Improved assemblies and comparison of two ancient *Yersinia pestis* genomes

Nina Luhmann[1,2⋆], Daniel Doerr[2,3] and Cedric Chauve[4]

[1] International Research Training Group "Computational Methods for the Analysis of the Diversity and Dynamics of Genomes", Bielefeld University, Germany
[2] Genome Informatics, Faculty of Technology and Center for Biotechnology, Bielefeld University, Germany
[3] School of Computer and Communication Sciences, EPFL, 1015 Lausanne, Switzerland
[4] Department of Mathematics, Simon Fraser University, Burnaby (BC), Canada

Abstract. *Yersinia pestis* is the causative agent of the bubonic plague, a disease responsible for several dramatic historical pandemics. Progress in ancient DNA (aDNA) sequencing rendered possible the sequencing of whole genomes of important human pathogens, including the ancient *Yersinia pestis* strains responsible for important outbreaks of the bubonic plague in London in the 14th century and in Marseille in the 18th century among others. However, aDNA sequencing data are still characterized by short reads and non-uniform coverage, so assembling ancient pathogen genomes remains challenging and prevents in many cases a detailed study of genome rearrangements. It has recently been shown that comparative scaffolding approaches can improve the assembly of ancient *Yersinia pestis* genomes at a chromosome level. In the present work, we address the last step of genome assembly, the gap-filling stage. We describe an optimization-based method AGapEs (Ancestral Gap Estimation) to fill in inter-contig gaps using a combination of a template obtained from related extant genomes and aDNA reads. We show how this approach can be used to refine comparative scaffolding by selecting contig adjacencies supported by a mix of unassembled aDNA reads and evolutionary parsimony signal. We apply our method to two ancient *Yersinia pestis* genomes from the London and Marseilles outbreaks of the bubonic plague. We obtain highly improved genome assemblies for both the London strain and Marseille strain genomes, comprised of respectively five and six scaffolds, with 95% of the assemblies supported by ancient reads. We analyze the genome evolution between both ancient genomes in terms of genome rearrangements, and observe a high level of synteny conservation between these two strains.

1 Introduction

Yersinia pestis is the pathogen responsible for the bubonic plague, a disease that marked human history through several dramatic pandemics, including the Justinian Plague and the Black Death. It diverged a few thousands years ago from a relatively non-virulent human pathogen, *Yersinia pseudotuberculosis*. The precise timing of the divergence between these two pathogens is still controversial [Rasmussen et al., 2015], but it is widely accepted that the emergence of *Yersinia pestis* as a virulent human pathogen was characterized by the acquisition of numerous repeat sequences, especially Insertion Sequences (IS) that triggered an extensive chromosomal rearrangement activity [Chain et al., 2004, Darling et al., 2008]. This has led to consider the *Yersinia* family as an important model for the study of genomes rearrangements during pathogen evolution [McNally et al., 2016].

Traditionally, the study of genome rearrangements relies on a comparative approach using the genomes of related extant organisms. Under appropriate models of evolution, this

⋆ Corresponding author

comparison provides indirect insight into genomic features of ancient species and their evolution toward extant species, see [Darling et al., 2008] for example for the specific case of genome rearrangements in *Yersinia*. In contrast, sequenced ancient DNA (aDNA) extracted from conserved remains can give direct access to the sequence of ancient genomes and thus, theoretically, allows to study the evolution from ancestors to descendants directly. Following advances in aDNA high-throughput sequencing technologies and protocols [Gasc et al., 2016, Orlando et al., 2015, Yoshida et al., 2015, Hagelberg et al., 2015, Hofreiter et al., 2015, Pääbo et al., 2004], the genomes of several ancient human, animal and plant pathogens have recently been sequenced at the level of complete or almost complete chromosomes, including the agents of potato blight [Martin et al., 2013, Yoshida et al., 2013], brucellosis [Kay et al., 2014], tuberculosis [Bos et al., 2014], leprosis [Schuenemann et al., 2013], *Helicobacter pylori* [Maixner et al., 2016], cholera [Devault et al., 2014] and of the bubonic plague [Wagner et al., 2014, Bos et al., 2011, Bos et al., 2016], leading to important historical and evolutionary discoveries. However, unlike extant DNA high-throughput sequencing that is experiencing a breakthrough transition towards long-reads, aDNA sequencing methods generate extremely short reads with low and non-uniform coverage [Yoshida et al., 2015]. As a result, aside of rare exceptions [Schuenemann et al., 2013], the assembly of aDNA reads generates numerous short contigs. For example, the reference-based assembly of the Black Death pandemic agent resulted in several thousand contigs [Bos et al., 2011], two thousand of them of length 500bp and above. While short aDNA reads can be mapped onto one or several extant reference genomes to detect important evolutionary signals such as SNPs and small indels [Schubert et al., 2014, Peltzer et al., 2016], highly fragmented assemblies make it challenging to exploit aDNA sequencing data to analyze the evolution of pathogen genome organization, including important features such as the evolution of repeats and large scale genome rearrangements.

Without long-read sequencing data, comparative scaffolding based on the comparison of the contigs of a genome of interest with related assembled genomes has proven to be a useful approach to improve the assembly of fragmented genomes, especially bacterial genomes [Rissman et al., 2009, Rajaraman et al., 2013, Kolmogorov et al., 2014]. In particular, the FPSAC method [Rajaraman et al., 2013] was introduced to improve ancient genome assemblies within a phylogenetic context. It was applied to aDNA contigs from the *Yersinia pestis* strain responsible for the medieval London bubonic plague outbreak – that was shown to be ancestral to several extant *Yersinia pestis* strains [Bos et al., 2011] – and resulted in an improvement of the initial contig assembly from thousands of contigs to a chromosome-scale scaffolding. Moreover, taking advantage of the high sequence conservation in *Yersinia pestis* genomes, the inter-contigs gaps of the ancient *Yersinia pestis* strain were filled with putative sequences reconstructed from multiple sequence alignments of conserved extant gaps. This gap-filling step shed an interesting light on genomic features hidden within the assembly gaps, in particular IS and their correlation with rearrangement breakpoints reuse. However, the scaffolding of adjacencies and gap sequences obtained in [Rajaraman et al., 2013], that accounted for roughly 20% of the genome size, were inferred through computational methods within a parsimony framework. This can be sensitive to convergent evolution that cannot be ruled out for genomes with a high rate of genome rearrangements such as *Yersinia pestis* [Darling et al., 2008]. In the present work, we address this issue by using the large set of aDNA reads that are unassembled after the contig assembly stage, to both confirm scaffolding adjacencies and fill-in inter-contigs gaps.

