FPSAC: fast phylogenetic scaffolding of ancient contigs

Ashok Rajaraman[1,2], Eric Tannier[3,4] and Cedric Chauve[1,5,*]

[1]Department of Mathematics, Simon Fraser University, Burnaby (BC) V5A1S6, Canada, [2]International Graduate Training Center in Mathematical Biology, Pacific Institute for the Mathematical Sciences, Vancouver (BC), Canada, [3]INRIA Grenoble Rhône-Alpes, Montbonnot 38334, France, [4]Université de Lyon 1, Laboratoire de Biométrie et Biologie Évolutive, CNRS UMR5558 F-69622 Villeurbanne, France and [5]LaBRI, Université Bordeaux I, 33405 Talence, France

ABSTRACT

Motivations: Recent progress in ancient DNA sequencing technologies and protocols has lead to the sequencing of whole ancient bacterial genomes, as illustrated by the recent sequence of the *Yersinia pestis* strain that caused the Black Death pandemic. However, sequencing ancient genomes raises specific problems, because of the decay and fragmentation of ancient DNA among others, making the scaffolding of ancient contigs challenging.

Results: We show that computational paleogenomics methods aimed at reconstructing the organization of ancestral genomes from the comparison of extant genomes can be adapted to correct, order and orient ancient bacterial contigs. We describe the method FPSAC (fast phylogenetic scaffolding of ancient contigs) and apply it on a set of 2134 ancient contigs assembled from the recently sequenced Black Death agent genome. We obtain a unique scaffold for the whole chromosome of this ancient genome that allows to gain precise insights into the structural evolution of the *Yersinia* clade.

Availability and Implementation: Code, data and results are available at http://paleogenomics.irmacs.sfu.ca/FPSAC.

Contact: cedric.chauve@sfu.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Palaeomicrobiology aims at analyzing ancient microorganisms, especially pathogens obtained from the remains of infected hosts (Donoghue and Spigelman, 2006; Drancourt and Raoult, 2005). Aside of an historical interest in characterizing precisely past infectious diseases (Drancourt, 2012), understanding the evolution of pathogens and their relation with their hosts is of primary interest for modern microbiology (Donoghue, 2011; Wilson, 2012). Initially based on polymerase chain reaction techniques, ancient DNA (aDNA) sequencing benefited from advances in sequencing technologies and the development of new protocols, that lead to breakthroughs, such as the sequencing of whole molecules from the *Yersinia pestis* strain that caused the Black Death pandemic, including a plasmid (Schuenemann *et al.*, 2011) and the main chromosome (Bos *et al.*, 2011).

Recently, Bos *et al.* (2011) extracted >10 million short single reads (average length of 53 nt) from the dental pulp of an individual infected by the Black Death pathogen, using the genome of an extant *Y.pestis* strain (CO92) as a bait. They were assembled, using Velvet (Zerbino and Birney, 2008), into >130 000 contigs, including 2134 contigs of length ≥500 nt from the main chromosome of the Black Death agent. This first sequencing of the chromosome of an extinct prokaryote helped to clarify the causes of the Black Death pandemic (Bos *et al.*, 2011; Parkhill and Wren, 2011; Wilson, 2012). However, the 2134 larger contigs cover only 85% of the expected length of the ancestral chromosome and their organization along this ancestral chromosome is still unknown, keeping out of reach a detailed genome-scale study of the evolution of the structural organization of *Yersinia* genomes, whose impact on pathogenicity is still an open question (Chain *et al.*, 2004).

Current scaffolding methodologies can hardly be applied to fully assemble and finish an ancient bacterial genome from a dataset such as the one described by Bos *et al.* (2011), aside of short molecules like plasmids (Schuenemann *et al.*, 2011). These methods, aimed at ordering and orienting the contigs, and estimating the lengths of inter-contig gaps, rely on data such as mate-pair libraries with mixed insert sizes (Bashir *et al.*, 2012; Chapman *et al.*, 2011; Donmez and Brudno, 2013; Gao *et al*, 2011; Ribeiro *et al.*, 2012; Salmela and *et al.*, 2011), genome maps (Lin *et al.*, 2012) or comparison with one or several closely related genomes (Kim *et al.*, 2013; Gnerre *et al.*, 2009). Owing to the decay and fragmentation of aDNA molecules [whose length depends on many factors, but that can be as short as 300 nt (Drancourt and Raoult, 2005)], reads from ancient genomes are expected to be short, and genome maps or mate-pair libraries with long inserts are not available. This leaves the comparative approach as the only possibility. The usual setting of the comparative approach involves the comparison of the contigs with one, or a few, closely related genomes sequences or maps (Husemann and Stoye, 2010; Munoz *et al.*, 2010; Rissman *et al.*, 2009). For an ancestral genome, comparison with a single reference genome, either a descendant or an outgroup, is likely to predict derived syntenic features as ancestral (Rissman *et al.*, 2009), which is a problem for genomes such as the *Y.pestis* genomes that contains many repeats and are highly rearranged (Darling *et al.*, 2008). There exists only one scaffolding method that allows to compare with several related genomes while using a phylogenetic tree (Husemann and Stoye, 2010), but it is not designed to scaffold an ancient genome. We address this specific

---

*To whom correspondence should be addressed

problem here and describe a phylogenetic approach to scaffold ancient bacterial contigs that adapts existing methods initially designed to predict ancestral genome features from the comparison of extant genomes.

