analyzing the evolution of DNA sequences. We propose in the next section a set of methodological principles that can be used for ancestral genome reconstructions, and how to apply them on *M. canettii* and *M. tuberculosis* data. Obtained results and further perspectives are discussed in Section 4. This research work ends with a conclusion section in which the article is summarized and intended future work is outlined.

## 2 Scientific background

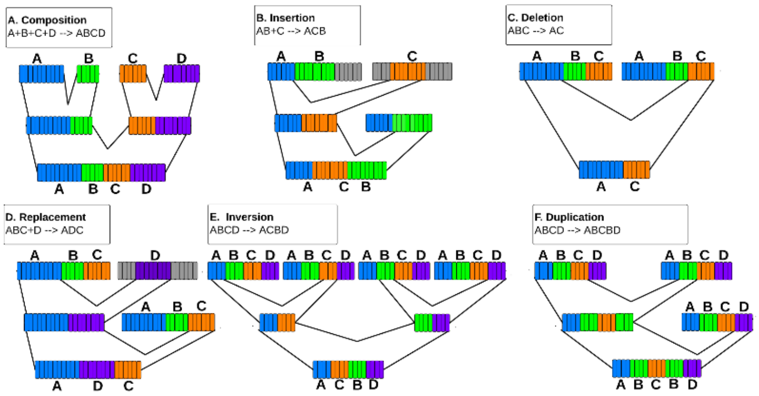


**Fig. 1.** Various genome rearrangement events.

### 2.1 On genomic evolution

It is well-known that DNA sequences change over time due to local mutations, which are either single nucleotide polymorphism (SNP) or insertion-deletion (indel) of one nucleotide. Mutations that affect the organization of genes are called genome rearrangements, which include inversions, transpositions, and chromosome fusions and fissions. Example of such large scale modifications are illustrated in Figure 1. During evolution, such large-scale mutations rearranging the genome have occurred, and both gene order and content have been modified accordingly, which may represent a meaningful role in speciation [11].

One important problem in molecular evolution, which is targeted by this study, is that of reconstructing ancestral genomic sequences. In this problem, an evolutionary tree of organisms is provided, together with genomic sequences for the leaf species. The aim is to infer the genomic sequences of the ancestral nodes in the tree, that is, those of the organisms that no longer exist. Various methods have been developed to infer such ancestral sequences and they already have been used in various biological studies.

More precisely, evolution of biomolecules over time have mainly been computationally studied in two directions, namely through ancestral genome reconstruction problem and through the evolution of pan and core genomes over time. A short overview of these topics is provided below.

### 2.2 Ancestral genome reconstruction

Ancestral reconstruction may focus at sequence level or at gene order level, the former being quite resolved [12–20], at least if we do not consider indels and mutation neighborhood, while the latter is more difficult in general, due to its combinatorial complexity. More precisely, given an alignment of DNA sequences and a tree, ancestral nucleotides of extant species can be obtained by modeling the evolution of a trait through time as a stochastic process (Markov chain). Using it as the basis for statistical inference, both maximum likelihood or Bayesian inference approaches can be applied to estimate ancestral configuration.

Well known software like RAxML [21], BEAST2 [22], or PAML [23] can be used for such reconstructions. However, most of the time, like in the R package [24], indels are not considered in such ancestral state reconstruction, even if researches have recently been realized via the so-called "Poisson Indel Process" [25]. Such process is a significant improvement, if we compare it with the parsimony approach that can be found in PHAST software, or with the Thorne-Kishino-Felsenstein model of indel evolution. Large scale modifications, for its part, is most of the time regarded in a combinatorial framework by modeling genomes as permutations of genes or homologous regions. Indeed, this genome rearrangement problem [26] is usually formulated as follows: "given two genomes (permutations) and a set of allowable operations (like inversion, deletion, or transposition), what is the shortest sequence of operations that will transform one genome into the other?". However, even in the case of three genomes, such a problem is NP-hard [27], although it has received much attention in mathematics and computer science [11].

An important remark, motivating our proposal, is that the NP-hard character of this problem only appears if we consider a very large number of operations in very large sequences. On our side, we will consider quite small sequences and a relatively small number of large scale recombinations. So we face tractable problems in various real situations, on which simple and pragmatic approaches may work.

### 2.3 Core and pan genome extraction

An early study about finding the common genes in chloroplasts has been realized by Stoebe et al. in 1998 [28]. They established the distribution of 190 identified genes and 66 hypothetical protein-coding genes (ysf) in all nine photosynthetic algal plastid genomes available (excluding non-photosynthetic Astasia tonga) from the last update of plastid genes nomenclature and distribution. The distribution reveals a set of approximately 50 core protein-coding genes retained in all taxa. In 2003, Grzebyk et al. [29] have studied the core genes among 24 chloroplast sequences extracted from public databases, 10 of them being algae plastid genomes. They broadly clustered the 50 genes from Stoebe et al. into three major functional domains: (1) genes encoded for ATP synthesis (atp genes); (2) genes encoded for photosynthetic processes (psa and psb genes); and (3) housekeeping genes that include the plastid ribosomal proteins (rpl and rps genes). The study shows that all plastid genomes were rich in housekeeping genes with the rbcL gene involved in photosynthesis. Other examples of such core and pan studies can be found in, e.g., [30–32].

Concerning bacterias, many studies have recently achieved the extraction of core and pan genomes using NCBI annotations, which are mainly based on generic annotation tools like Glimmer, GeneMarkS, or Prodigal (see for instance [33–35], the Pseudomonas aeruginosa case being resolved by us in [36]). In most of these studies, considered genomes have been annotated with various different annotation algorithms, mixing human curated and automated coding sequence prediction tools that are not specific to the genus under consideration. This large variety of manners to detect coding sequences and their functionality leads to large variability in gene boundaries (start and stop codons) and naming process, which obviously severely biases the core and pan genomes determination.

