# Improving ancient DNA genome assembly (#13078)

First revision

Please read the **Important notes** below, the **Review guidance** on page 2 and our **Standout reviewing tips** on page 3. When ready **submit online**. The manuscript starts on page 4.

## Important notes

### Editor and deadline

Thomas Rattei / 29 Jan 2017

| | |
|---|---|
| **Files** | 1 Tracked changes manuscript(s)<br>1 Rebuttal letter(s)<br>6 Figure file(s)<br>3 Latex file(s)<br>1 Table file(s)<br>Please visit the overview page to **download and review** the files not included in this review PDF. |
| **Declarations** | **One or more DNA sequences were reported.** |

# Review guidelines

Please read in full before you begin

## How to review

When ready **submit your review online**. The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

🗋 You can also annotate this PDF and upload it as part of your review

To finish, enter your editorial recommendation (accept, revise or reject) and submit.

### BASIC REPORTING

- ✓ Clear, unambiguous, professional English language used throughout.
- ✓ Intro & background to show context. Literature well referenced & relevant.
- ✓ Structure conforms to **PeerJ standards**, discipline norm, or improved for clarity.
- ✓ Figures are relevant, high quality, well labelled & described.
- ✓ Raw data supplied (see **PeerJ policy**).

### EXPERIMENTAL DESIGN

- ✓ Original primary research within **Scope of the journal**.
- ✓ Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- ✓ Rigorous investigation performed to a high technical & ethical standard.
- ✓ Methods described with sufficient detail & information to replicate.

### VALIDITY OF THE FINDINGS

- ℹ Impact and novelty not assessed. Negative/inconclusive results accepted. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- ✓ Data is robust, statistically sound, & controlled.
- ✓ Conclusions are well stated, linked to original research question & limited to supporting results.
- ✓ Speculation is welcome, but should be identified as such.

The above is the editorial criteria summary. To view in full visit **https://peerj.com/about/editorial-criteria/**

# 7 Standout reviewing tips

The best reviewers use these techniques

| Tip | *Example* |
|---|---|
| **Support criticisms with evidence from the text or from other sources** | *Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.* |
| **Give specific suggestions on how to improve the manuscript** | *Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).* |
| **Comment on language and grammar issues** | *The English language should be improved to ensure that your international audience can clearly understand your text. I suggest that you have a native English speaking colleague review your manuscript. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult.* |
| **Organize by importance of the issues, and number your points** | *1. Your most important issue*<br>*2. The next most important item*<br>*3. …*<br>*4. The least important points* |
| **Give specific suggestions on how to improve the manuscript** | *Line 56: Note that experimental data on sprawling animals needs to be updated. Line 66: Please consider exchanging "modern" with "cursorial".* |
| **Please provide constructive criticism, and avoid personal opinions** | *I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC* |
| **Comment on strengths (as well as weaknesses) of the manuscript** | *I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.* |

# Improving ancient DNA genome assembly

**Alexander Seitz** Corresp., [1] , **Kay Nieselt** [1]

[1] Center for Bioinformatics (ZBIT), Integrative Transcriptomics, Eberhard-Karls-Universität Tübingen, Tübingen, Germany

Corresponding Author: Alexander Seitz
Email address: alexander.seitz@uni-tuebingen.de

Most reconstruction methods for genomes of ancient origin that are used today require a closely related reference. In order to identify genomic rearrangements or the deletion of whole genes, *de novo* assembly has to be used. However, because of inherent problems with ancient DNA, its *de novo* assembly is highly complicated. In order to tackle the diversity in the length of the input reads, we propose a two-layer approach, where multiple assemblies are generated in the first layer, which are then combined in the second layer. We used this two-layer assembly to generate assemblies for two different ancient samples and compared the results to current *de novo* assembly approaches. We are able to improve the assembly with respect to the length of the contigs and can resolve more repetitive regions.

# Improving ancient DNA genome assembly

**Alexander Seitz**[1] **and Kay Nieselt**[1]

[1]**Center for Bioinformatics (ZBIT), Integrative Transcriptomics,**
**Eberhard-Karls-Universität Tübingen**

Corresponding author:
Alexander Seitz[1]

Email address: alexander.seitz@uni-tuebingen.de

## ABSTRACT

Most reconstruction methods for genomes of ancient origin that are used today require a closely related reference. In order to identify genomic rearrangements or the deletion of whole genes, *de novo* assembly has to be used. However, because of inherent problems with ancient DNA, its *de novo* assembly is highly complicated. In order to tackle the diversity in the length of the input reads, we propose a two-layer approach, where multiple assemblies are generated in the first layer, which are then combined in the second layer. We used this two-layer assembly to generate assemblies for two different ancient samples and compared the results to current *de novo* assembly approaches. We are able to improve the assembly with respect to the length of the contigs and can resolve more repetitive regions.

## INTRODUCTION

The introduction of next generation sequencing (NGS) made large scale sequencing projects feasible (Bentley et al., 2008). Their high throughput allows fast and also cheap sequencing of arbitrary genomic material. It revolutionized modern sequencing projects and made the study of ancient genomes possible (Der Sarkissian et al., 2015). However, the resulting short reads pose several challenges for the reconstruction of the desired genome when compared to the longer Sanger reads (Li et al., 2010; Sawyer et al., 2012). For modern DNA samples, the problem of having only short reads can be mitigated by the sheer volume of sequenced bases and usage of long fragments with paired-end and mate-pair sequencing. The insert size is used to determine the distance between the forward and the reverse read, which are sequenced from both ends of the fragments. These distances can be important for *de novo* assembly as they are used for repeat resolution and scaffolding. However, samples from ancient DNA (aDNA) mostly contain only very short fragments between 44 and 172 bp (Sawyer et al., 2012). Paired-end sequencing of these short fragments therefore often results in overlapping forward and reverse reads (thus actually negative inner mate pair distances). Because of these short fragments, mate-pair sequencing as well as sequencing technologies producing long reads (like PacBio) does not result in the same information gain that can be achieved on modern samples. Additionally, post-mortem damage of aDNA, most importantly the deamination of cytosine to uracil, can result in erroneous base incorporations (Rasmussen et al., 2010). Using reference based approaches, these errors can be detected, as they always occur at the end of the fragments. This is not possible using *de novo* assembly approaches and these errors can lead to mistakes in the assembly. However, treating the sample with *Uracil-DNA Glycosylase* (UDG) can resolve most of these errors (Briggs et al., 2010). Deeper sequencing does not always yield better results as the amount of endogenous DNA contained in aDNA samples is often very low (Sawyer et al., 2012).