We introduce the method AGapEs (Ancestral Gap Estimation) which, for a potential adjacency between two ancient contigs, attempts to fill the inter-contig gap by selecting a set of overlapping aDNA reads that minimizes the edit distance to a template sequence obtained from the extant genomes sequences that support the adjacency. We directly include annotations of potential Insertion Sequences in the extant genomes in the analysis. In particular, when the presence of an IS in the ancient genome is doubtful, due to a mixed signal of presence/absence in the supporting extant genomes, a pair of templates can be considered, respectively including and excluding the IS. We apply this strategy to two data sets of ancient DNA reads for ancestors of the human pathogen *Yersinia pestis* [Chain et al., 2004, McNally et al., 2016]. This bacterium is the causative agent of the bubonic plague and responsible for three major epidemics, the last one still on-going. The first aDNA data was obtained from London victims of the Black Death pandemic in the 14th century [Bos et al., 2011], and the second consists of five samples from victims of Great Plague of Marseille around 400 years later [Bos et al., 2016]. For both data sets, we obtain an assembly with reduced fragmentation and are able to fill a large number of inter-contig gaps with aDNA reads. We identify several genome rearrangements between the ancient strains and extant *Yersinia pestis* genomes, however observe only a single small inversion between both ancient strains, suggesting that the genome organization of the agent of the second major plague pandemic was highly conserved.

2 Materials and Methods

We first describe the input to our analysis, namely ancient sequencing data, ancient and extant assemblies and annotations of IS, before outlining the general pipeline we used to improve the assembly of the ancient genomes.

Sequencing data and reference genomes. The first aDNA data set was obtained from a London victim of the Black Death pandemic in the 14th century [Bos et al., 2011] (individual 8291), the second consists of five samples from victims of Great Plague of Marseille around 400 years later [Bos et al., 2016]. The average read length is 53 bp in the London dataset and 75bp in the five Marseille samples (Figure S3). We rely on seven extant *Yersinia pestis* and four *Yersinia pseudotuberculosis* as reference genomes (see Table S1). The phylogeny of the considered strains is depicted in Figure S1 and is taken from [Bos et al., 2016, Bos et al., 2011].

Contig assembly and preprocessing. Bos et al [Bos et al., 2011] describe a reference-based assembly of the London strain consisting of 2,134 contigs of length at least 500bp. It was obtained with the assembler Velvet [Zerbino and Birney, 2008] using the extant strain *Yersinia pestis CO92* as a reference. In order to assess the influence of the reference sequence in the assembly of the ancient genome, we de novo assembled aDNA reads into contigs using

Minia [Chikhi and Rizk, 2013] for both aDNA data sets (London outbreak and Marseille outbreak). Minia is a conservative assembler based on an efficient implementation of the de Bruijn graph methodology. In general, Minia produces shorter contigs, as it avoids assembly decisions in case of ambiguity in the sequence data. We will refer to the assembly by Bos et al. as *reference-based* and the Minia assemblies as *de novo* assemblies in the following. To allow the comparison with extant genomes, contigs above a minimum length threshold were aligned with the extant genomes to define families of homologous synteny blocks (called markers from now) as described in [Rajaraman et al., 2013]. Marker families were then filtered to retain only one-to-one orthologous families, *i.e.* families that contain one and exactly one marker in each considered extant and ancient genome.

Insertion sequence annotation. Insertion Sequences (IS) are strongly related to rearrangements in *Yersinia pestis* evolution, and their annotation in the considered extant genomes is crucial. In order to annotate IS, we designed our own annotation pipeline. Because IS elements in the original Genbank files were rather disparately annotated, we relied on automated annotations from the Basys annotation server [Van Domselaar et al., 2005]. Basys identified 11 families of IS transposase proteins (see Table S2). For each of these families, we produced a multiple alignment of their annotated sequences using `muscle` [Edgar, 2004] which was subsequently used to train Hidden Markov Model (HMM) profiles. Using `hmmer` [Eddy et al., 2009], we then annotated those regions as associated to IS elements that showed significant correlation to any of the HMM profiles. We eventually combined the Genbank annotations with these derived annotations. The number of these IS annotations per reference genome ranges from 151 in *Yersinia pestis KIM10+* to 293 in *Yersinia pestis Antiqua* (see Table S1). The length of the annotations ranges from 60bp to 2,417bp; some short annotations deviate from the expected length for IS, however, in order to avoid filtering any true annotations, we include them all as potential IS coordinates in the following analysis.

Ancestral marker adjacencies. Each marker can be defined by a pair of marker extremities. An adjacency consists of two markers extremities that are contiguous along a genome, *i.e.* are not separated by a sequence containing another marker. For extant genomes, extant adjacencies can be observed directly, while for an ancestral genome of interest, we infer potential ancestral adjacencies using the Dollo parsimony principle as in [Rajaraman et al., 2013]: two ancient marker extremities are potentially adjacent if there exist two extant genomes whose evolutionary path contains the most common recent ancestor of the London and Marseille strains and where the two corresponding extant marker extremities are contiguous (see Figure S5 for an example). Hence every potential ancestral adjacency is supported by a set of extant adjacencies. A *gap* is the sequence between the two marker extremities defining an adjacency. Therefore each ancestral gap is likewise supported by a set of extant gap sequences.

We say that two potential ancestral adjacencies are *conflicting* if they share a common marker extremity. An *IS-annotated* adjacency is supported by at least one extant adjacency whose gap contains an IS annotation. An adjacency that is neither conflicting nor IS-annotated is said to be *simple*.

Assembly of ancestral gap sequences from aDNA reads. The main methodological contribution we introduce is a template-based method to assess the validity of a potential ancestral adjacency. The general principle is to associate to every ancestral gap a template sequence obtained from the supporting extant gaps sequences. We can then map aDNA reads onto this template and assemble the mapped aDNA reads into a sequence that minimizes the edit distance to the template sequence. The rationale for this template-based approach is that, due to the low coverage of the aDNA reads and their short length, existing gap-filling methods fail to fill a large number of ancestral gaps. For example, the method gap2Seq [Salmela et al., 2015], a recent efficient gap-closing algorithm based on finding a path of given length in a de Bruijn graph, is not able to fill roughly half of the ancestral gaps of the Black Death data set (see Table S6).