## 3 A concrete semi-automatic ancestral reconstruction

### 3.1 General presentation

By a phylogenetic study, it is possible to reconstruct the evolutionary relationship of a set of organisms in the form of a binary tree, in which the given set of organisms are descendants placed at the leaves, while internal nodes stand for extinct ancestors connected by edges. We argue that, knowing this tree, ancestral genomes can be completely reconstructed in some easy cases, by aligning extant genomes and finding homologies between them, and then inferring various scenarii of evolutionary events during history [37]. This ancestral reconstruction can be achieved by mixing state-of-the-art algorithms and manual investigations, if the considered genomes have not evolved so much. To illustrate the feasibility of the proposal, an example of such reconstruction is provided in this section, in the case of the MTB complex.

```
Sequence1   -TCAGGA-TGAAC----
Sequence2   ATCACGA-TGAACC---
Sequence3   ATCAGGAATGAATCC--
Sequence4   -TCACGATTGAATCGC-
Sequence5   -TCAGGAATGAATCGCM
```

**Fig. 2.** Representation of a multiple sequence alignment.

**Table 1.** Information about some *Mycobacterium* genomes

| Organism name | Accession | Sequence length | Number of genes |
|---|---|---|---|
| *Mycobacterium tuberculosis W-148* | NZ_CP012090.1 | 4,418,548 bp | 4,133 |
| *Mycobacterium tuberculosis H37Rv* | NC_018143.2 | 4,411,709 bp | 4,132 |
| *Mycobacterium africanum GM041182* | NC_015758.1 | 4,389,314 bp | 4,089 |
| *Mycobacterium africanum strain 25* | CP010334.1 | 4,386,422 bp | 4,798 |
| *Mycobacterium microti strain 12* | CP010333.1 | 4,370,115 bp | 4,321 |
| *Mycobacterium canettii CIPT 140010059* | NC_015848.1 | 4,482,059 bp | 4,137 |
| *Mycobacterium canettii CIPT 140070008* | NC_019965.1 | 4,420,197 bp | 4,103 |
| *Mycobacterium bovis strain ATCC BAA-935* | NZ_CP009449.1 | 4,358,088 bp | 4,095 |
| *Mycobacterium bovis BCG str. Tokyo 172* | NZ_CP014566.1 | 4,371,707 bp | 4,076 |

To illustrate this claim, the complete sequences of 65 *Mycobacterium* genomes, which are available on the NCBI[1] have been downloaded. Listed according to their species, 42 genomes of *tuberculosis*, 15 *bovis*, 2 *africanum*, 5 *canettii*, and 1 *microti* have been recovered. Table 1 shows information about some of these *Mycobacterium* genomes. Among this MTBC, we particularly focused on *tuberculosis* and on *canettii*, as there are enough of them, and because the virulent *tuberculosis* species is supposed to have emerged from *canettii* forty thousand years ago. To verify such an evolutionary hypothesis, the first task of our approach, proposed to achieve an ancestral reconstruction of close genomes, is to perform a multiple sequence alignment of the sequences. This task is described in the next section.

### 3.2 Multiple sequence alignment

the first stage of this alignment stage, is to identify a common starting point in these complete circular genomes. In order to do so, we searched for a reference sequence of 200 nucleotides from *M. tuberculosis H37Rv*, and we found it or its transconjugate in each genome using a local blast. Then, a circular rotation (together with a transconjugate operation if needed) has been performed on each complete genome, so that each sequence starts with the same 200 nucleotides, if we except SNPs. Once these sequences have been operated to share the same orientation and starting location, the overall alignment of each chromosome has been performed.

Alignment of large sets of sequences is a common task during biological investigations and has a wide variety of applications incorporating homology detection [38], finding evolutionarily relevant sites, and phylogenetics. A multiple sequence alignment, as depicted in Figure 2, may explain many aspects about a gene: which regions are constrained, which sites undergo positive selection [39], and potentially the

[1] ftp://ftp.ncbi.nih.gov/genomes

structure of its gene product [40]. Furthermore, aligning sequences can help to detect events of mutations or recombination in couples of close genomes, which is valuable for what we intend to do. To achieve such an alignment, we thus have considered the *AlignSeqs* function from Decipher R package [41]. Indeed, after various tests on well known alignment tools, this latter was the only one that achieved to align complete bacterial genomes with a good accuracy.

This *AlignSeqs* function takes as input two aligned sets of DNA sequences and returns a merged alignment. It can be used to achieve multiple sequence alignment in a progressive or iterative manner on sequences of the same kind. Indeed, multiple alignments are accomplished by aligning two sequences, merging with another sequence, combining with another set of sequences, and so on until all the sequences are aligned [42, 43]. We thus obtained a first representation of synteny of the whole 65 *Mycobacterium* genomes, which is depicted in Figure 3. It can be observed that these 65 genomes have a high sequence similarity with low recombination events.

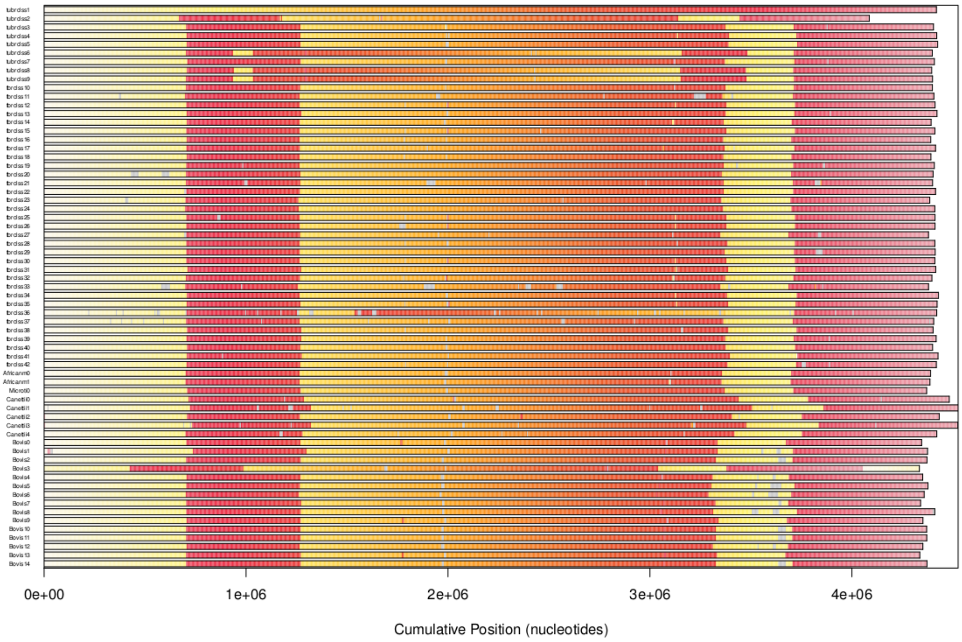


**Fig. 3.** A synteny representation of all available *Mycobacterium* strains

### 3.3 Phylogenetic study

This first representation of synteny blocks, obtained thanks to the multiple sequence alignment of the whole *Mycobacterium* genus, has allowed us to detect the location of a few large scale inversions. We thus have been able to manually invert again these inversions, so that the multiple alignment became quite perfect, if we except small indels and SNPs. It is then possible to use all the 65 complete genomes in the next stage, namely the phylogenetic study.