In order to achieve a higher content of endogenous DNA, samples are often subject to enrichment using capture methods (Avila-Arcos et al., 2011). The principle of these capture methods relies on selection by hybridization (Maricic et al., 2010). Regions of interest are fixed to probes prior to sequencing. These probes can be immobilized on glass slides, called array capture (Hodges et al., 2007), or recovered by affinity using magnetic beads, referred to as in-solution capture (Gnirke et al., 2009). Using these capture methods, only DNA fragments that can bind to the probes are used for amplification, which increases the amount of the desired DNA. However, as these methods amplify sequences that are contained on the probes, regions that were present in ancient samples and lost over time are not amplified and thus cannot be identified as they are not specifically targeted (Khan et al., 2013). Nevertheless, many aDNA projects use these capture methods (Shapiro and Hofreiter, 2014).

Currently, there are two ways to reconstruct a genome from sequencing data, *de novo* and reference-based (Hofreiter et al., 2015). If there is a known, closely related genome, it can be used as a reference. Mapping programs like BWA (Li and Durbin, 2009) can then be used to align the reads against the reference genome. Single nucleotide variations (SNVs) or short indels between the DNA sequence of the sample and reference can be identified after all reads are aligned.

Because of the inherent characteristics of aDNA, specialized mapping pipelines for the reconstruction of aDNA genomes, such as EAGER (Peltzer et al., 2016) and PALEOMIX (Schubert et al., 2014), have recently been published. The mapping against a reference genome allows researchers to easily eliminate non-endogenous DNA and identify erroneous base incorporations. These errors can be identified after the mapping (e.g. by mapDamage (Ginolhac et al., 2011) or PMDtools (Skoglund et al., 2014)) and used to verify that the sequenced fragments stem from ancient specimen.

The reference-based mapping approaches cannot detect large insertions or other genomic architectural rearrangements. In addition, if the ancient species contained regions that are no longer present in the modern reference, these cannot be identified via mapping against modern reference genomes. In these cases a *de novo* assembly of the genome should be attempted. This is also true for modern samples, if no closely related reference is available. The introduction of NGS has lead to new assembly programs, such as SOAPdenovo2 (Luo et al., 2012), SPADES (Bankevich et al., 2012) and many more, that can handle short reads. However, if the ancient sample was sequenced after amplification through capture arrays, genomic regions that are not contained on the probes also can't be identified. Using shotgun sequencing, reads originating from species that colonized the sample post-mortem are often more abundant (Knapp and Hofreiter, 2010). However, if shotgun data are available an effort for assembly can be made to identify longer deletions or genomic rearrangements.

The assembly of modern NGS data is still a challenging problem (Chao et al., 2015) and methods to improve it are still being developed. Among these is ALLPATHS-LG (Gnerre et al., 2011), arguably the winner of the so-called Assemblathon (Earl et al., 2011). ALLPATHS-LG uses the information provided by long fragments from paired-end and mate-pair sequencing to improve the assembly, and has therefore been shown to be one of the best assembly programs that are available today (Utturkar et al., 2014). However, because of the short fragments contained in aDNA samples, this approach is not feasible for aDNA samples and other methods have to be employed.

*De Bruijn* graph assemblers highly rely on the length of the *k*-mer to generate the graph (Li et al., 2012). The choice of an optimal value is even a difficult problem for modern sequencing projects (Durai and Schulz, 2016).

Because of the short fragments of aDNA samples, the sequencing adapter is often partially or fully sequenced. After the adapter is removed, the length of the resulting read is then equal to the length of the fragment. Furthermore, overlapping forward and reverse reads can be merged to generate longer reads, which is usually done in aDNA studies to improve the sequence quality (Peltzer et al., 2016). Thus the length distribution of reads from aDNA samples is often very skewed. This implies that the choice of one single fixed *k*-mer size in *de Bruijn* graph-based assembly approaches is not ideal in aDNA studies. Long *k*-mers miss all reads that are shorter than the value of *k* and shorter *k*-mers cannot resolve repetitive regions.

In order to overcome the problem of the different input read lenghts, we have developed a two-layer assembly approach. In the first layer, the contigs are assembled from short reads using a *de Bruijn* graph approach with multiple *k*-mers. These contigs are then used in the second layer in order to combine overlapping contigs contained in the different assemblies resulting from the first layer. This is done using an overlap-based approach.

The next section contains the methods we used to improve and compare the *de novo* assembly for aDNA samples. In the results section, we used our two-layer assembly to improve the assembly of two ancient DNA samples and compare our approach to different assembly programs.

## METHODS

The general structure of our two-layer assembly approach is to use multiple assemblies in a first layer with different *k*-mers, which are then merged in a second layer assembly using an overlap-based assembly program (see Figure 1).