We describe now the AGapEs algorithm. Assume we are given a template sequence t for a gap in an adjacency $a = \{m_1, m_2\}$ between two marker extremities. We define $R = m_1{+}t{+}m_2$ as the concatenated nucleotide sequence of the oriented markers and the respective template. We first align the aDNA reads onto R, using BWA [Li and Durbin, 2009], where we only consider mappings whose start and/or end position is in t (i. e. either fully included in t or overlapping the junction between a marker and the gap template). Next, we construct a graph $G(V, E)$ where vertices are mappings $m \in \mathcal{M}_t$ and there is an edge between two vertices (i. e. mappings) if the two mapping coordinates (segments of R) overlap. For each such edge/overlap, we define s as the non-overlapping suffix of the mappings with the highest end coordinate. We can then associate a weight to each edge given by the edit distance between s and the subsequence R_s of R it aligns to. A sequence of overlapping reads that minimizes the distance to t can then be found by searching for a shortest path between the vertex labeled with the smallest start position (i. e. the first mapping covering the junction between m_1 and t) and the vertex labeled with the largest start position (i. e. the last mapping covering the junction between t and m_2). See Figure S8 for an illustration.

If such a path exists, it can be found with Dijkstra's algorithm [Dijkstra, 1959] implemented based on a min-priority queue in $O(|E|+|V|\log|V|)$ time. If no such path exists, then there are either regions in R that are not covered by any mapped aDNA read or breakpoints in the mapping, where two consecutive bases in the sequence are covered, but not both by the same read. In these cases, uncovered regions and breakpoints need to be identified in the mapping beforehand to identify start and end vertex of the shortest path. We can then obtain a partial gap filling, precisely for the regions covered by mapped reads.

The key element of the approach described above lies in defining the template sequence or set of alternative template sequences associated to each ancestral gap. We follow the general approach described in [Rajaraman et al., 2013], that computes a multiple sequence alignment of the supporting extant sequence gaps and applies the Fitch-Hartigan parsimony algorithm [Fitch, 1971] to each alignment column to reconstruct a most parsimonious ancestral sequence. If the multiple sequence alignment of extant gaps shows little variation, as is the case for most gaps in our data sets, then a single template sequence can be considered, as we expect that minor variations compared to the true ancestral sequence (substitutions, small indels) will be corrected during the local assembly process. Alternatively, if larger variations are observed, such as larger indels or a contradicting pattern of presence/absence of an IS in the supporting extant gaps, then alternative templates can be considered, under the hypothesis that the true variant can be recovered from the mapped aDNA reads.

Hence in the following analysis, we separate all potential ancestral gaps into groups of simple, conflicting and IS-annotated gaps. For simple and conflicting gaps without IS annotation, we can follow the process described above directly. For IS-annotated gaps, we reduce the described large variations in the multiple alignment by further dividing its supporting extant gaps into sets of annotated and non-annotated sequences respectively. Building the multiple alignment on each of these sets separately allows us to define two alternative templates that can be used as a basis to fill the gap. Ideally, differences in read coverage or breakpoints naturally identified by AGapEs then point to one of the alternative templates for each IS-annotated gap. Further, for each template that is only partly covered by mapped reads, we will correct the covered parts according to the read sequence and revert to the template sequence otherwise.

The implementation of AGapEs is available at `http://github.com/nluhmann/AGapEs`, the data underlying the following results can be downloaded from `http://paleogenomics.irmacs.sfu.ca/DOWNLOADS/AGAPES_data_results.zip`.

3 Results and discussion

3.1 The London strain

Contig assembly and segmentation into markers. In order to assess the impact of the initial contig assembly on the final result, we considered two contig assemblies of the aDNA reads. The reference-based assembly consists of 2,134 contigs of length 500bp and above that cover 4,013,159 bp. As expected, the de novo assembly is more fragmented with 4,183 contigs of length at least 300bp that cover 2,631,422 bp (see Supplementary Material subsection A.4).We compared both contig assemblies by aligning them with MUMmer [Kurtz et al., 2004]. Unaligned bases mostly belong to regions in the reference-based assembly that have not been assembled in the conservative de novo assembly, and only an extremely low amount of nucleotide variations can be observed (Table S3), together with no observed genome rearrangement.

Subsequently, we obtain 2,207 markers that cover 3,463,281 bp in total for the reference-based assembly. For the de novo assembly, we obtain 3,691 markers covering 2,215,596 bp in total. All markers for the de novo assembly are contained in or overlapping with markers from the reference-based assembly.

Reconstructing potential ancestral adjacencies. For the reference-based assembly, we inferred 2,208 potential adjacencies: 1,991 are simple, 207 IS-annotated but non-conflicting, and 10 are conflicting. Among the conflicting adjacencies 8 are also IS-annotated, illustrating that most rearrangements in *Yersinia pestis* that can create ambiguous signal for comparative scaffolding, are associated with IS elements. For the de novo assembly, we obtain 3,691 potential ancestral adjacencies: 3,483 are simple, 201 are IS-annotated and non-conflicting, and only 7 are conflicting, including 5 IS-annotated adjacencies (see also Table S4).

For most potential ancestral adjacencies, the lengths of the sequences in extant genomes associated with the supporting extant adjacencies are very similar, indicating well conserved extant gaps (Figures S7(a) and S7(b)). We have 28 and 21 gaps in the reference-based and de novo assembly respectively whose lengths difference falls into the length range of potential annotated IS elements, thus raising the question of the presence of an IS within these adjacencies in the ancestral genome. We note a small number of potential ancestral adjacencies with strikingly large extant gap length differences (7 and 5 in the respective assemblies). All of these gaps accumulate more than one IS annotation in some extant genomes. Most problematic are two gaps with length differences of more than 100.000 bp. As these gaps are not well conserved in general (apart from the inserted sequences), it is difficult to obtain a good template sequence based on a very fragmented multiple alignment at this point. We will get back to these special gaps in the next paragraphs.