Indeed, the evolutionary history of our population of genomes can be represented as a phylogenetic tree using the multiple sequence alignment combined with manual local inversions previously obtained. Various methods are well established in the literature to investigate the best phylogenetic tree for a given set of aligned sequences. Well-known techniques for phylogenetic analysis include parsimony methods, maximum likelihood, distance-based methods, and even artificial intelligence based ones [44, 45]. On our side, we decided to consider the use of RAxML as a default phylogenetic tree reconstruction toolkit, a well known and reputed software based on maximum likelihood [21, 46].

As we reversed the inversions, our phylogenetic investigations are based on the whole genome. This leads to well supported and trustworthy trees of strains, on which we can reliably consider to reconstruct ancestral states. As an illustrative example, we represent the phylogenetic trees of *M. canettii* species with a relevant outgroup in Figure 4a. This very well supported tree has been obtained using RAxML with GTR Gamma model as advised by JModelTest 2.0. The *Mycobacterium tuberculosis* phylogeny, for its part, leads to bootstrap supports larger than 98%, as shown in Figure 4b.

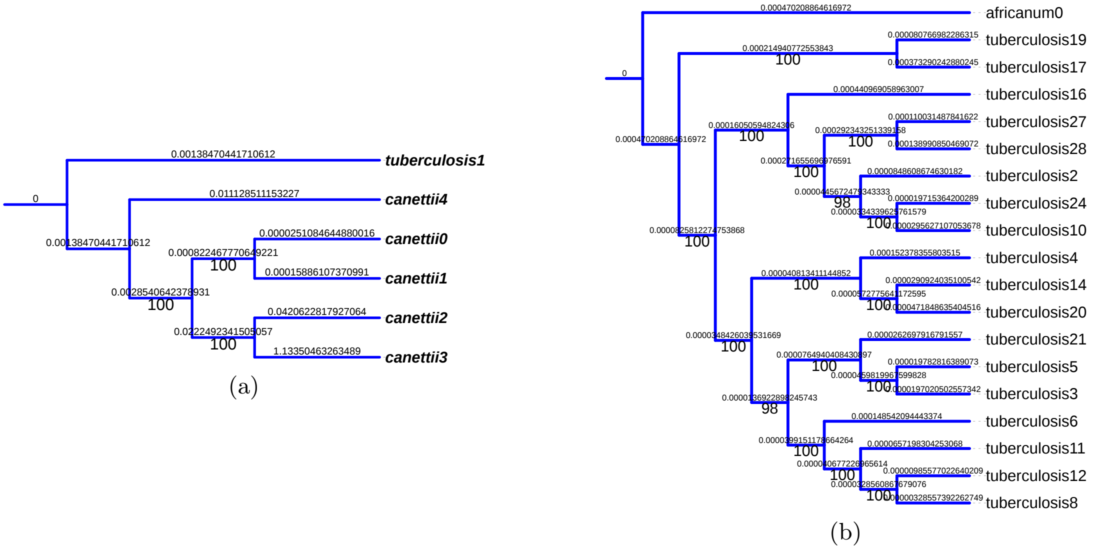


**Fig. 4.** Well-supported phylogenies: (a) *M. canettii* species using a *M. tuberculosis* as outgroup, (b) *M. tuberculosis* species with *M. africanum* as outgroup. Phylogenetic trees have been calculated on the entire genomes with RAxML and GTR Gamma model.

Having a confidential representation of the general evolution of MTBC strains due to this phylogenetic study, we are then left to reconstruct the ancestral states of the alignment at each internal node of the tree. This final ancestral reconstruction will be applied in two stages, considering first the variants of length 1 in the alignment (namely, single nucleotide polymorphism and indels of 1 nucleotide), and then larger variants that mainly consist of insertion or deletion of a subsequence at a location in the tree.

### 3.4 Ancestral reconstruction: mononucleotidic variants

Focusing on mononucleotidic variants, we separated the treatment of single nucleotide polymorphisms (SNPs) versus insertion-deletions (indels). For the former, the situation seems quite simple, the only problem being to prevent confusion between a "true" SNP and a SNP induced by a recombination of the indel kind. For the latter, future challenges encompass to determine which indels are related to tandem repeats, which are associated with mobile elements, or which are due to repeated sequences. Let us detail each case hereafter.

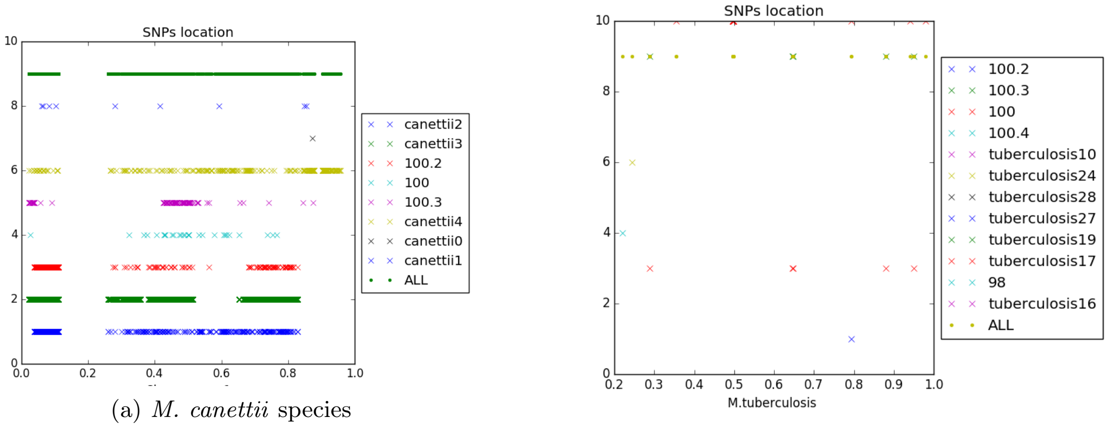
Regarding SNPs, the ancestral reconstruction is achieved as follows. The marginal probability distributions for bases at ancestral nodes in the phylogenetic tree are first calculated. These distributions are obtained using the sum-product message passing algorithm [47], assuming independence of sites. The ancestral reconstruction is done by using PHAST software [48], which reconstructs indels too by parsimony, also assuming site independence. Obtained results on mononucleotidic variants are then carefully visually checked, as the number of such variants is not excessive, see Tables 2 and 3.