The design of predictive methods to reconstruct ancestral genomic features is a relatively ancient field of computational genomics, dating back to methods such as Fitch's algorithm for reconstructing ancestral genomic sequences (Fitch, 1971). Advances in computational paleogenomics include improved methods for reconstructing ancestral genome sequence (Blanchette *et al.*, 2004; Diallo *et al.*, 2010; Liberles, 2007), gene content (Cohen *et al.*, 2010; Csurös, 2010; Szöllősi *et al.*, 2012) and gene order. The latter ones have been used for reconstructing ancestral genomes organization of bacteria (Fremez *et al.*, 2007; Wang *et al.*, 2006), animals (Alekseyev and Pevzner, 2009; Chauve and Tannier, 2008; Ma *et al.*, 2006; Muffato *et al.*, 2010; Ouangraoua *et al.*, 2011; Putnam *et al.*, 2007), plants (Murat *et al.*, 2010; Sankoff *et al.*, 2009), yeasts (Bertrand *et al.*, 2010; Gordon *et al.*, 2009) or protists (Ma *et al.*, 2008). Recent developments provide exact and fast algorithms that handle repeats as well as diverse types of genome rearrangements and chromosome structures (Bérard *et al.*, 2012; Chauve *et al.*, 2013; Jones *et al.*, 2012; Manuch *et al.*, 2012).

We describe here how to adapt and integrate some of these methods to process ancient bacterial contigs. We apply our method to the Black Death agent genome, using the genomes of eleven closely related descendants and outgroups from the *Y.pestis* and *Yersinia pseudotuberculosis* clades, whose phylogeny is given in Bos *et al.*, (2011): we correct, order and orient the ancient contigs of the of the medieval Black Death agent chromosome into a single scaffold and estimate the inter-contigs DNA sequences, and we describe a preliminary analysis of this reconstructed ancestral genome.

2 METHODS AND ALGORITHMS

We are given a set of contig sequences for an ancestral genome A, together with a set of related extant sequenced genomes, descendants and outgroups of A, organized into a phylogenetic tree T. Our scaffolding method FPSAC relates to a generic scheme for reconstructing ancestral genome organization (Chauve and Tannier, 2008; Ma *et al.*, 2006; Jones *et al.*, 2012), and is composed of four phases:
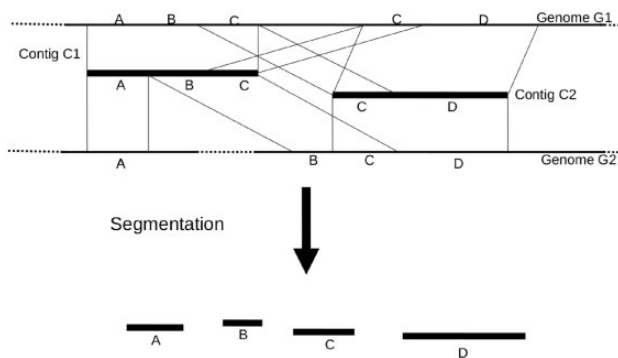
(1) *Computing homologous families*. A homologous family is composed of at least one contig segment (*ancestral marker*) and several non-overlapping extant genomes segments (*extant markers*), that pairwise align, with high similarity, along their whole length. Each homologous family is assigned a *multiplicity* bounding the number of occurrences (copy number) of ancestral marker(s) from this family in the ancestral genome A.

(2) *Computing putative ancestral adjacencies*. An ancestral adjacency is composed of two ancestral markers that are believed to have been contiguous in A. We predict them using a Dollo parsimony principle that takes advantage of the internal position of A in the considered phylogenetic tree. All adjacencies are weighted according to their phylogenetic conservation, defining a weighted *adjacency graph*.

(3) *Scaffolding from ancestral adjacencies*. If the set of all ancestral adjacencies is not compatible with a multichromsomal circular chromosomal structure that respects the multiplicity constraints of homologous families, we compute a maximum weight subset of adjacencies that is compatible with such a circular chromosomal structure. Next, as adjacencies alone can define several contig orders, due to repeated ancestral markers forming tangles in the adjacency graph, conserved intervals spanning repeats are used to clear the ambiguities, in a way similar to the use of mate-pairs to scaffold extant genomes.

(4) *Estimating inter-markers gap lengths and sequences*. For each ancestral adjacency, the length of the ancestral gap between the two involved markers is estimated from the length of the gap between the corresponding extant adjacencies (extant gaps). The sequences of the extant gaps whose length agrees with the estimated ancestral gap length are aligned into a multiple sequence alignment that is used to reconstruct a putative ancestral gap sequence.

2.1 Computing homologous markers families

We map the ancient contigs onto the extant genomes. Every significant hit (defined here by a length of at least 100 nt with 95% of identity) indicates two homologous sequences, one located on a contig and one located on an extant genome. Owing to rearrangements and repeats, some contigs do not align over their whole length to every extant genome, indicating potential evolutionary breakpoints. To detect families of homologous segments, we apply an iterative segmentation procedure, which produces contig and extant genome segments such that (1) contig segments align over their whole length to extant genomes segments and (2) pairs of extant genome segments do not overlap (*i.e.* either they have the same coordinates, or they are completely disjoint).