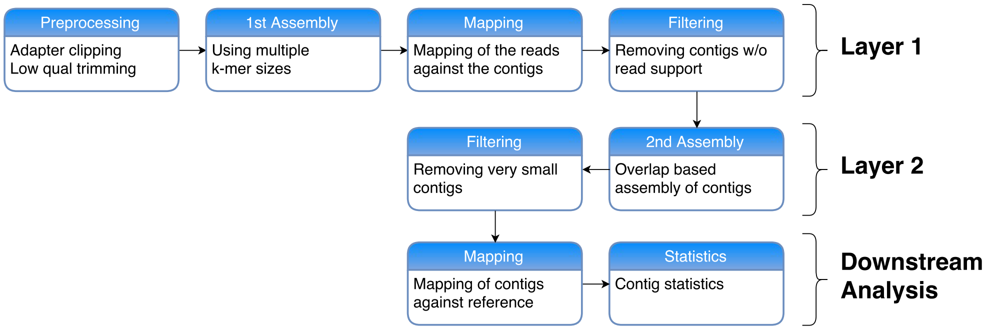

Preprocessing
Adapter clipping
Low qual trimming
1st Assembly
Using multiple k-mer sizes
Mapping
Mapping of the reads against the contigs
Filtering
Removing contigs w/o read support
Layer 1
Filtering
Removing very small contigs
2nd Assembly
Overlap based assembly of contigs
Layer 2
Mapping
Mapping of contigs against reference
Statistics
Contig statistics
Downstream Analysis

**Figure 1.** Workflow of our two-layer assembly approach. First the reads are preprocessed by removing sequenced adapters and clipping low-quality bases. After that, multiple *de novo* assemblies are generated using a *de Bruijn* graph approach with multiple values for *k*. The reads are then mapped back against each of these resulting contigs and the contigs with no read support are filtered out. In Layer 2, these filtered contigs are then combined and assembled again using an Overlap-Layout-Consensus approach. Very short contigs are removed. The resulting contigs are mapped against a reference genome and contig statistics are calculated in order to assess the quality of the assembly.

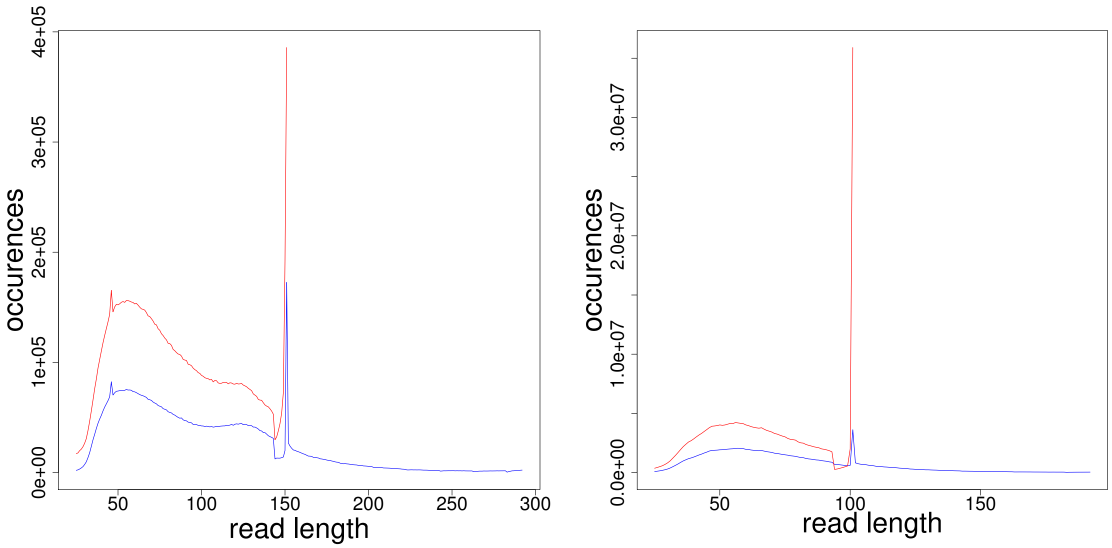
We used the tool *Clip & Merge* (Peltzer et al., 2016) for the preprocessing of the reads. In order to evaluate how different preprocessing affects the assembly, the reads were all adapter clipped, quality trimmed, and then treated using three different methods: First, *Clip & Merge* was used with default parameters to merge overlapping forward and reverse reads. Second, the parameter `-no_merging` was used to perform only adapter-clipping and quality-trimming without the merging of the reads, leaving the paired-end information (reads with no partner were removed). Third, after processing the reads as described in the second method, we gave each read a unique identifier and combined all forward and reverse reads in one file. Here reads without partners were kept. After the first and third method, a single-end assembly was performed, whereas the reads from the second preprocessing method were used in a paired-end assembly.

The different preprocessing methods result in reads of different length. The reason for this is the different fragment lengths contained in the sample. To resolve problems originating from these different lengths, we propose assembly of aDNA using a two-layer approach. In the first layer, we use a *k*-mer based assembly program. For our analysis here, we used SOAPdenovo2 (Luo et al., 2012) and MEGAHIT (Li et al., 2014) in the first layer, but any other assembly program, for which different values for *k* can be chosen, can be used. In order to cover a broad range of *k*-mers representing both short and long reads contained in the input, we used ten different *k*-mer sizes $(37, 47, 57, \ldots, 127)$.

*De Bruijn* based programs first generate all possible *k*-mers based on the input reads. Matching *k*-mers are used to generate the *de Bruijn* graph. This can lead to random overlaps of *k*-mers contained in different reads and therefore to read incoherent contigs (Myers, 2005). To filter out the contigs generated by random overlaps, we used BWA-MEM (Li, 2013) to map the reads against contigs. Contigs that are not supported by any read were removed before the next step. After removing contigs with no read support, the contigs were then reassembled with SGA. To identify contigs belonging to the desired genome, the results were mapped against the respective reference genome using BWA-MEM and extracted in order to compare the different assemblies.

To merge the results of the different assemblies of the first layer, each contig is given a unique identifier before they are combined into one file. This file is the input of the second layer assembly. Here, the assembly is based on string overlaps instead of *k*-mers, a concept originally introduced by Myers (2005). An assembly program that uses this approach is the String Graph Assembler (SGA) (Simpson and Durbin, 2012). It efficiently calculates all overlaps of the input using suffix arrays (Manber and Myers, 1993). These overlaps are then used to generate an overlap graph and the final contigs are generated based on this graph. We used this method to merge the contigs from the different assemblies based on their overlap.