At the end, 2,956 SNPs and 166 indels have been found in the alignment of the clade constituted by the 5 strains of *M. canettii*, as shown in Figure 5a. Figure 5b, for its part, represents the location of the 394 SNPs and of the 25 indels that have been found in the alignment of the clade constituted by 8 genomes of *M. tuberculosis*.

### 3.5 Ancestral reconstruction of larger variants

*Mycobacterium* species considered in this article are highly conserved, with really similar regions and without rearrangement. As previously evoked, we found only a few significant inversions, like the one at the last common ancestor of strains *CIPT 140010059*, *140070010*, *140060008*, *140070017*, and *140070008*, as shown in Figure 6a. Figure 6b, for its part, is a dotplot representing these homologous regions, as identified by the FindSynteny function in R. synteny blocks of the 42 *M. tuberculosis* are finally depicted in Figure 7, where we have obtained 99% of DNA sequence identity. To sum up, if we except a large scale inversion, we can only report some small indels at this recombination level.

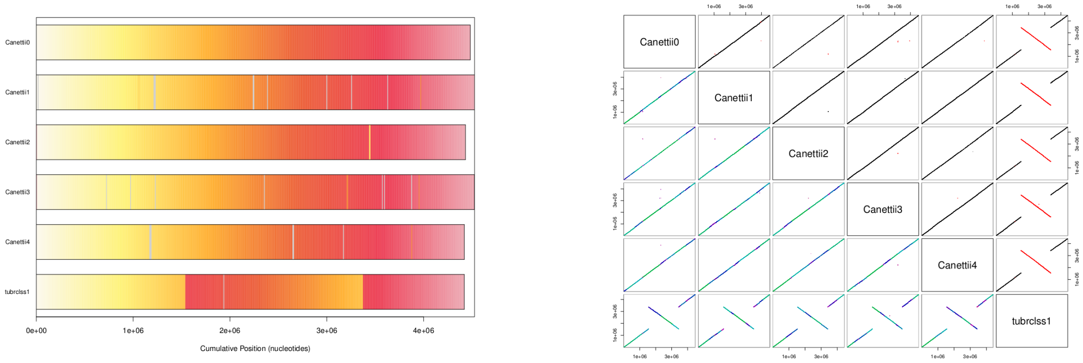
Ad hoc algorithms have then been designed to deal with mid size variants. More specifically, we have written first a string algorithm that detect small and noisy inversions, but the latter, distributed on our



(a) *M. canettii* species

(b) *M. tuberculosis* species

**Fig. 5.** SNPs location of mononucleotidic variants



(a)

(b)

**Fig. 6.** (a) synteny blocks in *canettii*. Each genome is colored according to the position of the corresponding region in the first genome (gray if a region is unshared). (b) Dot plots provide an alternative representation of the synteny map of *M. canettii*. Black diagonal lines show syntenic regions sharing the same orientation, whereas red anti-diagonal ones represent blocks of synteny between opposite strands. The description of all of these species tends to show a high sequence similarity with little recombination events.

supercomputer facilities, was only able to detect artifacts. So either the MTBC genomes have not faced inversion events during its recent history, or this recombination case still needs further investigations. Authors tend to prefer the first possibility, as *Mycobacterium* genomes evolve in a clonal manner (which is not the case, for instance, with *Yersinia* genus, in which a large amount of mobile elements has led to a large number of reported inversions) [49]. Duplication, for its part, has not yet been investigated but, as for inversions, the analysis of synteny blocks tends to show that such events are rare, at least if we consider the large scale ones.

Both indels of midsize and SNPs are rare, for its part, has been deeply studied, using PHAST software as detection tool. From obtained results, we can conclude the following points. (1) Such events are quite rare in some lineages of the MTB complex like *tuberculosis*, as described in Table 4. (2) Most of the times, the situation is very easy to understand manually, leading either to an insertion or to a deletion at an obvious internal node of the tree, as illustrated in Figure 9. (3) Most of the times, the inserted motif has not faced mutations during evolution: leaves that contain the motif have no mutation in it, thereby contributing to an easy to resolve situation. (4) Surprisingly, ancestral states recovered by PHAST and its parsimony approach leads to disappointing results. Similarly, obviously wrong results have been obtained with state-of-the-art competitor software. To sum up, a manual reconstruction of mid size indels is possible, due to the low number of these recombinations that are mainly very easy to resolve, while automatic tools from the literature are not currently able to do it.

All these steps are summarized in figure 10. In this one, gray boxes correspond to manual steps whereas all the other ones are automatically executed.