Ancestral gaps filling. We apply AGapEs to all potential ancestral gaps. We assume a gap to be filled, if we find a sequence of reads that covers the whole ancestral gap. As we test two alternative templates for an IS-annotated gap, we consider it filled if only one alternative is covered or if both templates are covered but the IS is only annotated in a single extant genome. In the latter case, we expect the non-IS gap version to be ancestral, as the IS was most likely obtained along the edge to the annotated extant genome. If otherwise both alternative template sequences are covered, we cannot recover the true positive gap at this point and mark it as not filled. If a gap template sequence is only partially covered by mapped aDNA reads, we correct the covered regions as described above and use the template sequence of the uncovered regions to complete filling the gap. Figure 1 summarizes the gap-filling results (see also Table S4).

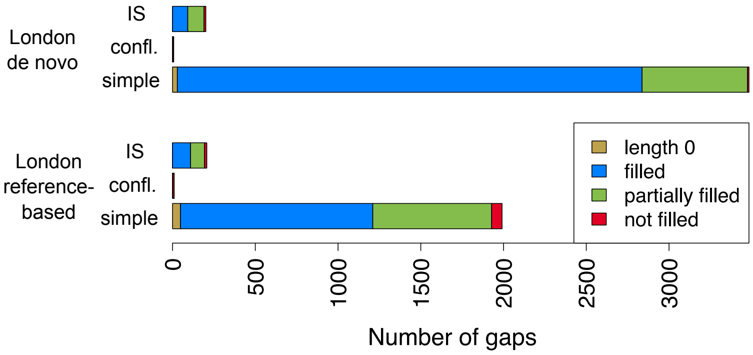

Fig. 1. Results of gap filling for both assemblies. Note that if a gap is conflicting and IS-annotated, we assign it to the conflicting group. We differentiate between gaps of length 0 (i. e. both markers are directly adjacent), completely and partially filled gaps and not filled gaps.

For both assemblies, a high number of gaps is supported by sufficient read coverage that enables us to fill the gap with a sequence of overlapping aDNA reads. Especially considering partially covered gaps for the de novo assembly improves the length of the genome that is supported by reads. Note that we also find covering reads for all gaps of length 0, spanning the breakpoint between directly adjacent markers.

We further computed the edit distance between the reconstructed gap sequence and the previous gap template. For IS-annotated gaps, we computed the distance to a template sequence based on all extant gap occurrences, i. e. without considering the alternative templates as described previously. We identified one case where the parsimonious gap sequence based on all extant occurrences of the adjacency excludes the IS. However if aDNA reads are mapped separately to alternative templates based on IS and non-IS annotated extant gaps, only the IS-annotated gap template is covered.

For IS-annotated gaps, in both assemblies 95 ancestral gaps contain the IS, while 112 resp. 106 ancestral gaps are reconstructed without the IS. From these 95 IS gaps, 22 contain annotations that are shorter than 400bp, however they all contain additional longer annotations in the same gap. Analyzing the number of ancestral IS with a Dollo parsimony criterion considering only the extant IS annotations, we have 96 ancestral gaps that contain an IS, indicating a large agreement between the IS that are conserved by the parsimony criterion and the IS supported by aDNA reads.

Conflicting adjacencies. Conflicting adjacencies are related by the marker extremities they share, defining clusters of related conflicting adjacencies. For the reference-based assembly, we identified three such clusters (see Figure S10). Two of them consist of three adjacencies that are all annotated with IS elements, while the other consists of four adjacencies, including two IS-annotated adjacencies. In total, only two of these conflicting adjacencies are supported by aDNA reads. All other adjacencies contain uncovered regions indicating potential breakpoints. So in order to propose a conflict-free scaffolding, we chose to remove all unsupported conflicting adjacencies. Note that filling these gaps only partially does not provide much information, as uncovered regions can be either breakpoints or not sequenced regions of the ancestral genome. For the de novo assembly, there are only two clusters of conflicting adjacencies that match with the clusters observed in the reference-based assembly according to the coordinates of the supporting extant gaps. As the same adjacencies are covered by aDNA reads, we resolve the scaffolding conflicts identically to the reference-based assembly by keeping the two supported adjacencies and removing all other conflicting adjacencies. See Figure S11 for the read coverage of discarded adjacencies.

For the reference-based assembly, the set of ancestral adjacencies can then be ordered into seven Contiguous Ancestral Regions (CARs), while we obtain five CARs for the de novo assembly. We convert the reconstructed sequences of markers back to genome sequences by filling the gaps with the read sequences if possible and resorting to the template sequence otherwise.

As mentioned earlier, we observe two gaps with highly differing extant gap lengths and very little conservation in both reconstructions. While the extant gap coordinates are similar for both gaps, the multiple alignment is in both cases very fragmented and hence the resulting template sequences are not similar, even though they are based on mainly the same extant gap sequences. The mapping of reads onto these templates is poor: in the de novo assembly, the gap contains 211 uncovered regions of 9319 bp in total. See Figure S12 for an overview over the read coverage for this gap in the de novo assembly. As the reconstructed sequences have a high edit distance after partial gap filling, so we cannot reconstruct a coinciding sequence in both reconstructions, we remove these gap sequences completely at this point to avoid dubious and non-robust reconstructed ancestral sequences.

Comparing the two improved assemblies. We compared the two sets of CARs obtained from both initial assemblies by aligning the resulting genome sequences using MUMmer [Kurtz et al., 2004]. As seen in Figure 2, we observe no rearrangements between both resulting sets of CARs, showing that, in terms of large-scale genome organization, the final result does not depend on the initial contig assembly.