From a set of pairwise contigs/genomes alignments, we cut the contigs and the corresponding extant genome segments if either (1) or (2) is not satisfied. Assume first that (1) is violated: there is a segment $[a, b]$ from a contig of length ℓ that aligns to an extant genome, with $a > 1$ or $b < \ell$ or both. We assume that $a > 1$; the other case is treated symmetrically. The contig is cut into two segments, with coordinates $[1, a - 1]$ and $[a, \ell]$ and the corresponding genome segments are cut accordingly. All others alignments of segments from this contig overlapping coordinate a are also cut into two subsegments at this position in the same way as previously mentioned. We iteratively apply this procedure until (1) is verified for all pairwise alignments, thus defining a new set of pairwise contigs/genomes alignments. Next, assume that (2) is violated: two different contigs have segments aligning to two overlapping regions of an extant genome, say $[a, c]$ and $[b, d]$, with $a < b < c < d$. In this case, the two contigs are cut into two segments so that the four resulting segments align to genome segments with coordinates $[a, b]$, $[b, c]$ (for two of them) and $[c, d]$ (see Fig. 1). After iteratively applying this procedure until (2) is satisfied, it is possible that (1) is violated again. To make the procedure converge, we remove short alignments (below the length threshold used to define significant hits) and repeat the

Fig. 1. Illustration of the segmentation procedure to obtain homologous families of markers. For this example, we consider two contigs C1 and C2 and their alignments on two genomes G1 and G2. Part C of C1 and C2 aligns to the same positions in both genomes, including two different positions on G1. Parts A and B of C1 align at two different positions of G2. So the segmentation produces four families, with non-overlapping ancestral markers A, B, C and D. For these four segments, properties (1) and (2) are satisfied, whereas both were violated for C1 and C2. The family containing segment C contains two ancestral segments, two extant segments from G1 and one from G2. According to the number of occurrences in other genomes, this family may have a multiplicity >1.

two procedures until (1) and (2) are both satisfied. Then all aligned sequences naturally cluster into sets of highly similar ancient and extant sequences forming homologous families.

2.2 Multiplicity of homologous families

Next, we assign to each homologous family a *multiplicity* that is the expected number of occurrences of the ancestral marker of the family in the ancestral genome. The multiplicity of a family is computed from the number of occurrences of the extant markers in the extant genomes (the family profile) to minimize the number of evolutionary gain/loss along the branches of the considered phylogenetic tree. It is computed by a linear time dynamic programming algorithm [see Csürös (2010) for example].

2.3 Computing ancestral adjacencies

To account for the orientation of markers in predicted ancestral syntenic features (adjacencies and intervals), we decompose each marker (ancestral or extant) into two marker extremities, its head and its tail, a standard approach in genome rearrangement studies (Chauve *et al.*, 2010).

Adjacencies are then defined in terms of marker extremities instead of markers, and are computed following a Dollo parsimony principle described in Chauve and Tannier (2008): two ancestral marker extremities form an ancestral adjacency if they are contiguous (no other marker is between them in the chromosome) in at least two extant genomes whose evolutionary path in T contains A.

Adjacencies are weighted according to their patterns of phylogenetic conservation as described in Ma *et al.* (2006) [see also Chauve and Tannier (2008)]. The weighted adjacency graph is defined as follows: its vertices are the markers extremities and its edges are the weighted adjacencies.

2.4 Computing ancestral scaffolds

An ancestral scaffold is a linear or circular order of ancestral markers. The set of ancestral adjacencies might not translate into an unambiguous set of ancestral scaffolds for two reasons: (1) there might not exist a set of circular or linear markers orders that contain all adjacencies and respect the multiplicity of each marker, and (2) even if ancestral adjacencies can be organized in ancestral scaffolds, several sets of scaffolds can exist because of marker multiplicities (Fig. 2).

To address point (1), we compute a maximum weight subset of ancestral adjacencies such that every marker extremity belongs to a number of adjacencies that is at most the multiplicity of the marker family (Wittler *et al.*, 2011; Manuch *et al.*, 2012): for an ancestral marker of multiplicity m, each of its extremities can belong to at most m ancestral adjacencies. Such a selected subset of ancestral adjacencies, that is computed in polynomial time Manuch *et al.* (2012), is compatible with an order of the markers into a set of linear and/or circular scaffolds which respects the copy number constraint given by the ancestral marker multiplicities.

It is important to note that, although bacterial genomes can be composed of several circular molecules (chromosomes and plasmids), the algorithm we use does not control the resulting chromosomal structure (in terms of the number of scaffolds and of their linearity/circularity). The problem of computing a maximum weight subset of adjacencies that can be realized into a constrained chromosomal structure is NP-hard, as it includes the Maximum Weight Path Cover Problem (Ma *et al.*, 2008). Relaxing the constraints on the chromosomal structure leads to a tractable problem (Manuch *et al.*, 2012); moreover, if the resulting adjacencies can be realized into a set of linear segments, then this defines an optimal solution to the Maximum Weight Path Cover Problem, and so, an optimal set of scaffolds.

To address point (2), we rely on *conserved intervals* that span markers with multiplicity >1 (see Fig. 2 for an illustration of this principle). More precisely, we define a *repeat cluster* as a maximal connected subgraph of the adjacency graph induced by extremities of ancestral markers with multiplicity >1. A *repeat spanning interval* of R in a given genome G is a sequence of markers in G of the form $a\, x_1 \dots x_k\, b$ such that the multiplicity of a and b is 1 and the x_i's all belong to the repeat cluster R. A repeat spanning interval is conserved if it appears, up to a complete reversal, in two genomes whose evolutionary path in T contains A. Identifying all conserved repeat spanning intervals can be done in time linear in the total size of all repeat clusters. Next, repeat spanning intervals are weighted using the same method as ancestral adjacencies, and for each repeat cluster R, we greedily select repeat spanning intervals that are both compatible with the adjacencies selected during the previous step, which contain markers of R, and satisfy the multiplicity constraints of the markers of R (Chauve *et al.*, 2013).