As SGA uses string-based overlaps and modern sequencing techniques are not error-free, it provides steps to correct for these errors. There is a preprocessing step that removes all bases that are not A,G,C or T. There is also a correction step that performs a *k*-mer based error correction and a filtering step that removes input reads with a low *k*-mer frequency. Because the input for SGA are already pre-assembled



(a) Jorgen625 (*Mycobacterium leprae*)

(b) OBS137 (*Yersinia pestis*)

**Figure 2.** Read length distribution for the different preprocessed `fastq` files. red: RAW reads, blue: reads after merging.

contigs, these errors should already be averaged out (Schatz et al., 2010) and these steps were not used for the assembly of the second layer. However, the assemblies with the different *k*-mers produce similar contigs, which is why the duplicate removal step of SGA is performed. SGA can also use the Ferragina Manzini (FM) index (Ferragina and Manzini, 2000) to merge unambiguously overlapping sequences, which was used to further remove duplicate information. Afterwards the overlap graph was calculated and the new contigs were assembled. All these steps were performed using the standard parameters provided by SGA. Afterwards, contigs shorter than 1 000 bp were removed from the final assembly. In order to evaluate our two-layer assembly method, the resulting contigs were then aligned with the reference genome of interest. We used again BWA-MEM for this step. Finally various statistics for the assembly were computed.

To evaluate our approach, the results are compared to other *de Bruijn* assembly programs that can use information from multiple *k*-mer sizes to generate their assembly graph. Both SOAPdenovo2 and MEGAHIT can use the information from several *k*-mers, which is why we also evaluate against these results. Additionally, we use the "interactive *de Bruijn* graph de novo assembler" (IDBA) (Peng et al., 2010), in order to get results from an assembly program that was not part of our two-layer assembly evaluation and also uses multiple *k*-mers for the generation of the assembly graph. To evaluate the results using only an overlap-based approach, we also assembled the preprocessed input reads directly with SGA.

## RESULTS

In order to evaluate our two-layer assembly approach, we applied it to two different published ancient samples. One is the sample Jorgen625, published by Schuenemann et al. (2013) containing DNA from ancient *Mycobacterium leprae*, the other one is the sample OBS137, published by Bos et al. (2016) containing DNA from ancient *Yersinia pestis*. There are two sequencing libraries available for the sample Jorgen625. In order to evaluate the two leprosy libraries as well as the OBS137 sample, we used the EAGER pipeline (Peltzer et al., 2016) to map the libraries against the respective reference genome (*Mycobacterium leprae TN* and *Yersinia pestis CO92*). One of the two libraries of Jorgen625 contained relatively long fragments with a mean fragment length of 173.5 bp and achieved an average coverage on the reference genome of 102.6X. The other library was sequenced on an Illumina MiSeq with a read length of 151 bp. It was produced from shorter fragments with a mean fragment length of 88.1 bp and a mean coverage of 49.3X. With its shorter fragments and lower achieved coverage, the second library better reflects typical sequencing libraries generated from aDNA samples (Sawyer et al., 2012), so we focused our experiments on this library. The OBS137 sample was sequenced on an Illumina HiSeq 2000

**Table 1.** Results using our two-layer assembly with SOAPdenovo2 and MEGAHIT compared to the separate assemblies of SGA, SOAPdenovo2, MEGAHIT and IDBA. The results show only values for contigs that could be mapped against the respective reference genome. Only the best assemblies (w.r.t. the longest mapped contig) for the different preprocessing methods and *k*-mers are shown. "SOAP" represents the results using multiple *k*-mers for the generation of their graph structure. "MEGAHIT" and "IDBA" alone also represent an assembly using multiple internal *k*-mers. The assemblies next to "Lyr X" represent the best assemblies generated by our approach in Layer X=1 or 2. Preprocessing refers to how the reads were preprocessed before assembly and gaps represent the number of gaps that result after the contigs were mapped against the reference genome. Values in bold represent the best value that could be achieved. All other statistical values can be found in the supplementary material.

| | name | preprocessing | # contigs | N50 | mean contig length | longest contig | # gaps |
|---|---|---|---|---|---|---|---|
| | Jorgen625 (*Mycobacterium leprae*) | | | | | | |
| separate | SOAP | single | 249 | 21909 | 13210.3 | 99866 | 103 |
| | MEGAHIT | merged | 175 | 28410 | 16777.5 | 91499 | 106 |
| | IDBA | paired | 164 | 35419 | 20152.7 | 118220 | 118 |
| | SGA | single | 1157 | 2199 | 1997.3 | 8640 | 952 |
| Lyr 1 | SOAP K57 | single | 215 | 24962 | 14918.6 | 72345 | 120 |
| | MEGAHIT K77 | merged | 253 | 21863 | 12765.4 | 87880 | 108 |
| Lyr 2 | SOAP + SGA | single | **133** | **42136** | **25225.0** | **135656** | 88 |
| | MEGAHIT + SGA | merged | 668 | 19758 | 12245.3 | 109259 | **80** |
| | OBS137 (*Yersinia pestis*) | | | | | | |
| separate | SOAP | single | 1745 | 2263 | 2098.9 | 8641 | 1034 |
| | MEGAHIT | merged | 1090 | 4042 | 3267.1 | 9972 | 640 |
| | IDBA | merged | **779** | **5196** | **3839.1** | 9988 | 498 |
| | SGA | merged | 3 | 1126 | 1291.7 | 1633 | 6 |
| Lyr 1 | SOAP K47 | merged | 91112 | 131 | 118.3 | 6425 | 901 |
| | MEGAHIT K77 | single | 4940 | 1321 | 898.6 | 6307 | 1980 |
| Lyr 2 | SOAP + SGA | merged | 1960 | 2633 | 2281.0 | **13420** | 842 |
| | MEGAHIT + SGA | single | 3104 | 1884 | 1816.7 | 11478 | 967 |