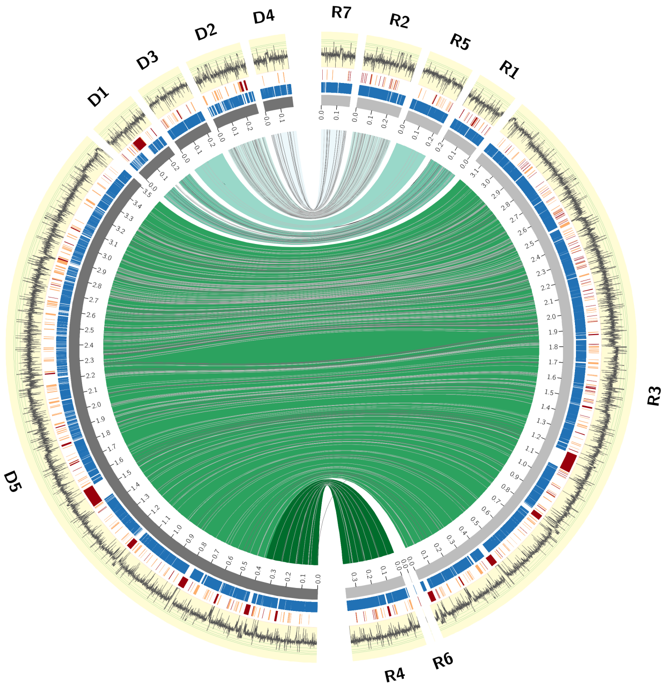

Fig. 2. Comparison between the de novo assembly (left) and the reference-based assembly (right) for the London data set. The inner links connect corresponding CARs in the reconstructions. The grey lines indicate substitutions and InDels observed. The positions in both assemblies covered by markers are indicated in blue. All gaps that have IS annotation in the extant genomes are shown in orange. In addition, gaps that are only partially filled or have very unconserved extant gap lengths are indicated in red. Finally, the most outer ring shows the average read coverage in windows of length 200bp in log-scale. Figure done with Circos [Krzywinski et al., 2009].

In addition, we aligned all reads again to the final assembly to assess the amount of uncovered regions in the reconstructed sequences. In total, only 85,578bp in the reference-based assembly and 88,529bp in the de novo assembly are not covered by any read; however most uncovered regions are rather short (see Figures 2 and S15). Based on this mapping, we ran the assembly polishing tool Pilon [Walker et al., 2014] on the final assembly. It identified several positions where the assembled base (also present in the template) is the minority in comparison to all reads mapping at this position. As Pilon is not taking the respective bases of the extant genomes into account, it runs the risk of correcting the assembly according to sequencing errors in the reads. In fact, the most frequent proposed substitutions correspond to the common damage pattern of cytosine deamination observed in aDNA [Pääbo et al., 2004]. As a consequence, we only keep small indel corrections by Pilon but reject all single-base corrections.

We achieve a high similarity between both sets of CARs. While the improved de novo assembly contains a larger amount of filled gap sequences, we align nearly all of both sequences and observe only a low number of SNPs and insertions and deletions between both assemblies (see Figure 2). The observed differences are often located in gaps with low read coverage regions. If short regions in the gaps are only covered by a single read, in order to find a shortest path in the mappings, this read has to be included at all costs and can cause corrections to the template that are not supported by any other read. Further re-sequencing of these regions could clear which variant is present in the ancient genome.

In the improved reference-based assembly, 78.74% of the resulting sequence is defined by markers and hence directly adopted from the initial assembly, while for the improved de novo assembly only 49.88% of the improved assembly is based on marker sequences and a larger part is based on the filled gap sequences. Together with the gaps that have been filled by read sequences, we can say that for the reference-based assembly in total 94.46% and for the de novo assembly in total 95.25% are reconstructed using only the available aDNA reads.

3.2 The Marseille strain

This data set consists of five samples as described in [Bos et al., 2016] that we assembled separately with Minia [Chikhi and Rizk, 2013] and parameter k=21 (unlike for the Black Death data set, there was no available reference-based assembly). We first compared the quality of the resulting assemblies by mapping contigs with a minimal length to the genome of the extant strain *Yersinia pestis CO92* and summing the total length of the mappings as seen in Figure 3. While restricting the minimal contig length, two of the samples cover an extensively larger part of the reference and thus indicate a better sequencing quality. Figure 4 shows that if we restrict the minimal contig length, only a small part of the reference genomes are covered by contigs from all five samples.

We use the assembly of sample *OBS116* with a minimal contig length of 500bp to segment the extant genomes into markers. The assembly consists of 3,089 contigs with a total length of 3,636,663bp. The segmentation results in 2,859 markers with a total length of 3,143,627bp and we analyze 2,859 potential adjacencies: 27 of these gaps have a length of 0, leaving 2,832

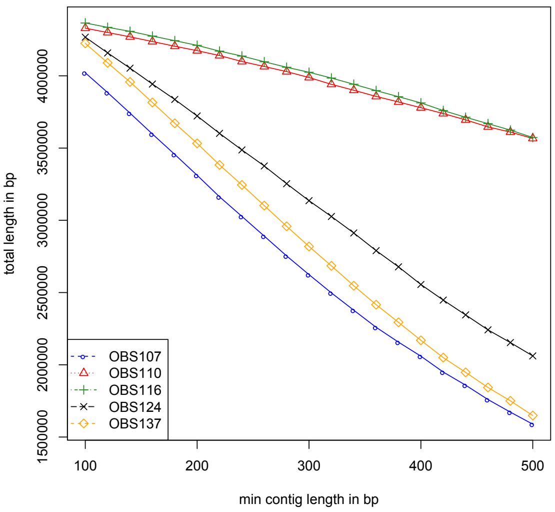

Fig. 3. Total length of contigs mapped to *Yersinia pestis CO92* greater than a minimum contig length.

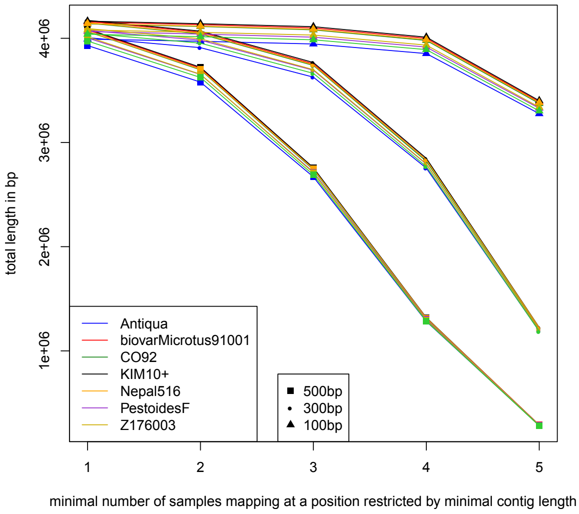

Fig. 4. Comparison of the assembled contigs by mapping to different reference sequences. While most of the references are covered by at least one sample, only a small part of the reference genomes are covered by all five samples.

gaps to fill. Based on the observations above, we joined all sample reads sets for filling the gaps in the reconstruction to achieve a better coverage.