Provided all repeats are spanned by enough conserved intervals, this results into an unambiguous scaffolding that includes all ancestral markers, including repeated ones. Otherwise, this means that the evolutionary signal present in the considered extant genomes is not sufficient to resolve repeats in the ancestral genome, in which case, adjacencies composed of two repeats that
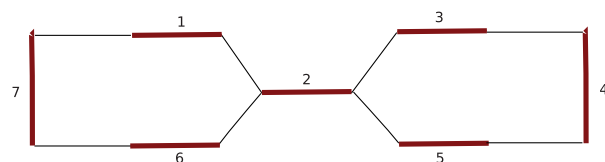
Fig. 2. Illustration of the ambiguity in ordering ancestral markers with multiplicities >1 and of the use of intervals to address it. Here is a toy example where we have markers 1, ..., 7, drawn with bold segments, and adjacencies between their extremities, drawn with thin lines. Assume every marker has multiplicity 1 except marker 2, which has multiplicity 2. Then every marker extremity has as many adjacencies as its multiplicity predicts. But there are two possible circular orderings of these markers according to these adjacencies: 1,2,3,4,5,2,6,7, or 1,2,5,4,3,2,6,7. Suppose we have in addition repeat spanning intervals 1.2.3 and 5.2.6, then only the first ordering is compatible with them

Fig. 3. Phylogeny of the considered genomes from Bos *et al.* (2011)

do not belong to a repeat spanning interval are discarded, resulting in a more fragmented, but unambiguous, set of scaffolds.

2.5 Estimating inter-contig gaps lengths and sequences

An *ancestral gap* in an ancestral scaffold is the sequence located between two consecutive ancestral markers (say X and Y). For each ancestral gap, we consider the extant genomes in which occurrences of X and Y are consecutive (no extant marker is between them) and in the same respective orientations as in the ancestor, thus defining an *extant gap $X - Y$*. We define a *conserved extant gap* as an extant gap whose length is equal in two extant genomes whose evolutionary path in T contains A, following a Dollo criterion. The lengths of conserved extant gaps $X - Y$ define a *length interval* for the ancetsral gap $X - Y$. If there is no conserved extant gap, the ancestral gap length interval is defined by the minimum and maximum extant gap lengths between X and Y. We align all sequences of extant gaps between markers X and Y whose length is in this interval into a multiple sequence alignment. A parsimonious estimation of each ancestral gap sequence is computed from the corresponding alignment of extant gap sequences using the classical Fitch algorithm (Fitch, 1971).

3 RESULTS

We describe here the result of our method FPSAC applied to the dataset described in Bos *et al.* (2011), followed by a preliminary analysis of the resulting scaffolded chromosome.

3.1 Data

The input data are the 2134 larger assembled contigs (500 nt and above) described in Bos *et al.* (2011), and the DNA sequences of the fully assembled chromosomes of four *Y.pseudotuberculosis* genomes and seven *Y.pestis* genomes, of which five are believed to descend from the *Y.pestis* strain that was involved in the Black Death pandemic (Fig. 3).

3.2 Contig segmentation and homologous families

The sequences of the 2134 contigs were mapped to the full genome sequences of the 11 selected extant genomes using
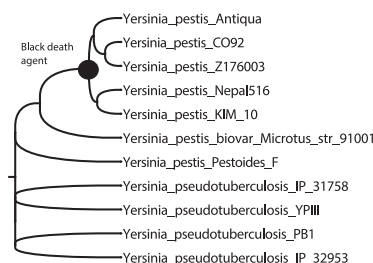
Megablast (Zhang *et al.*, 2000) with default parameters. As already noted by Bos *et al.* (2011), 29 contigs did not map on the *Yersinia* genomes, leaving 2105 ancestral contigs to analyze. The segmentation step resulted in 2616 homologous families. Almost all families have multiplicity 1, but 21 of them have multiplicity greater than 1, and among them, 20 have multiplicity 2 or 3, which indicates that most repeated parts of the genomes were not present in the larger contigs. We removed the last family, which corresponded to the 5S ribosomal protein family, because of its combined short length (133 nt) and high multiplicity (8). The amount of DNA encoded by the ancestral markers, when multiplicity is accounted for, is 3 846 616 nt of ancestral DNA, whereas the initial contigs encode 4 013 159 nt.

3.3 Comparative scaffolding

We detected 2634 putative ancestral adjacencies. Only 6 adjacencies of these 2634 putative ancestral adjacencies needed to be discarded to obtain a maximum weight subset of adjacencies compatible with a set of linear/circular scaffolds. There were 29 conserved repeat spanning intervals, and 2 of them needed to be discarded to extract a maximum weight subset that defined an unambiguous set of three large linear scaffolds, in which all contigs are represented.

There are six possibilities for joining these three scaffolds into one circular scaffold. Extant adjacencies between markers located at the scaffolds extremities were computed and defined an order and orientation for the three scaffolds: two adjacencies between scaffold extremities were supported by all outgroup species, whereas no adjacency between scaffold extremities was supported by ingroup species, and the last adjacency was supported by one outgroup (*Y.pestis Microtus*) and involved a marker absent from all *Y.pseudotuberculosis* genomes.

3.4 Gap lengths and sequences

Out of 2636 ancestral gaps only 22 did not have a length interval supported according to the Dollo criterion. In most other cases, length intervals were narrow: 2561 of the gaps (out of 2636) have a length interval whose bounds differ by at most 10 nt.