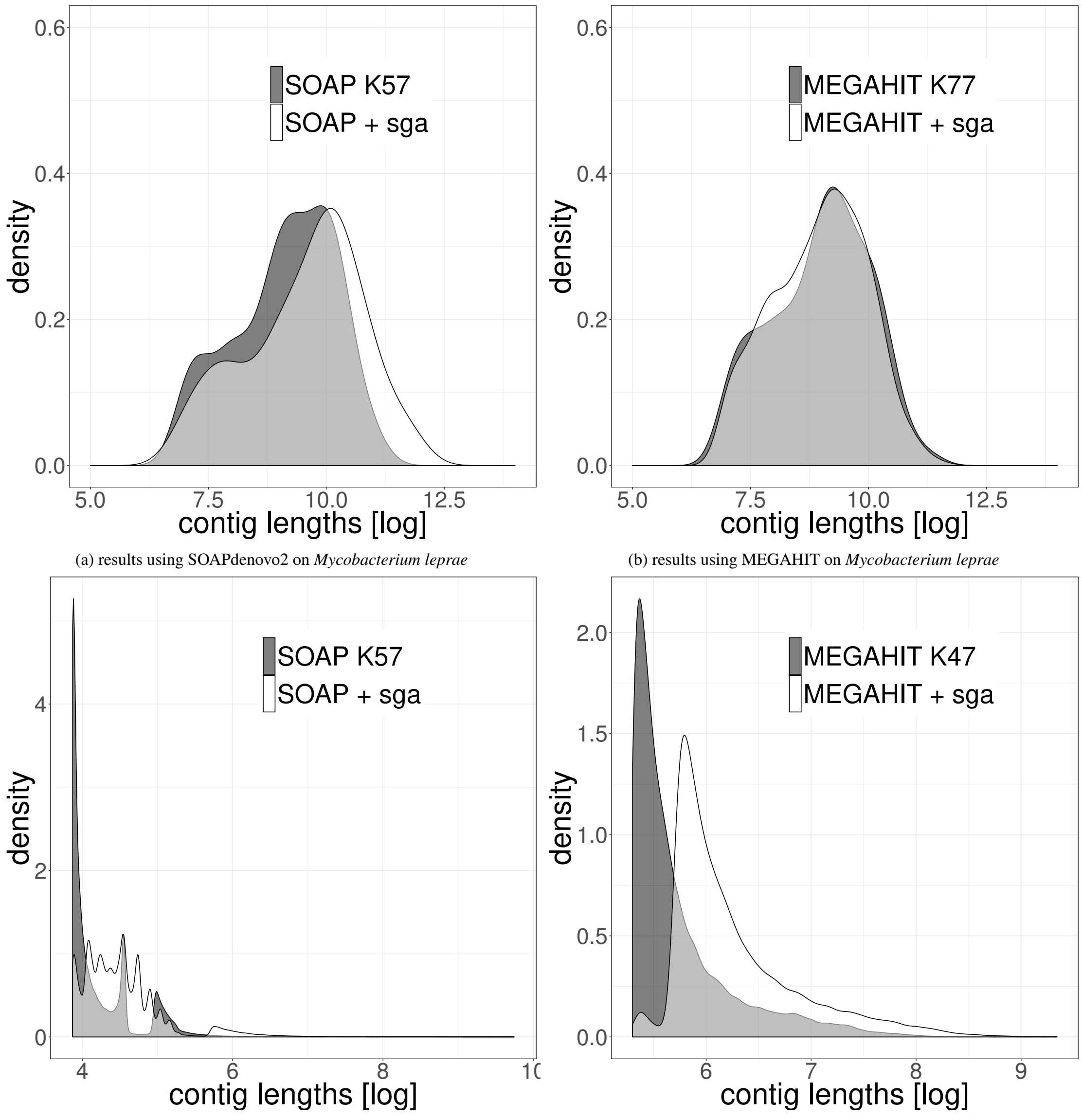
with a read length of 101 bp. The mean fragment length of this library is 69.2 bp and achieved a mean coverage of 279.5X. It is important to note that the leprosy data was generated using shotgun sequencing, whereas the pestis data was first amplified using array capture methods. Both samples were treated with UDG.

The distribution of the read lengths after the preprocessing steps (see Figure 2) shows that the resulting read lengths are highly variable. The peak at read length 151 (in the leprosy case) and 101 (in the pestis case), respectively, are attributed to those reads that were sequenced from fragments longer than the read length. For these no adapter and no low quality bases had to be removed. Therefore, after preprocessing they have the original read length performed in the respective experiment.

For the comparison of the different assembly programs, we extracted the contigs that can be mapped against the respective reference genome (*Mycobacterium leprae TN* and *Yersinia pestis OBS137*, resp.) and calculated several statistics (see Table 1). The results that were generated in the second layer are shown as well as the assembly that generated the longest contig in the first layer using the respective assembly program. Additionally, results from SGA applied to the reads themselves as well as results from programs that can use multiple *k*-mers in their assembly are shown. The complete result table with all intermediate steps is available in the supplementary material.

For both samples, the longest contig, the N50, and the mean contig length could be improved by up to 100% by our two-layer approach. On the leprosy sample, the best result was achieved using all clipped input reads in one single-end assembly without merging. On the pestis sample, the best result was achieved using the merged input reads. Using both SOAPdenovo2 and MEGAHIT with multiple *k*-mers for the generation of the assembly graph, the overall assembly was improved by up to 30% compared to the single *k*-mer assembly. Using SGA directly on the preprocessed reads did not result in good assembly results when compared to SOPA, MEGAHIT or IDBA. IDBA produced the best results when compared to any other assembly using only one layer. On the pestis data, it also overall produced the best results except when comparing the length of the longest contig. Here the longest contig produced by our two-layer approach was up to 35% longer than the one computed by IDBA. For the the leprosy data, all statistical metrics were lower compared to our two-layer assembly.

The length distribution of the resulting leprosy contigs shows a clear shift towards longer contigs (see Figure 3). Because the contigs generated from the pestis data were very short, we did not filter them for a



(a) results using SOAPdenovo2 on *Mycobacterium leprae*

(b) results using MEGAHIT on *Mycobacterium leprae*

(c) results using SOAPdenovo2 on *Yersinia pestis*

(d) results using MEGAHIT on *Yersinia pestis*

**Figure 3.** Distribution of the length of the contigs generated by the different assemblies. The results generated by the second layer assembly with SGA are shown in white. The results of one first layer assembly is shown in dark grey. The light grey part represents the overlap of both methods. 3a shows the results using SOAPdenovo2 in the first layer and 3b shows the results using MEGAHIT in this layer for the leprosy data. 3c and 3d show the same results on the pestis data. In order to highlight the differences, the data was logarithmized.