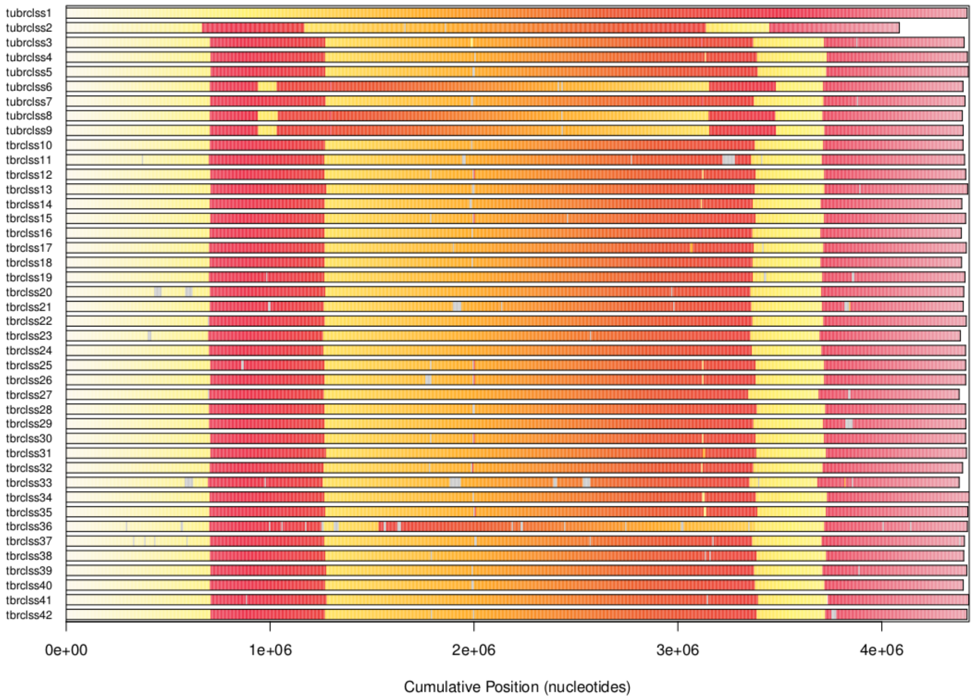


**Fig. 7.** A representation of *M. tuberculosis* genomes species tends to show more than 95% nucleotide similarity with little recombination events.

**Table 2.** Number of alignment columns with polymorphism, by pair of strains, on *M. canettii* genomes. Note that, when a large string is deleted at some location in the tree, all the characters of this deletion are counted here.

| | ***canettii0*** | ***canettii1*** | ***canettii2*** | ***canettii3*** | ***canettii4*** | ***tuberculosis1*** |
|---|---|---|---|---|---|---|
| ***canettii0*** | 0 | 3524 | 27256 | 60957 | 4833 | 3354 |
| ***canettii1*** | 3524 | 0 | 27260 | 61233 | 7971 | 1150 |
| ***canettii2*** | 27256 | 27260 | 0 | 62717 | 27468 | 27437 |
| ***canettii3*** | 60957 | 61233 | 62717 | 0 | 60987 | 61346 |
| ***canettii4*** | 4833 | 7971 | 27468 | 60987 | 0 | 7510 |
| ***tuberculosis1*** | 3354 | 1150 | 27437 | 61346 | 7510 | 0 |

**Table 3.** Variations in the alignment of *M. tuberculosis*

| | ***tuberculosis4*** | ***tuberculosis19*** | ***tuberculosis17*** | ***tuberculosis16*** | ***tuberculosis27*** | ***tuberculosis28*** | ***tuberculosis24*** | ***tuberculosis10*** |
|---|---|---|---|---|---|---|---|---|
| ***tuberculosis4*** | 0 | 199770 | 214401 | 219205 | 216387 | 217235 | 216919 | 217186 |
| ***tuberculosis19*** | 199770 | 0 | 212403 | 219039 | 216908 | 216672 | 216726 | 216953 |
| ***tuberculosis17*** | 214401 | 212403 | 0 | 216808 | 216534 | 217011 | 216786 | 216882 |
| ***tuberculosis16*** | 219205 | 219039 | 216808 | 0 | 216669 | 216916 | 216251 | 216678 |
| ***tuberculosis27*** | 216387 | 216908 | 216534 | 216669 | 0 | 142974 | 189148 | 199505 |
| ***tuberculosis28*** | 217235 | 216672 | 217011 | 216916 | 142974 | 0 | 189460 | 199412 |
| ***tuberculosis24*** | 216919 | 216726 | 216786 | 216251 | 189148 | 189460 | 0 | 194315 |
| ***tuberculosis10*** | 217186 | 216953 | 216882 | 216678 | 199505 | 199412 | 194315 | 0 |

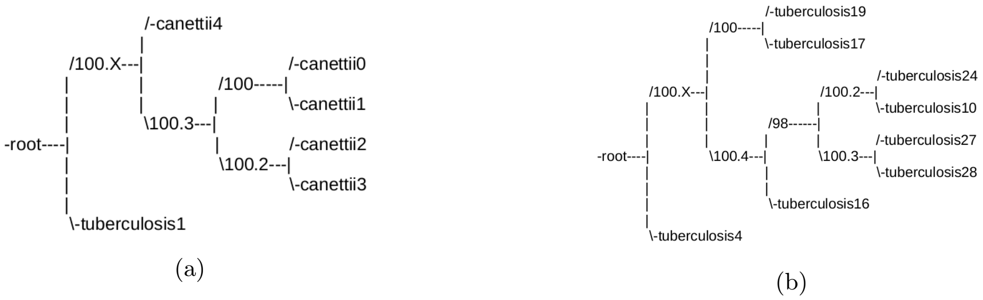


**Fig. 8.** Example of phylogenetic tree, show the ancestor nodes (internal node), and the relation with their children (a) *M. canettii* species, (b) *M. tuberculosis* species.

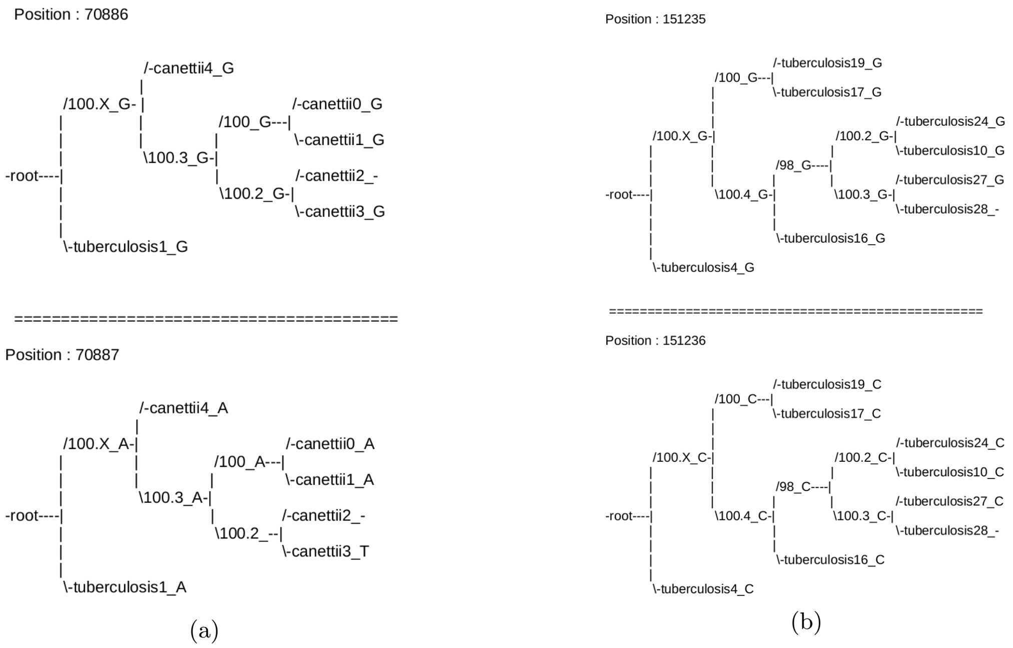


**Fig. 9.** The insertions and deletions of nucleotides (indels) on the internal node of the tree (a) represent the nucleotides contain the ancestor nodes and their children on *M. canettii* species, (b) *M. tuberculosis* species.