Next for each ancestral gap, we aligned all extant gaps whose lengths fell in the ancestral gap length interval, using Muscle (Edgar, 2004) (version 3.8.31), and constructed an ancestral sequence from each alignment using Fitch's algorithm (Fitch, 1971). This resulted into a single sequence containing alternating sequenced ancestral contig segments and estimated ancestral gap sequences, illustrated in Figure 4.
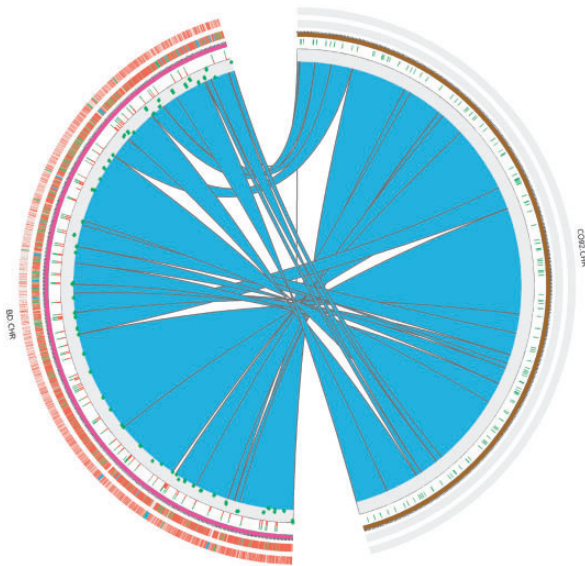
Fig. 4. Comparison of the reconstructed Black Death agent chromosome (left) and of the *Y.pestis CO92* chromosome (right). Outside tracks of the Black Death agent chromosome represents gaps (outer track) and markers (inner track), with red (respectively green, blue) indicating small (resp. mid-length, large) elements. The first two inside tracks represent annotated (green) and inferred (green) insertion sequences. The scattered inside track represents the level of breakpoint reuse in evolutionary scenario between the ancestor and the strains *Y.pestis Antiqua*, *Y.pestis KIM10* and *Y.pestis biovar Microtus str. 91001*. Blue ribbons join colinear chromosome segments. Figure computed using Circos (Krzywinski *et al.*, 2009)

3.5 Assessing accuracy with simulations

To assess the validity and accuracy of FPSAC, we simulated 50 datasets as follows (full details of the simulation and results are given in Supplementary Material). First, for each dataset, one of the current extant genomes was randomly chosen as the ancestral genome and it was allowed to evolve along the *Yersinia* phylogeny by performing up to X random inversions along each branch, with $X \in \{10, 20, 30, 40, 50\}$; note these numbers are all greater than the estimated rearrangement numbers for the real data, thus resulting in 11 simulated extant genomes expected to be more scrambled than the real data. Next, 2134 contigs were selected along the genome following the length distribution of the real contigs, and 10 pairs were used to create chimeric contigs. Finally, the FPSAC pipeline was applied on the resulting 50 dataset (ancient contigs and extant genomes) with the same parameters than on the real *Yersinia* data.

We obtained on average 2808.42 families, 130.64 having a multiplicity >1. The scaffolding resulted into a single scaffold except in five cases (average number of scaffolds of 1.18); there were few adjacencies (two in total over the fifty data sets) and repeat spanning intervals (three in total) that needed to be discarded. To assess the accuracy of the contig order implied by the scaffolds, we looked at occurrences of the non-chimeric contigs in the reconstructed sequence (including the reconstructed gaps) and at the length of the gaps between these occurrences. We

found that 99.47% of the initial contigs appear in the reconstructed sequence with at least 95% of identity over 95% of their length, and that 98.66% of the gaps between consecutive contigs were reconstructed with the exact length in the reconstructed sequence. Regarding chimeric contigs, 99.14% of them were detected as chimeric. These high accuracy numbers are consistent with previous simulations on the reconstruction of ancestral gene orders from randomly rearranged extant genomes (Ma *et al.*, 2006), although here we can also observe a high accuracy in the reconstructed gap lengths, which was not considered in previous simulations.

3.6 Analysis of the reconstructed ancestor

The pipeline described previously resulted in an ancestral genome sequence of length 4.6 Mb showing that roughly 775 kb were added to the ancestral marker sequences by the gap sequences estimation step.

In the resulting scaffold, each occurrence of an ancestral marker corresponds to one or several segments of the initial contigs. The ordering of these segments is mostly compatible with the initial contigs. We found only one *chimeric* contig (see Fig. 5), split into two non-adjacent markers in the ancestral genome organization. Also four contig segments were found to be *duplicated*: a large part (>500 nt) of each is probably present in more than one occurrence in the ancestral genome, whereas the initial assembly predicted only one occurrence. Finally, 63 contigs have a sequence that is found, up to small variations, inside another contig, whereas their number of extant occurrences suggest they have multiplicity 1, so we believe they are *redundant*. An alternative explanation is that they are derived mutations of the ancient genome, which, in such a case, would not be ancestral to current strains.

Regarding the six discarded adjacencies, two of them point toward a possible large-scale inversion. Both, the selected adjacencies and intervals, as well as the discarded ones, have similar phylogenetic support. So this alternative structure cannot be ruled out as non-ancestral, which raises the question of the possible coexistence of different genome architectures among the *Y.pestis* infecting the host individual whose remains were used for sequencing.