We can see in Figure 5 that with the combined set of reads, we can fill nearly all simple gaps by read sequences. In addition, we obtain a higher number of IS-annotated gaps that are filled in comparison to the London data set. For the IS-annotated gaps, 95 are reconstructed containing the IS, 21 contain IS annotations shorter than 400bp. Hence we identified the same number of potential ancestral IS as for the London strain.

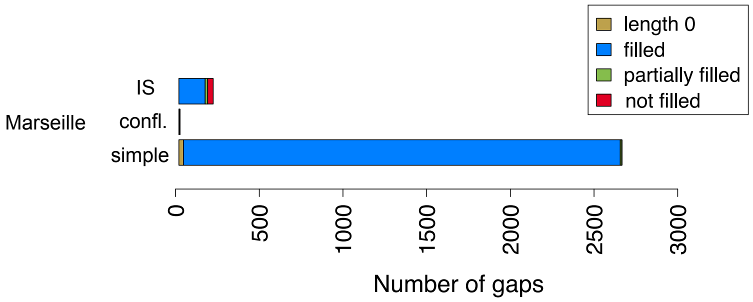

Fig. 5. Result of gap filling for de novo assembly. Note that if a gap is conflicting and IS-annotated, we assign it to the conflicting group. We differentiate between gaps of length 0 (i. e. both markers are directly adjacent), completely and partially filled gaps and not filled gaps.

We identified two conflicting components in this set of potential adjacencies (see Figure S13). Both of them align in terms of gap lengths and extant occurrences with the two components shared by the assemblies for the London strain. In the first component, again only one conflicting adjacency is covered by reads. However, this is a different adjacency in comparison to both reconstructions for the London strain, while on the other hand we have no read support for the gap that is covered in the London data set. This could indicate a potential point of genome rearrangement (see discussion in next section). In the second component, all involved adjacencies are covered by reads from the five samples. In order to obtain a set of high confidence ancestral CARs, we removed all conflicting adjacencies in this component from the set of potential adjacencies. The coverage of all discarded adjacencies is shown in Figure S14.

This results into 6 CARs for the ancestral genome. Again, we used *BWA* [Li and Durbin, 2009] to align reads from all five samples again to the assembly to assess the amount of uncovered regions in the reconstructed sequences. In total, only 54,672bp in this mapping are not covered by any read and the length of the uncovered regions is rather short (see Figure S15).

3.3 Comparison of the London and Marseille strains genomes

As the Marseille *Yersinia pestis* strain is assumed to be a direct descendant of the London Black Death strain [Bos et al., 2016], we aligned the obtained CARs in both reconstructions to identify genome rearrangements. As shown in Figure 6, apart from one larger deletion and one larger insertion in the Marseille strain related to the removed gap sequence in the London strain and a small inversion of length 4138bp marked in black, the reconstructed CARs show no larger rearrangements between both genomes (grey links).

The difference in conflicting adjacencies kept is a possible indication for a rearrangement that however cannot be explicitly identified at this point. It causes the split pattern observed between CAR3 and CAR1 in the London strain and CAR2 and CAR5 in the Marseille strain. Given that the available read data does not allow us to further order the resulting CARs into a single scaffold, additional potential rearrangements could be assumed to be outside of the reconstructed CARs. In contrast, Figure 6 depicts several inversions and translocations between both ancient sets of CARs and the extant *Yersinia pestis CO92* (red and blue links respectively).

3.4 Discussion

In this paper, we present a method to fill the gaps between contigs assembled from aDNA reads that combines comparative scaffolding using related extant genomes and direct aDNA sequencing data, and we apply it to two ancient *Yersinia pestis* strains isolated from the remains of victims of the second plague pandemic.

The comparison of the two assemblies for the London strain illustrates that relying on a shorter initial de novo contig assembly does not impact significantly the final result. The results we obtain with the Marseille data set illustrates that if a good coverage of reads

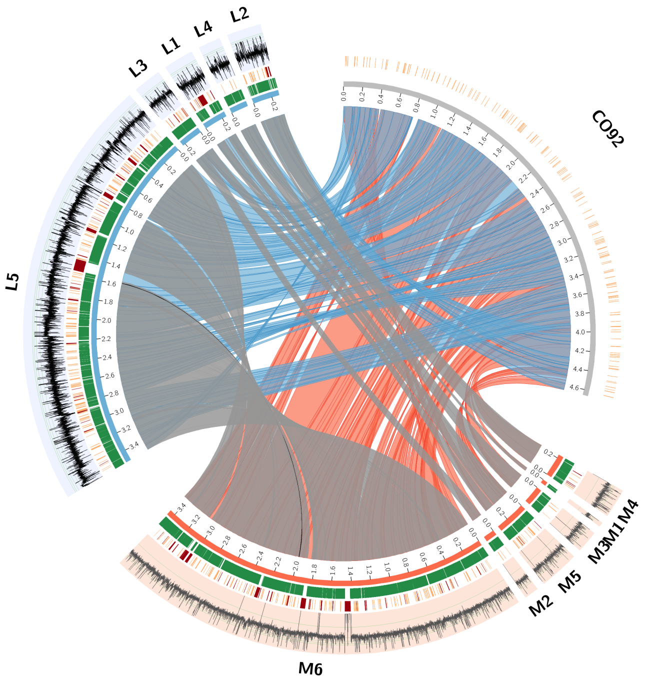

Fig. 6. Comparison between the de novo assembly of the London strain (blue) and the Marseille strain (red) with the reference *Yersinia pestis CO92*. The inner links connect corresponding CARs in the reconstructions and the reference. Note that there is only a small inversion marked in black among the grey links. The positions in both reconstructions covered by markers are indicated in green. All gaps that have IS annotations in the extant genomes are shown in orange. For CO92, all IS annotations are shown as well. In addition, gaps that are only partially filled or have very unconserved extant gap lengths are indicated in red. Finally, the most outer ring shows the average read coverage in windows of length 200bp in log scale. Figure done with Circos [Krzywinski et al., 2009].

over the whole genome can be provided (as through multiple sequencing experiments for multiple samples), even a cautious initial contig assembly can be improved in such a way that most gaps are filled using unassembled aDNA reads. With both data sets, we obtain largely improved genome assemblies, with a reduced fragmentation (from thousand of contigs to a handful of CARs) and a very small fraction of the final assembly that is not supported by aDNA reads.