*M. canettii* SNPs

| Fathers | Children | No. of SNPs |
|---|---|---|
| *100.2* | *canettii2* | 1041 |
| | *canettii3* | 12398 |
| *100* | *canettii0* | 1 |
| | *canettii1* | 9 |
| *100.3* | *100* | 28 |
| | *100.2* | 735 |
| *100.X* | *100.3* | 111 |
| | canettii4 | 438 |

*M. tuberculosis* SNPs

| Fathers | Children | No. of SNPs |
|---|---|---|
| *100* | *tuberculosis19* | 5 |
| | *tuberculosis17* | 14 |
| *100.2* | *tuberculosis24* | 1 |
| | *tuberculosis10* | 0 |
| *100.3* | *tuberculosis27* | 0 |
| | *tuberculosis28* | 0 |
| *98* | *100.2* | 1 |
| | *100.3* | 0 |
| *100.4* | *98* | 0 |
| | *tuberculosis16* | 1 |
| *100.X* | 100 | 5 |
| | 100.4 | 1 |

**Table 4.** Number of SNPs in the considered species (100.X refers to an ancestral node, as in the tree)

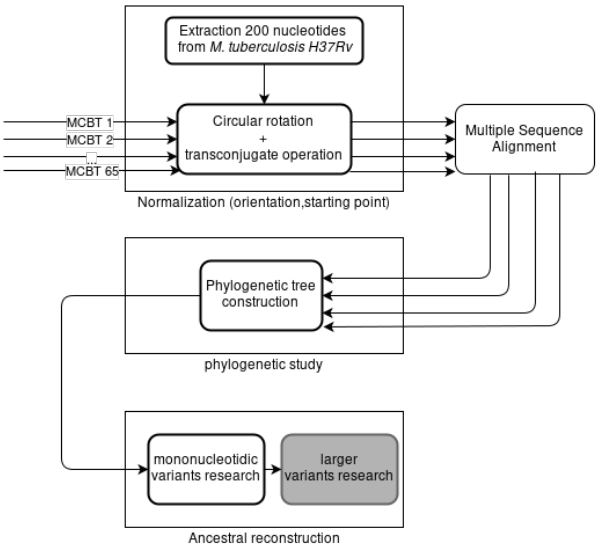

Extraction 200 nucleotides from *M. tuberculosis H37Rv*
MCBT 1
MCBT 2
...
MCBT 65
Circular rotation + transconjugate operation
Normalization (orientation,starting point)
Multiple Sequence Alignment
Phylogenetic tree construction
phylogenetic study
mononucleotidic variants research
larger variants research
Ancestral reconstruction

**Fig. 10.** Flowchart of the proposed approach.

## 4 Discussion

The obtained ancestors have not yet been studied in this work-in-progress. They will be investigated with updated and improved algorithms, encompassing mobile elements and gene content evolution analyzes.

Indeed, an important category of genome modification is the loss of functional genes, for instance because they become ineffective. In order to do so, we will consider the phylogenetic tree whose leaves will contain sets of genes, and we will compute core and pan genomes at each internal node of the tree. Having this core and pan tree, we will design an algorithm to investigate more deeply the evolution of these pan and core genomes over the tree, to see if some branches can be related to hot spots of evolution. We thus intend to determine at which rate such loss or gain occur, and which kinds of functionality are concerned. We will finally compute how much mutations fall inside a coding sequence, by studying which kind of genes has evolved on the phylogenetic tree, by wondering if the mutation rate has changed over time, and if such mutability can be related to environmental events. In other words, we will wonder which variations have been potentially significant among the numerous events that have been found when aligning these sequences.

With such a pipeline, we intend to investigate the following questions. Are some recombinations at the origin of severe tuberculosis epidemics? Are transposases responsible of such recombinations like inversions [50, 51]? Are transposases in general more present in *M. tuberculosis* (affecting humans) than in *M. africanum*, *M. bovis*, or *M. bovis BCG*? Are they related to the virulence of the strain? How core and pan genomes have evolved over time in this complex? Finally, we will compare the last common ancestor of this complex to a *M. canettii*, to see if the *canettii* ancestor hypothesis can be verified by the ancestral reconstruction way.

At this point, our partial conclusion is that the reconstruction of ancestral sequences is possible, at least in the case of close and clonal bacterias. Furthermore, elements being part of this reconstruction have already be designed, at least in their first revision (for instance to detect and deal with mononucleotidic variants). However, the MTB complex seems to be a little too complicated for a first deep investigation of semi-automatic reconstruction of ancestral sequences of bacteria, and a genus like *Brucella* may be more easy to deal with in a first concrete investigation of this problem.

## 5 Conclusion

In this article, we have firstly emphasized that, even if various algorithms and software already exist to face the NP-hard character of the ancestral genome reconstruction problem, they do not work perfectly, in particular when SNPs or indels fall into repeated sequences. We have then argued that, when regarding the relatively low number of mutation and recombination events in such *Mycobacterium*, a pragmatic approach is possible. We have proposed to reconstruct all ancestors of all complete available genomes of *Mycobacterium tuberculosis* and of *M. canettii*. The study has started by investigating single nucleotide polymorphism level, while indels and large scale recombination are regarded in a second stage. Our conclusion is that, by mixing automatic reconstruction of obvious situations with human interventions on signaled problematic cases, it may be possible to achieve a concrete, complete, and really accurate reconstruction of some specific bacteria lineages. We can thus investigate how these genomes have evolved from their last common ancestors.