We also took advantage of the availability of a full chromosome sequence for the main chromosome of the Black Death agent to analyze its structure and evolution at the whole-genome scale. We first analyzed insertion sequence (IS) elements that have been suspected to be involved into the high rearrangement rate of *Y.pestis* genomes (Chain *et al.*, 2004). We mapped extant IS to the reconstructed ancestral chromosome (see Supplementary Material). This resulted in 92 ancestral gaps and markers containing IS. We confirmed this comparative annotation with an automatic annotation of the reconstructed chromosome sequence. We could also observe that a large proportion of these IS (at least 58) were already present in the last common ancestor of all *Y.pestis* strains, whereas they are almost completely absent from the genomes of the considered *Y.pseudotuberculosis*, thus providing more evidence that the *Y.pestis* speciation from its *Y.pseudotuberculosis* ancestor was characterized by a burst of IS insertion (Chain *et al.*, 2004).
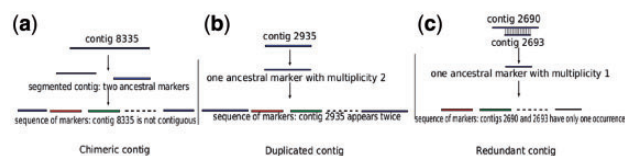
Fig. 5. Contig correction: (**a**) the contig is cut during the segmentation procedure, but not joined during the marker ordering; (**b**) the contig is found to have two occurrences in the marker ordering; (**c**) two contigs contain the same DNA sequence and this sequence is predicted to have only one occurrence in the marker ordering

We also analyzed the genome rearrangements between the reconstructed ancestral sequence and the extant genome sequences by sampling inversion scenarios between the ancestral genome and the extant genomes using the software DCJ2HP (Miklós and Tannier, 2010). There are 8–9 inversions between the *Y.pseudotuberculosis* strains and the medieval genome, and 9–22 inversions when compared with the (evolutionarily closer) *Y.pestis* strains, showing a clear acceleration of evolutionary rearrangement following the Black Death *Y.pestis* divergence (see Supplementary Material). As noticed by Darling *et al.* (2008), we can also observe that inversion breakpoints are not randomly distributed and used (Fig. 4): highly used ones are concentrated in one-third of the chromosome, around its probable replication origin. The positions of the inversion breakpoints are also highly correlated with IS, as remarked earlier (Deng *et al.*, 2002): 76 of the 118 mapped breakpoints are close (<1000 nt distant) to some predicted IS, whereas this number drops to 39 for uniformly sampled random coordinates (*P*-value $< 10^{-3}$). Rearrangements are numerous in all *Y.pestis* branches, strongly suggesting that they could be driven by the IS.

4 DISCUSSION

4.1 Contig segmentation and marker multiplicities

Aligning contigs to extant genomes and using these alignments to segment contigs might at first seem counterintuitive, as it increases the fragmentation of the initial assembly. However, it allows us to take advantage of the available fully assembled extant genomes to identify potential chimeric contigs and to extract potential repeated sequences from the contigs, which would have been collapsed into a single contig, a well-documented issue with assembling from short reads (Treangen and Salzberg, 2012). Our approach follows a recent suggestion by Roy *et al.* (2012) to rely on shorter contigs of higher quality (here in terms of mapping to related genomes). This phase benefited from the high sequence conservation in the *Y.pestis* clade that allowed us to rely on high similarity pairwise alignments as input of the segmentation phase. Less conserved data would likely require more involved methods to compute a segmentation into non-overlapping homologous families (Angiuoli and Salzberg, 2011;Minkin *et al.*, 2013).

Finally, the possibility to infer the multiplicity of contig segments from the alignment on extant genomes, using comparative genomics methods designed to study the evolution of gene families, offers an elegant alternative, specific to aDNA assembly, however, to current copy number estimation methods that rely on the depth of coverage, which can be uneven when sequencing highly fragmented aDNA.

4.2 Estimating ancestral gap sequences

The key idea is that conserved adjacencies are also likely to define conserved gaps. In the data processed, we can observe that for most ancestral gaps, a strict Dollo parsimony criterion identifies conserved gaps. Moreover, again benefiting from the high sequence conservation of the *Y.pestis* genomes, we could estimate most of the ancestral gap sequences from the multiple alignments of the corresponding extant gaps using a standard ancestral character reconstruction method. If greater sequence variation was observed, more powerful methods designed to infer ancestral DNA from a multiple alignment would be appropriate (Blanchette *et al.*, 2004; Diallo *et al.*, 2010; Liberles, 2007). In a future work, we aim to use the reconstructed gap sequences as a template to exactly assemble these gaps from the sequenced reads. However, optimally mapping aDNA reads onto extant DNA requires specific protocols that have recently been developed for eukaryotic aDNA, but still needs to be established for bacterial aDNA (Schubert *et al.*, 2012).

4.3 Scaffolding and comparative genomics

The FPSAC method follows principles similar to most existing scaffolding methods designed for extant genomes. It relies on extracting a genome structure from a graph (the adjacency graph), whose vertices are sequence elements and edges indicate connectivity between pairs of vertices. In most scaffolding algorithms, edges of this graph are defined by mate-pair reads, whereas we rely on adjacencies and intervals that are conserved under a Dollo parsimony criterion. The main difference we can observe is the low number of tangles in the graph we obtained compared with the usual large number observed in graphs based on mate-pairs, in part because of the absence of repeated sequences in the analyzed contigs. It is interesting to observe that, despite the fact that *Y.pestis* genomes are highly rearranged, FPSAC was able to capture a clear signal regarding the organization of markers along the ancestral chromosome. Also important is the use of recently developed polynomial time exact algorithms to extract a consistent set of adjacencies while accounting for the multiplicity of repeated segments (Manuch *et al.*, 2012) and to assess the compatibility of repeat spanning intervals with a given adjacency graph (Chauve *et al.*, 2013).

4.4 Applicability to other datasets

We applied FPSAC to a dataset with specific characteristics. The assembled ancestral contigs were obtained using a library-enrichment approach and are assumed to belong to the genome of an internal node of a phylogeny, whose leaves are sequenced and assembled. Moreover, the clade of interest contains high sequence conservation and highly rearranged genomes. We address the impact of these different points below and discuss the applicability of FPSAC to a wider range of datasets.