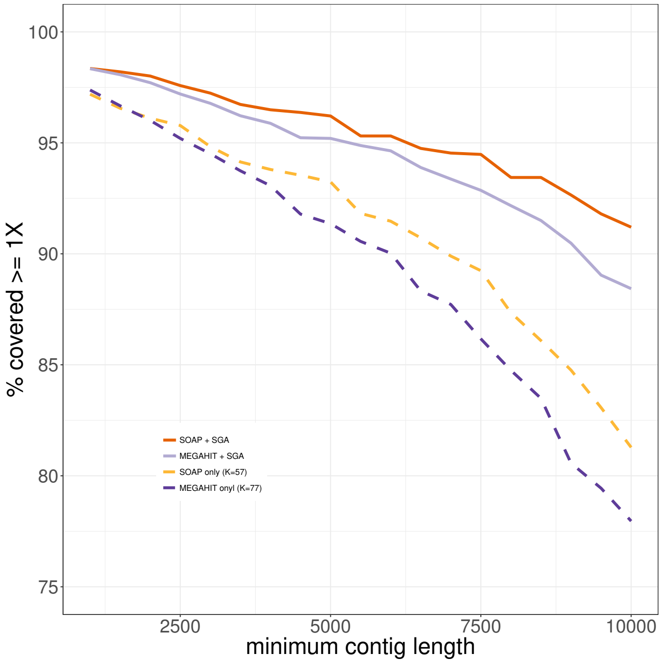


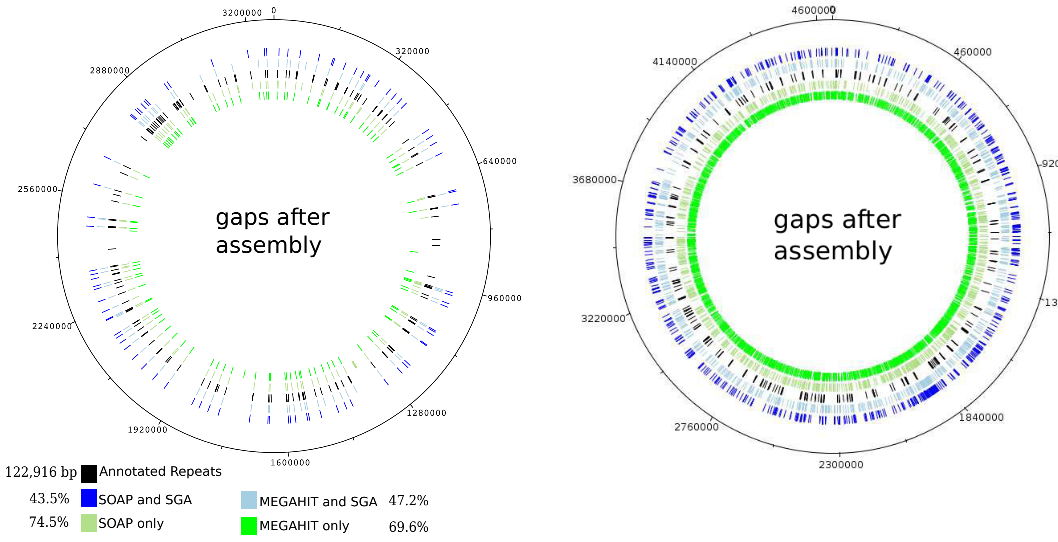
**Figure 4.** The percentage of the reference genome of *Mycobacterium leprae TN* that could be covered using only contigs longer than the minimum contig length. Results from the first and second layer assemblies are shown.

minimum length of 1000 bp. It can be seen that even when all contigs are used, there is a shift towards longer contigs after our two-layer assembly method.

Since one normally is interested in one genome of interest, we computed the genome coverage after mapping all contigs of length at least 1000 bases against the reference genome of *Mycobacterium leprae TN*. We used Qualimap2 (Okonechnikov et al., 2015) for the analysis of the mapping. We also analyzed the coverage of the leprosy genome, that could be achieved using only contigs longer than $1,000, 1,500, \ldots, 10,000$ bp (see Figure 4). It shows that the percentage of the genome that could be covered is always higher after the second layer assembly than using only the results generated in the first layer assemblies. This becomes more and more pronounced with increasing filter threshold for the minimum contig length. When using only contigs longer than $1,000$ bp, the results are almost the same. Using only contigs longer than $10,000$ bp, around 90% of the genome can be covered using the second layer assembly with SGA, whereas at most 80% of the genome is covered by contigs from assemblies generated in the first layer. This means that the same percentage of coverage of the reference genome can be achieved with longer contigs in comparison to the results generated in the first layer. When filtering the pestis data for contigs with a minimum length of 1,000 bp, the best coverage by assemblies of the first layer that could be achieved was 60%. The coverages that could be achieved by the second layer assemblies range between 70 and 83%, where each assembly improved on the ones of the first layer by at least 16% (see supplementary material). Analyzing the mapped contigs that were generated by the second layer, we found that they mapped almost perfectly (with some small insertions and deletions) against the reference genome.

The percentage of the genome that was covered more than once is around 1% for the assemblies generated in the first layer with SOAPdenovo2 and MEGAHIT. This value has increased after the second layer assembly where the contigs were assembled again with SGA, showing that not all overlapping contigs could be identified and merged by SGA.