In future work, we intend to reconstruct all ancestors of all complete available genomes of specific bacteria strains, namely, and ordered by complexity: *Brucella* genus, *Yersinia pestis*, and *Pseudomonas aeruginosa*. Moreover, we intend to compare them with ancient DNA when available (like for *Y. pestis*). In parallel, original mathematical description of some recombination mechanisms will be proposed, encompassing branching process and partial differential equation approaches for modeling mobile elements. Finally, we may try to correlate the evolutionary history of microorganisms to epidemiological data: events of genomic recombination may be related to epidemic outbreaks. And such putative correlations may be learnt by deep learning algorithms, leading to a new way to predict epidemic risks.

## Acknowledgments

Computations presented in this article were realised on the supercomputing facilities provided by the Mésocentre de calcul de Franche-Comté.

## References

1. Noel H Smith, Stephen V Gordon, Ricardo de la Rua-Domenech, Richard S Clifton-Hadley, and R Glyn Hewinson. Bottlenecks and broomsticks: the molecular evolution of mycobacterium bovis. *Nature Reviews Microbiology*, 4(9):670–681, 2006.

2. IC Shamputa, Cho SangNae, J Lebron, LE Via, H Mukundan, MA Chambers, WR Waters, MH Larsen, et al. Introduction and epidemiology of mycobacterium tuberculosis complex in humans. *Tuberculosis, leprosy and mycobacterial diseases of man and animals: the many hosts of mycobacteria*, pages 1–16, 2015.
3. Roland Brosch, Stephen V Gordon, M Marmiesse, P Brodin, C Buchrieser, K Eiglmeier, T Garnier, C Gutierrez, G Hewinson, K Kremer, et al. A new evolutionary scenario for the mycobacterium tuberculosis complex. *Proceedings of the national academy of Sciences*, 99(6):3684–3689, 2002.
4. Michaela M Gutacker, James C Smoot, Cristi A Lux Migliaccio, Stacy M Ricklefs, Su Hua, Debby V Cousins, Edward A Graviss, Elena Shashkina, Barry N Kreiswirth, and James M Musser. Genome-wide analysis of synonymous single nucleotide polymorphisms in mycobacterium tuberculosis complex organisms: resolution of genetic relationships among closely related microbial strains. *Genetics*, 162(4):1533–1543, 2002.
5. Serge Mostowy, Debby Cousins, Jacqui Brinkman, Alicia Aranaz, and Marcel A Behr. Genomic deletions suggest a phylogeny for the mycobacterium tuberculosis complex. *Journal of infectious Diseases*, 186(1):74–80, 2002.
6. Makiko Yamada-Noda, Kiyofumi Ohkusu, Hiroyuki Hata, Mohammad Monir Shah, Pham Hong Nhung, Xiao Song Sun, Masahiro Hayashi, and Takayuki Ezaki. Mycobacterium species identification–a new approach via dnaj gene sequencing. *Systematic and applied microbiology*, 30(6):453–462, 2007.
7. Michel Fabre, Yolande Hauck, Charles Soler, Jean-Louis Koeck, Jakko Van Ingen, Dick Van Soolingen, Gilles Vergnaud, and Christine Pourcel. Molecular characteristics of "mycobacterium canettii" the smooth mycobacterium tuberculosis bacilli. *Infection, Genetics and Evolution*, 10(8):1165–1173, 2010.
8. RD Fleischmann, D Alland, Jonathan A Eisen, L Carpenter, O White, J Peterson, R DeBoy, R Dodson, M Gwinn, D Haft, et al. Whole-genome comparison of mycobacterium tuberculosis clinical and laboratory strains. *Journal of bacteriology*, 184(19):5479–5490, 2002.
9. Thierry Wirth, Falk Hildebrand, Caroline Allix-Béguec, Florian Wölbeling, Tanja Kubica, Kristin Kremer, Dick van Soolingen, Sabine Rüsch-Gerdes, Camille Locht, Sylvain Brisse, et al. Origin, spread and demography of the mycobacterium tuberculosis complex. *PLoS Pathog*, 4(9):e1000160, 2008.
10. Gregory I Lang and Andrew W Murray. Estimating the per-base-pair mutation rate in the yeast saccharomyces cerevisiae. *Genetics*, 178(1):67–82, 2008.
11. Guillaume Fertin. *Combinatorics of genome rearrangements*. MIT press, 2009.
12. Jian Ma, Aakrosh Ratan, Brian J Raney, Bernard B Suh, Louxin Zhang, Webb Miller, and David Haussler. Dupcar: reconstructing contiguous ancestral regions with duplications. *Journal of computational biology*, 15(8):1007–1027, 2008.
13. Yves Gagnon, Mathieu Blanchette, and Nadia El-Mabrouk. A flexible ancestral genome reconstruction method based on gapped adjacencies. *BMC bioinformatics*, 13(Suppl 19):S4, 2012.
14. Bradley R Jones, Ashok Rajaraman, Eric Tannier, and Cedric Chauve. Anges: reconstructing ancestral genomes maps. *Bioinformatics*, 28(18):2388–2390, 2012.
15. Jian Ma, Louxin Zhang, Bernard B Suh, Brian J Raney, Richard C Burhans, W James Kent, Mathieu Blanchette, David Haussler, and Webb Miller. Reconstructing contiguous regions of an ancestral genome. *Genome research*, 16(12):1557–1565, 2006.
16. Fei Hu, Jun Zhou, Lingxi Zhou, and Jijun Tang. Probabilistic reconstruction of ancestral gene orders with insertions and deletions. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(4):667–672, 2014.
17. Mathauieu Blanchette, Abdoulaye Baniré Diallo, Eric D Green, Webb Miller, and David Haussler. Computational reconstruction of ancestral dna sequences. In *Phylogenomics*, pages 171–184. Springer, 2008.
18. Virginie Lopez Rascol, Pierre Pontarotti, and Anthony Levasseur. Ancestral animal genomes reconstruction. *Current opinion in immunology*, 19(5):542–546, 2007.
19. Bret Larget, Donald L Simon, Joseph B Kadane, and Deborah Sweet. A bayesian analysis of metazoan mitochondrial genome arrangements. *Molecular Biology and Evolution*, 22(3):486–495, 2005.
20. Sridhar Hannenhalli, Colombe Chappey, Eugene V Koonin, and Pavel A Pevzner. Genome sequence comparison and scenarios for gene rearrangements: A test case. *Genomics*, 30(2):299–311, 1995.
21. Alexandros Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.
22. Remco Bouckaert, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A Suchard, Andrew Rambaut, and Alexei J Drummond. Beast 2: a software platform for bayesian evolutionary analysis. *PLoS Comput Biol*, 10(4):e1003537, 2014.
23. Ziheng Yang. Phylogenetic analysis by maximum likelihood (paml), 2000.
24. Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics*, 20(2):289–290, 2004.
25. Alexandre Bouchard-Côté and Michael I Jordan. Evolutionary inference via the poisson indel process. *Proceedings of the National Academy of Sciences*, 110(4):1160–1166, 2013.
26. GA Watterson, Warren J Ewens, Thomas Eric Hall, and A Morgan. The chromosome inversion problem. *Journal of Theoretical Biology*, 99(1):1–7, 1982.
27. Shimon Even and Oded Goldreich. The minimum-length generator sequence problem is np-hard. *Journal of Algorithms*, 2(3):311–313, 1981.