In the case where descendants of the ancient genome of interest are not available, either due to lineage extinction or because they have not been sequenced, other comparative methods can be used, that do not rely on a Dollo parsimony principle

(Husemann and Stoye, 2010). So the only important requirement to use FPSAC is the availability of the genome sequences of at least two related genomes whose evolutionary path contains the ancestor of interest. From there, the performances of the scaffolding obtained with FPSAC will depend both on the level of sequence conservation of both sequence identity and synteny in the considered related genomes. Diverged extant genomes might result in difficulty to obtain homologous families, which adequately span the initial set of ancient contigs, as well as computing their multiplicities. High rearrangement rates might result in wrong adjacencies because of convergent evolution. Useful indicators to assess the results obtained with FPSAC are thus both the coverage of contigs and extant genomes by ancestral markers and in the number of discarded adjacencies during the scaffolding phase.

The method we describe can be applied as is even if some of the chosen closely related extant genomes are not fully assembled. The impact of unassembled extant genomes is likely to be more fragmented scaffolding than what we observed on the Black Death agent dataset because of undetected ancestral adjacencies or repeat spanning intervals.

Finally, if the initial ancient contigs originate from a mixture of microbial backgrounds, for example if they result from *de novo* assembly of reads obtained through shotgun sequencing, then FPSAC can be used to assemble contigs subsets that have been identified, through comparison with extant genomes sequences, as belonging to well identified clades and satisfy the requirements described previously. However, even this initial step is currently a challenge, for example because of repeated sequences belonging, up to variations, to several genomes (Pell *et al.*, 2012). The problem of applying a method such as FPSAC to a whole set of contigs originating from a mixture of genomes, that is, for scaffolding an ancient metagenome, is an important research avenue.

5 CONCLUSION

Technological sequencing advances can now provide sequences from whole ancient bacterial genomes, which promises to be an invaluable source of knowledge for understanding pathogen evolution. However, assembling ancient bacterial genomes poses specific issues, in particular because of high fragmentation in numerous contigs. In parallel, computational paleogenomics by comparative methods has grown tremendously, and computational methods can now provide ancestral genome sequences accounting for substitutions or small indels, ancestral gene content or ancestral genome organizations, at the level of full chromosomes, but, until now, were never combined to scaffold and estimate the sequence of an ancient bacterial chromosome. In the present work, we described a general method to combine both sequencing and computational reconstruction, and illustrated its potential on a real dataset.

Conflict of Interest: none declared.

REFERENCES

Alekseyev,M.A. and Pevzner,P.A. (2009) Breakpoint graphs and ancestral genome reconstructions. *Genome Res.*, **19**, 943–957.
Angiuoli,S.V. and Salzberg,S.L. (2011) Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics*, **27**, 334–342.
Bashir,A. *et al.* (2012) A hybrid approach for the automated finishing of bacterial genomes. *Nat. Biotechnol.*, **30**, 701–707.
Bérard,S. *et al.* (2012) Evolution of gene neighborhoods within reconciled phylogenies. *Bioinformatics*, **28**, i382–i388.
Bertrand,D. *et al.* (2010) Reconstruction of ancestral genome subject to whole genome duplication, speciation, rearrangement and loss. In: Singh,M. and Moulton,V. (eds) *Algorithms in Bioinformatics, 10th International Workshop, WABI 2010, Liverpool, UK, September 6-8, 2010. Proceedings, volume 6293 of Lecture Notes in Bioinformatics.* pages 78–89. Springer Verlag.
Blanchette,M. *et al.* (2004) Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res.*, **14**, 2412–2423.
Bos,K.I. *et al.* (2011) A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature*, **478**, 506–510.
Chain,P. *et al.* (2004) Insights into the evolution of *Yersinia pestis* through whole-genome comparison with yersinia pseudotuberculosis. *Proc. Natl Acad. Sci. USA*, **101**, 13826–13831.
Chapman,J. *et al.* (2011) Meraculous: De novo genome assembly with short paired-end reads. *PLoS One*, **6**, e23501.
Chauve,C. and Tannier,E. (2008) A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. *PLoS Comput. Biol.*, **4**, e1000234.
Chauve,C. *et al.* (2010) Yeast ancestral genome reconstructions: the possibilities of computational methods II. *J. Comput. Biol.*, **17**, 1097–1112.
Chauve,C. *et al.* (2013) Hypergraph covering problems motivated by genome assembly questions (short abstract). To appear in the proceedings of. *International Workshop On Combinatorial Algorithms, 2013*.
Cohen,O. *et al.* (2010) Gloome: gain loss mapping engine. *Bioinformatics*, **26**, 2914–2915.
Csürös,M. (2010) Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, **26**, 1910–1912.
Darling,A.E. *et al.* (2008) Dynamics of genome rearrangement in bacterial populations. *PLoS Genet.*, **4**, e1000128.
Deng,W. *et al.* (2002) Genome sequence of *Yersinia pestis* KIM. *J. Bacteriol.*, **184**, 4601–4611.
Diallo,A.B. *et al.* (2010) Ancestors 1.0: a web server for ancestral sequence reconstruction. *Bioinformatics*, **26**, 130–131.
Djelouadji,Z. *et al.* (2011) Palaegenomics of *Mycobacterium tuberculosis*: epidemic burst with a degrading genome. *Lancet Infect. Dis.*, **11**, 641–650.
Donmez,N. and Brudno,M. (2013) Scarpa: scaffolding reads with practical algorithms. *Bioinformatics*, **29**, 428–434.
Donoghue,H. (2011) Insights gained from paleomicrobiology into ancient and modern tuberculosis. *Clin. Microbiol. Infect.*, **17**, 821–829.
Donoghue,H. and Spigelman,M. (2006) Pathogenic microbial ancient DNA: a problem or an opportunity? *Proc. R. Soc. B*, **273**, 641–642.
Drancourt,M. (2012) Plague in the genomic area. *Clin. Microbiol. Infect.*, **18**, 224–230.
Drancourt,M. and Raoult,D. (2005) Palaemicrobiology: current issues and perspectives. *Nat. Rev. Microbiol.*, **3**, 23–35.
Edgar,R.C. (2004) Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
Fitch,W.M. (1971) Toward defining the course of evolution: minimum change for a specified tree topology. *Syst. Zool.*, **20**, 406–416.
Fremez,R. *et al.* (2007) Phylogenetic exploration of bacterial genomic rearrangements. *Bioinformatics*, **23**, 1172–1174.
Gao,S. *et al.* (2011) Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. *J. Comput. Biol.*, **18**, 1681–1691.
Gnerre,S. *et al.* (2009) Assisted assembly: how to improve a de novo genome assembly by using related species. *Genome Biol.*, **10**, R88.
Gordon,J.L. *et al.* (2009) Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *saccharomyces cerevisiae* genome. *PLoS Genet.*, **5**, e1000485.
Husemann,P. and Stoye,J. (2010) Phylogenetic comparative assembly. *Algorithms Mol. Biol.*, **5**, 3.
Jones,B.R. *et al.* (2012) ANGES: Reconstructing ancestral genomes maps. *Bioinformatics*, **28**, 2388–2390.