Applied to the same data set for the London strain, the method FPSAC [Rajaraman et al., 2013] was able to obtain a single scaffold based on parsimonious optimization. Comparing our resulting assembly to this single scaffold, we can identify two breakpoints between both assemblies, hence both methods do not entirely support the same scaffold structure for the London strain. These disagreements should be seen as weak points in our assembly, as they are not reconstructed by different scaffolding objectives and would need to be confirmed more confidently by additional sequencing data.

We see a clear connection between conflicts in the set of potential adjacencies and the presence of IS elements in the corresponding gaps. Solving these conflicts based on aDNA read data provides a useful way to identify ancestral adjacencies in a conflicting component if the quality of the aDNA data is sufficient. The mapping of aDNA reads has shown to be mostly difficult at repetitive regions like Insertion Sequences, where the presence of the IS in the ancestral gap cannot be clearly detected by the aDNA sequencing data.

Interestingly, the improved assemblies of the London and Marseille strains show no explicit large genome rearrangements except for a small inversion. Even if potential genome rearrangement might not be observed due to the fragmentation of the assemblies into CARs, the synteny conservation between two strains separated by roughly 400 years of evolution is striking compared to the level of syntenic divergence with extant strains. This might be explained by the fact that both the London and Marseille strains belong to a relatively localized, although long-lasting, pandemic [Bos et al., 2016]. Also of interest is the observation that conflicting adjacencies in the Marseille data set were covered by aDNA reads, thus making it difficult to infer robust scaffolding adjacencies; this raises the question of the presence of several strains in the Marseille pandemic that might have differed by one or a few inversions.

Answering these questions with confidence would require additional targeted sequencing of a few regions of the genomes of the London and Marseille strains, or the sequencing of additional strains of the second plague pandemic, such as the *Yersinia pestis* genome sequenced from plague victims in Ellwangen [Spyrou et al., 2016] which is assumed to be an ancestor of the Marseille strains.

Authors contribution. C.C. designed the study. N.L. and C.C. designed the AGapEs method. N.L. implemented the method and analyzed data. D.D. designed and implemented the IS annotation method. C.C. and N.L. wrote the manuscript. All authors read and approved the paper.

Acknowledgements. We thank the group of Hendrik Poinar for help with the processing of the raw reads and discussions during a visit in Hamilton.

References

[Bos et al., 2014] Bos, K. I., Harkins, K. M., Herbig, A., Coscolla, M., Weber, N., Comas, I., Forrest, S. A., Bryant, J. M., Harris, S. R., Schuenemann, V. J., et al. (2014). Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature*, 514:494–497.

[Bos et al., 2016] Bos, K. I., Herbig, A., Sahl, J., Waglechner, N., Fourment, M., Forrest, S. A., Klunk, J., Schuenemann, V. J., Poinar, D., Kuch, M., et al. (2016). Eighteenth century yersinia pestis genomes reveal the long-term persistence of an historical plague focus. *eLife*, page e12994.

[Bos et al., 2011] Bos, K. I., Schuenemann, V. J., Golding, G. B., Burbano, H. A., Waglechner, N., Coombes, B. K., McPhee, J. B., DeWitte, S. N., Meyer, M., Schmedes, S., Wood, J., Earn, D. J. D., Herring, D. A., Bauer, P., Poinar, H. N., and Krause, J. (2011). A draft genome of yersinia pestis from victims of the black death. *Nature*, 478:506–510.

[Chain et al., 2004] Chain, P. S., Carniel, E., Larimer, F. W., Lamerdin, J., Stoutland, P., Regala, W., Georgescu, A., Vergez, L., Land, M., Motin, V., et al. (2004). Insights into the evolution of yersinia pestis through whole-genome comparison with yersinia pseudotuberculosis. *Proceedings of the National Academy of Sciences of the United States of America*, 101:13826–13831.

[Chikhi and Rizk, 2013] Chikhi, R. and Rizk, G. (2013). Space-efficient and exact de bruijn graph representation based on a bloom filter. *Algorithms for Molecular Biology*, 8:1.

[Darling et al., 2008] Darling, A., Miklós, I., and Ragan, M. (2008). Dynamics of genome rearrangement in bacterial populations. *PLoS Genetics*, 4:e1000128.

[Devault et al., 2014] Devault, A. M., Golding, G. B., Waglechner, N., Enk, J. M., Kuch, M., Tien, J. H., Shi, M., Fisman, D. N., Dhody, A. N., Forrest, S., Bos, K. I., Earn, D. J. D., Holmes, E. C., and Poinar, H. N. (2014). Second-Pandemic Strain of Vibrio cholerae from the Philadelphia Cholera Outbreak of 1849. *New Engl Journal Medicine*.

[Dijkstra, 1959] Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, 1:269–271.

[Eddy et al., 2009] Eddy, S. R. et al. (2009). A new generation of homology search tools based on probabilistic inference. In *Genome Informatics*, volume 23, pages 205–211.

[Edgar, 2004] Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–1797.

[Fitch, 1971] Fitch, W. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology*, 20:406–416.

[Gasc et al., 2016] Gasc, C., Peyretaillade, E., and Peyret, P. (2016). Sequence capture by hybridization to explore modern and ancient genomic diversity in model and nonmodel organisms. *Nucleic Acids Research*, Epub ahead of print.

[Hagelberg et al., 2015] Hagelberg, E., Hofreiter, M., and Keyser, C. (2015). Ancient dna: the first three decades. *Philosophical Transactions of the Royal Society B*, 370:20130371.

[Hofreiter et al., 2015] Hofreiter, M., Paijmans, J. L. A., Goodchild, H., Speller, C. F., Barlow, A., Fortes, G. G., Thomas, J. A., Ludwig, A., and Collins, M. J. (2015). The future of ancient dna: Technical advances and conceptual shifts. *Bioessays*, 37:284–293.

[Kay et al., 2014] Kay, G. L., Sergeant, M. J., Giuffra, V., Bandiera, P., Milanese, M., Bramanti, B., Bianucci, R., and Pallen, M. J. (2014). Recovery of a medieval brucella melitensis genome using shotgun metagenomics. *MBio*, 5(4):e01337–14.

[Kolmogorov et al., 2014] Kolmogorov, M., Raney, B. J., Paten, B., and Pham, S. K. (2014). Ragout - a reference-assisted assembly tool for bacterial genomes. *Bioinformatics*, 30:302–309.