28. Bettina Stoebe, William Martin, and Klaus V Kowallik. Distribution and nomenclature of protein-coding genes in 12 sequenced chloroplast genomes. *Plant Molecular Biology Reporter*, 16(3):243–255, 1998.
29. Daniel Grzebyk, Oscar Schofield, Costantino Vetriani, and Paul G Falkowski. The mesozoic radiation of eukaryotic algae: The portable plastid hypothesis1. *Journal of Phycology*, 39(2):259–267, 2003.
30. Itai Sharon, Ariella Alperovitch, Forest Rohwer, Matthew Haynes, Fabian Glaser, Nof Atamna-Ismaeel, Ron Y Pinter, Frédéric Partensky, Eugene V Koonin, Yuri I Wolf, et al. Photosystem i gene cassettes are present in marine virus genomes. *Nature*, 461(7261):258–262, 2009.
31. Matteo De Chiara, Derek Hood, Alessandro Muzzi, Derek J Pickard, Tim Perkins, Mariagrazia Pizza, Gordon Dougan, Rino Rappuoli, E Richard Moxon, Marco Soriani, et al. Genome sequencing of disease and carriage isolates of nontypeable haemophilus influenzae identifies discrete population structure. *Proceedings of the National Academy of Sciences*, 111(14):5439–5444, 2014.
32. Stefan Kurtz, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L Salzberg. Versatile and open software for comparing large genomes. *Genome biology*, 5(2):1, 2004.
33. Marie Touchon, Claire Hoede, Olivier Tenaillon, Valérie Barbe, Simon Baeriswyl, Philippe Bidet, Edouard Bingen, Stéphane Bonacorsi, Christiane Bouchier, Odile Bouvet, et al. Organised genome dynamics in the escherichia coli species results in highly diverse adaptive paths. *PLoS genet*, 5(1):e1000344, 2009.
34. Robert Boissy, Azad Ahmed, Benjamin Janto, Josh Earl, Barry G Hall, Justin S Hogg, Gordon D Pusch, Luisa N Hiller, Evan Powell, Jay Hayes, et al. Comparative supragenomic analyses among the pathogens staphylococcus aureus, streptococcus pneumoniae, and haemophilus influenzae using a modification of the finite supragenome model. *BMC genomics*, 12(1):1, 2011.
35. Hervé Tettelin, Vega Masignani, Michael J Cieslewicz, Claudio Donati, Duccio Medini, Naomi L Ward, Samuel V Angiuoli, Jonathan Crabtree, Amanda L Jones, A Scott Durkin, et al. Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13950–13955, 2005.
36. Benoît Valot, Christophe Guyeux, Julien Yves Rolland, Kamel Mazouzi, Xavier Bertrand, and Didier Hocquet. What it takes to be a pseudomonas aeruginosa? the core genome of the opportunistic pathogen updated. *PloS one*, 10(5):e0126468, 2015.
37. Jialiang Yang, Jun Li, Liuhuan Dong, and Stefan Grünewald. Analysis on the reconstruction accuracy of the fitch method for inferring ancestral states. *BMC bioinformatics*, 12(1):18, 2011.
38. Yong Wang, Ruslan I Sadreyev, and Nick V Grishin. Procain server for remote protein sequence similarity search. *Bioinformatics*, 25(16):2076–2077, 2009.
39. Carsten Kemena and Cedric Notredame. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*, 25(19):2455–2465, 2009.
40. Tandy Warnow. Large-scale multiple sequence alignment and phylogeny estimation. In *Models and Algorithms for Genome Evolution*, pages 85–146. Springer, 2013.
41. R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2014.
42. Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):1, 2004.
43. Erik S Wright. The art of multiple sequence alignment in r. 2014.
44. Bassam Alkindy, Christophe Guyeux, Jean-François Couchot, Michel Salomon, and Jacques Bahi. Using genetic algorithm for optimizing phylogenetic tree inference in plant species. In *MCEB15, Mathematical and Computational Evolutionary Biology*, Porquerolles Island, France, June 2015.
45. Bassam Alkindy, Bashar Al-Nuaimi, Christophe Guyeux, Jean-François Couchot, Michel Salomon, Reem Alsrraj, and Laurent Philippe. Binary particle swarm optimization versus hybrid genetic algorithm for inferring well supported phylogenetic trees. In *International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics*, pages 165–179. Springer, 2015.
46. Bassam AlKindy, Christophe Guyeux, Jean-François Couchot, Michel Salomon, Christian Parisod, and Jacques M Bahi. Hybrid genetic algorithm and lasso test approach for inferring well supported phylogenetic trees based on subsets of chloroplastic core genes. In *International Conference on Algorithms for Computational Biology*, pages 83–96. Springer, 2015.
47. Judea Pearl. Reverend bayes on inference engines: A distributed hierarchical approach. In *AAAI*, pages 133–136, 1982.
48. Melissa J Hubisz, Katherine S Pollard, and Adam Siepel. Phast and rphast: phylogenetic analysis with space/time models. *Briefings in bioinformatics*, page bbq072, 2010.
49. Marcel A Behr. Evolution of mycobacterium tuberculosis. In *The New Paradigm of Immunity to Tuberculosis*, pages 81–91. Springer, 2013.
50. Patricia Siguier, Jonathan Filée, and Michael Chandler. Insertion sequences in prokaryotic genomes. *Current opinion in microbiology*, 9(5):526–531, 2006.
51. Casey M Bergman and Hadi Quesneville. Discovering and detecting transposable elements in genome sequences. *Briefings in bioinformatics*, 8(6):382–392, 2007.