The mapping of the contigs generated by the first layer assemblies of SOAPdenovo2 and MEGAHIT against the reference genome of *Mycobacterium leprae TN* resulted in 108 and 120 gaps, depending on the assembly program (see Table 1). These values were reduced to 80 and 88 gaps, respectively, for the contigs generated by the second layer assembly with SGA. It can be seen that for the leprosy genome, the gaps in the mapping of the contigs mainly coincide with annotated repeat regions in the reference



(a) *Mycobacterium leprae TN*

(b) *Yersinia Pestis CO92*

**Figure 5.** Gaps in the mapping of the contigs against the reference genome of *Mycobacterium leprae TN* together with annotated repeat regions in the reference genome. The outer ring represents the gaps that occur after the mapping of the contigs that were generated by the second layer assembly with SGA after a first layer assembly with SOAPdenovo2. The second outer ring shows the same but for a first layer assembly using MEGAHIT. The middle ring represents the annotated repeat regions of the reference genome. The second inner and innermost ring represent the gaps after using the best individual SOAPdenovo2 and MEGAHIT assemblies, respectively. The percentages represent the relative number of unresolved bases in annotated repeat regions (in total 122,916 bp).

genome, as already shown by Schuenemann et al. (2013) (see Figure 5a). Altogether, the percentage of unresolved repetitive regions has dropped from 74.5% (when using only SOAPdenovo2) down to 43.5% using our two-layer approach.

For the pestis genome, this is not the case, as the resolved regions to not coincide with repetitive regions. However, it is apparent that after our two-layer approach, more genomic regions could be resolved. When analyzing the mapping of the raw reads against the reference genome of *Mycobacterium leprae TN* with Qualimap2 (Okonechnikov et al., 2015), 100% of the genome could be covered at least once and 99-98% of the genome was covered at least five times.

Up until now we showed that we were able to generate long, high quality contigs that can be mapped against the respective reference. Because the leprosy data was generated from shotgun sequencing, we analyzed whether the assembled contigs actually belong to the species of *Mycobacterium leprae* and not to other *Mycobacteria*. For this we took the ten longest contigs from each assembly and used BLASTN (Altschul et al., 1990) available on the NCBI webserver to align the contigs with all the genomes available from the genus *Mycobacterium*. All hits that generated the highest score for all of these 10 contigs belonged to a strain of *Mycobacterium leprae* (data not shown). As the pestis data was generated using a capture approach and *Yersinia pestis* typically can not survive longer than 72 hours in soil (Eisen et al., 2008), the contamination of other *Yersinia* bacteria can be excluded, which is why we did not perform this experiment on the pestis data.

Furthermore, we evaluated the scalability of our pipeline through subsampling. We used the library from the Jorgen625 sample with the longer fragments, as it contained more than twice as many reads (2 × 15,101,591 instead of 2 × 6,751,711 reads). We evaluated the whole pipeline using 1, 2, 5, 10 and all 15.1 million reads. The calculations were performed on a server with 500GB available memory and 32 CPUs of type Intel® XEON® E5-416 v2 with 2.30 GHz. We evaluated the pipeline using four threads wherever parallelization was possible. The results show that the runtime scales linearly with the number of input reads (see Supplementary Figure 1). The time it would take to assemble a human genome using our two-layer approach can be estimated using a linear regression. The ancient human LBK/Stuttgart sample published by Lazaridis et al. (2014) was sequenced using eight lanes, each containing between 200 and 230 million reads. The assembly of one such lane would take approximately one week and the assembly of all 1.74 billion reads almost two months.

# DISCUSSION AND CONCLUSIONS

With ancient genome assembly one faces a number of challenges. The underlying dataset stems from a metagenomic sample with short fragments. When performing a paired-end sequencing experiment, this results in mostly overlapping forward and reverse reads. Because of the highly different read lengths after the necessary preprocessing steps, including adapter removal and quality trimming, typical *de Bruijn* approaches using a fixed *k*-mer size cannot sufficiently assemble the sample. On the other hand, overlap-based approaches alone are also inferior. Our two-layer approach combining various assemblies using different *k*-mer sizes followed by a second assembly based on string overlaps is able to fuse the contigs generated in the first layer into longer contigs and reduce the redundancy. Additionally, we could show that longer, high quality contigs are generated after the second layer assembly. In particular, at least for our example genomes, we are able resolve more gaps. In the example of the *Mycobacterium leprae* genome, these gaps mainly span repetitive regions. The different values for *k* that are used in the first layer assembly lead to similar contigs that can be combined in the second layer assembly. The percentage of the genome that is covered more than once is increased after the second layer assembly of the leprosy data (see supplementary material). This shows that SGA is not able to identify and merge all overlapping contigs. One reason for this could be the underlying metagenomic sample combined with the shotgun sequencing approach. Multiple species in the sample share similar but not identical sequences. As SGA is not designed to assemble metagenomic samples, these differences cannot be distinguished from different sequences of the same genome containing small errors. This theory is supported by the fact that on the pestis data, which was enriched using a capture array, this additional coverage was reduced but not eliminated in comparison to the first layers (see supplementary material). This signifies that when assembling metagenomic and especially aDNA samples, the results always have to be regarded critically in order to avoid mistakes. In order to identify contigs belonging to our desired genome, we mapped them against a closely related reference genome. The contigs that are generated after the second layer map almost perfectly against the reference sequence that is known to be highly similar to the desired genome (Mendum et al., 2014), showing that even though we are assembling a metagenomic sample, the generated contigs of interest are highly specific. However, because of the metagenomic sample, contigs of other species are also present in the assembly and have to be excluded.

Another possibility could be sequencing errors in the sample, leading to distinct contigs using different *k*-mers. However, these errors can be excluded as a possible source of error, as they should be averaged out by the different assemblies (Schatz et al., 2010). Erroneous base incorporations are unlikely to be the source of these distinct contigs, as the sample was treated with *Uracil-DNA Glycosylase* (UDG), removing these errors. However, UDG does not repair methylated sites, so there may still be errors at sites of cytosine methylation (Briggs et al., 2010). Because the assemblies in the first layer are based on the majority of a base call at each position, given a high enough coverage (Schatz et al., 2010), these errors should also be accounted for.