Kim,J. *et al.* (2013) Reference-assisted chromosome assembly. *Proc. Natl Acad. Sci. USA*, **110**, 1785–1790.

Krzywinski,M. *et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.

Liberles,D.A. (ed.) (2007) *Ancestral Sequence Reconstruction.* Oxford University Press.

Lin,H. *et al.* (2012) AGORA: Assembly guided by optical restriction alignment. *BMC Bioinformatics*, **13**, 189.

Ma,J. *et al.* (2006) Reconstructing contiguous regions of an ancestral genome. *Genome Res.*, **16**, 1557–1565.

Ma,J. *et al.* (2008) Dupcar: reconstructing contiguous ancestral regions with duplications. *J. Comput. Biol.*, **15**, 1007–1027.

Manuch,J. *et al.* (2012) Linearization of ancestral multichromosomal genomes. *BMC Bioinformatics*, **13**(**Suppl. 19**), S11.

Miklós,I. and Tannier,E. (2010) Bayesian sampling of genomic rearrangement scenarios via double cut and join. *Bioinformatics*, **26**, 3012–3019.

Minkin,I. *et al.* (2013) Sibelia. http://bioinf.spbau.ru/sibelia.

Muffato,M. *et al.* (2010) Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics*, **26**, 1119–1121.

Munoz,A. *et al.* (2010) Scaffold filling, contig fusion and comparative gene order inference. *BMC Bioinformatics*, **11**, 304.

Murat,F. *et al.* (2010) Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res.*, **20**, 1545–1557.

Ouangraoua,A. *et al.* (2011) Reconstructing the architecture of the ancestral amniote genome. *Bioinformatics*, **27**, 2664–2671.

Parkhill,J. and Wren,B.W. (2011) Bacterial epidemiology and biology - lessons from genome sequencing. *Genome Biol.*, **12**, 230.

Pell,J. *et al.* (2012) Scaling metagenome sequence assembly with probabilistic de bruijn graphs. *Proc. Natl Acad. Sci. USA*, **109**, 13272–13277.

Putnam,N.H. *et al.* (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science*, **317**, 86–94.

Ribeiro,F.J. *et al.* (2012) Finished bacterial genomes from shotgun sequence data. *Genome Res.*, **22**, 2270–2277.

Rissman,A.I. *et al.* (2009) Reordering contigs of draft genomes using the mauve aligner. *Bioinformatics*, **25**, 2071–2073.

Roy,R.S. *et al.* (2012) Sliq: Simple linear inequalities for efficient contig scaffolding. *J. Comput. Biol.*, **19**, 1162–1175.

Salmela,L. *et al.* (2011) Fast scaffolding with small independent mixed integer programs. *Bioinformatics*, **27**, 3259–3265.

Sankoff,D. *et al.* (2009) Towards improved reconstruction of ancestral gene order in angiosperm phylogeny. *J. Comput. Biol.*, **16**, 1353–1367.

Schubert,M. *et al.* (2012) Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics*, **13**, 178.

Schuenemann,V.J. *et al.* (2011) Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of *Yersinia pestis* from victims of the black death. *Proc. Natl Acad. Sci. USA*, **108**, E746–E752.

Szöllősi,G.J. *et al.* (2012) Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc. Natl Acad. Sci. USA*, **109**, 17513–17518.

Treangen,T.J. and Salzberg,S.L. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.*, **13**, 36–46.

Wang,Y. *et al.* (2006) Reconstruction of ancient genome and gene order from complete microbial genome sequences. *J. Theoret. Biol.*, **239**, 494–498.

Wilson,D. (2012) Insights from genomics into bacterial pathogen populations. *PLoS Pathog.*, **8**, e1002874.

Wittler,R. *et al.* (2011) Consistency of sequence-based gene clusters. *J. Comput. Biol.*, **18**, 1023–1039.

Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Res.*, **18**, 821–829.

Zhang,Z. *et al.* (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.