[Krzywinski et al., 2009] Krzywinski, M. I., Schein, J. E., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*.

[Kurtz et al., 2004] Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, 5:R12.

[Li and Durbin, 2009] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25:1754–1760.

[Maixner et al., 2016] Maixner, F., Krause-Kyora, B., Turaev, D., Herbig, A., Hoopmann, M. R., Hallows, J. L., Kusebauch, U., Vigl, E. E., Malfertheiner, P., Megraud, F., et al. (2016). The 5300-year-old helicobacter pylori genome of the iceman. *Science*, 351(6269):162–165.

[Martin et al., 2013] Martin, M. D., Cappellini, E., Samaniego, J. A., Zepeda, M. L., Campos, P. F., Seguin-Orlando, A., Wales, N., Orlando, L., Ho, S. Y., Dietrich, F. S., et al. (2013). Reconstructing genome evolution in historic samples of the irish potato famine pathogen. *Nature Communications*, 4.

[McNally et al., 2016] McNally, A., Thomson, N. R., Reuter, S., and Wren, B. W. (2016). 'add, stir and reduce': Yersinia spp. as model bacteria for pathogen evolution. *Nature Reviews Microbiology*, 14:177–190.

[Orlando et al., 2015] Orlando, L., Gilbert, M. T. P., and Willersev, E. (2015). Reconstructing ancient genomes and epigenomes. *Nature Reviews Genetics*, 16:395–408.

[Pääbo et al., 2004] Pääbo, S., Poinar, H., Serre, D., Jaenicke-Després, V., Hebler, J., Rohland, N., Kuch, M., Krause, J., Vigilant, L., and Hofreiter, M. (2004). Genetic analyses from ancient dna. *Annual Review of Genetics*, 38:645–679.

[Peltzer et al., 2016] Peltzer, A., Jäger, G., Herbig, A., Seitz, A., Kniep, C., Krause, J., and Nieselt, K. (2016). Eager: efficient ancient genome reconstruction. *Genome Biology*, 17:60.

[Rajaraman et al., 2013] Rajaraman, A., Tannier, E., and Chauve, C. (2013). Fpsac: fast phylogenetic scaffolding of ancient contigs. *Bioinformatics*, 29:2987–2994.

[Rasmussen et al., 2015] Rasmussen, S., Allentoft, M., Nielsen, K., Orlando, L., Sikora, M., Sjgren, K.-G., Pedersen, A., Schubert, M., VanDam, A., Kapel, C., Nielsen, H., Brunak, S., Avetisyan, P., Epimakhov, A., Khalyapin, M., Gnuni, A., Kriiska, A., Lasak, I., Metspalu, M., Moiseyev, V., Gromov, A., Pokutta, D., Saag, L., Varul, L., Yepiskoposyan, L., Sicheritz-Pontn, T., Foley, R., Lahr, M., Nielsen, R., Kristiansen, K., and Willerslev, E. (2015). Early divergent strains of yersinia pestis in eurasia 5,000 years ago. *Cell*, 163:571 – 582.

[Rissman et al., 2009] Rissman, A. I., Mau, B., Biehl, B. S., Darling, A. E., Glasner, J. D., and Perna, N. T. (2009). Reordering contigs of draft genomes using the mauve aligner. *Bioinformatics*, 25:2071–2073.

[Salmela et al., 2015] Salmela, L., Sahlin, K., Mäkinen, V., and Tomescu, A. I. (2015). Gap filling as exact path length problem. In *Research in Computational Molecular Biology*, pages 281–292. Springer.

[Schubert et al., 2014] Schubert, M., Ermini, L., Sarkissian, C., Jónson, H., Ginolhac, A., Schaefer, R., and et al. (2014). Characterization of ancient and modern genomes by snp detection and phylogenomic and metagenomic analysis using paleomix. *Nature Protocols*, 9:1056–1082.

[Schuenemann et al., 2013] Schuenemann, V., Singh, P., Mendum, T., Krause-Kyora, B., Jäger, G., Bos, K., Herbig, A., Economou, C., Benjak, A., Busso, P., Nebel, A., Boldsen, J., Kjellström, A., Wu, H., Stewart, G., Taylor, G., Bauer, P., Lee, O., Wu, H., Minnikin, D., Besra, G., Tucker, K., Roffey, S., Sow, S., Cole, S., Nieselt, K., and Krause, J. (2013). Genome-wide comparison of medieval and modern mycobacterium leprae. *Science*, 341:179–183.

[Spyrou et al., 2016] Spyrou, M. A., Tukhbatova, R. I., Feldman, M., Drath, J., Kacki, S., de Heredia, J. B., Arnold, S., Sitdikov, A. G., Castex, D., Wahl, J., et al. (2016). Historical y. pestis genomes reveal the european black death as the source of ancient and modern plague pandemics. *Cell Host & Microbe*, 19(6):874–881.

[Van Domselaar et al., 2005] Van Domselaar, G. H., Stothard, P., Shrivastava, S., Cruz, J. A., Guo, A., Dong, X., Lu, P., Szafron, D., Greiner, R., and Wishart, D. S. (2005). Basys: a web server for automated bacterial genome annotation. *Nucleic acids research*, 33(suppl 2):W455–W459.

[Wagner et al., 2014] Wagner, D. M., Klunk, J., Harbeck, M., Devault, A., Waglechner, N., Sahl, J. W., Enk, J., Birdsell, D. N., Kuch, M., Lumibao, C., et al. (2014). Yersinia pestis and the plague of justinian 541–543 ad: a genomic analysis. *The Lancet Infectious Diseases*, 14:319–326.

[Walker et al., 2014] Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, 9(11):e112963.

[Yoshida et al., 2015] Yoshida, K., Sasaki, E., and Kamoun, S. (2015). Computational analyses of ancient pathogen dna from herbarium samples: challenges and prospects. *Frontiers in plant science*, 6.

[Yoshida et al., 2013] Yoshida, K., Schuenemann, V. J., Cano, L. M., Pais, M., Mishra, B., Sharma, R., Lanz, C., Martin, F. N., Kamoun, S., Krause, J., et al. (2013). The rise and fall of the phytophthora infestans lineage that triggered the irish potato famine. *Elife*, 2:e00731.

[Zerbino and Birney, 2008] Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research*, 18(5):821–829.