An important step is the preprocessing of the raw reads. We compared the performance using all reads as single reads, as paired reads or as merged reads. However, at least from our study, we can conclude that the results highly depends on the first layer assembler and probably also on the dataset itself. Interestingly, on the leprosy sample, SOAPdenovo2 produces better results when using all input reads in a single-end assembly than in a paired-end assembly. One possible explanation is that the information between the pairs does not contain additional information as almost all paired-end reads overlap and can be merged. It is possible that the program then disregards some overlaps in order to fulfill the paired-end condition. Overlaps that were disregarded this way could be used in the single-end assembly leading to a better assembly. Additionally, reads that did not have a partner were removed before the paired-end assembly. These reads are available in the single-end assembly. It could be that they contained some relevant information. On the pestis sample, the best results were achieved using the merged data. The reason for this is probably the length of the sequenced reads. In order to stay comparable, we used the same settings for the pestis data as for the leprosy data. However, because the pestis sample was sequenced with 101 bp reads, *de Bruijn* graph assemblers using a longer *k*-mer size than 101 bp can't assembly anything. This means that the assemblies in the first layer using a *k*-mer size of 107, 117, and 127 could not produce any results. This does not hold true for the merged data, because the merging of the reads resulted in longer reads (up to 192 bp). So there, these three assemblies could generate some contigs and contribute information to the second layer assembly.

The mapping of the assembled contigs from the leprosy dataset against the reference show that in our case, all gaps align with annotated repeat regions (for the assembly using SOAPdenovo2 in the first layer). Using our two-layer assembly approach, more of these regions could be resolved, but many still remain. In sequencing projects of modern DNA, repetitive regions are resolved using other sequencing technologies such as PacBio. It can produce much longer sequences that span these regions. However, these technologies are not applicable to aDNA as most of the fragments contained in the sample are even shorter than the sequences that can be produced using the Illumina platforms.

In general, it can be concluded that assembly of aDNA is highly dependent on the amount of endogenous DNA in the sample and thus the coverage of each base (Zerbino and Birney, 2008). We are able to improve results generated by current assembly programs. However, the information gain generated by the second layer assembly is dependent on the quality of the first layer assemblies. Thus if the first layer assemblies are of low quality, the second layer assembly cannot improve them significantly. In the example of the pestis data, the second layer assembly could improve on the contigs generated in the first layer assemblies but could not create an almost perfect assembly, as was the case on the leprosy dataset where the contigs in the first layer assemblies were already of high quality. First tests showed that in order to achieve an assembly covering all but the repetitive regions continuously, the input reads should achieve at least a coverage of 10-15X, where more than 90% of the genome should be covered more than 5 times. Of course this is not the only criteria, which can be seen from the pestis data, so more experiments have to be done in order to identify the reasons that make the assembly of a genome possible.

The runtime scales linearly with the number of input reads, which is no problem for small bacterial datasets. Since parallelization of our pipeline is straightforward, assembly of ancient human genome samples will also be feasible.

We have shown that our approach is able to improve the assembly of ancient DNA samples. However, this approach is not limited to ancient samples. In the paper by Arora et al. (2016), we used this two-layer assembly approach on modern, hard to cultivate *Treponema pallidum* samples. The processing of these samples also resulted in only short fragments similar to ancient DNA. There, we were able to identify and verify the missing of a gene using our assembly approach.

## SOFTWARE AVAILABILITY

We have developed an automated software pipeline, written in JAVA which will allow other researchers to use our methodology. This pipeline is available on github:
https://github.com/Integrative-Transcriptomics/MADAM

## ACKNOWLEDGMENTS

We thank Linus Backert and Vladimir Piven for their critical assessment and discussions on different ideas, as well as implementation specific details. Additionally we would to thank Alexander Peltzer, André Hennig, Michael Römer and Niklas Heinsohn for valuable insights and discussions on different ideas, implementation specific questions and for their comments on this paper.

## REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–10.

Arora, N., Schuenemann, V. J. J., Jäger, G., Peltzer, A., Seitz, A., Herbig, A., Strouhal, M., Grillová, L., Sánchez-Busó, L., Kühnert, D., Bos, K. I. I., Rivero Davis, L. R., Mikalová, L., Bruisten, S., Komericki, P., French, P., Grant, P. R. R., Pando, M. A., Gallo Vaulet, L., Rodríguez-Fermepin, M., Martinez, A., Centurión-Lara, A., Giacani, L., Norris, S. J. J., Šmajs, D., Bosshard, P. P. P., González-Candelas, F., Nieselt, K., Krause, J., and Bagheri, H. C. C. (2016). Origin of modern syphilis and emergence of a contemporary pandemic cluster. *bioRxiv*, (December):051037.

Avila-Arcos, M. C., Cappellini, E., Romero-Navarro, J. A., Wales, N., Moreno-Mayar, J. V. V., Rasmussen, M., Fordyce, S. L., Montiel, R., Vielle-Calzada, J.-P., Willerslev, E., and Gilbert, M. T. P. (2011).

Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. *Sci. Rep.*, 1:74.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. a., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. a., and Pevzner, P. a. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19(5):455–477.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. a., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. a., Benoit, V. a., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. a., Brown, R. C., Brown, A. a., Buermann, D. H., Bundu, A. a., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. a., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. a., Kang, G.-D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. a., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. a., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. a., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. a., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., and Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. - Supplement. *Nature*, 456(7218):53–9.

Bos, K. I., Herbig, A., Sahl, J., Waglechner, N., Fourment, M., Forrest, S. A., Klunk, J., Schuenemann, V. J., Poinar, D., Kuch, M., Golding, G. B., Dutour, O., Keim, P., Wagner, D. M., Holmes, E. C., Krause, J., and Poinar, H. N. (2016). Eighteenth century Yersinia pestis genomes reveal the long-term persistence of an historical plague focus. *eLife*, 5(JANUARY2016):1–11.

Briggs, A. W., Stenzel, U., Meyer, M., Krause, J., Kircher, M., and Pääbo, S. (2010). Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic acids research*, 38(6):1–12.

Chao, R., Yuan, Y., and Zhao, H. (2015). Recent advances in DNA assembly technologies. *FEMS Yeast Research*, 15(1):1–9.

Der Sarkissian, C., Allentoft, M. E., Ávila-Arcos, M. C., Barnett, R., Campos, P. F., Cappellini, E., Ermini, L., Fernández, R., da Fonseca, R., Ginolhac, A., Hansen, A. J., Jónsson, H., Korneliussen, T., Margaryan, A., Martin, M. D., Moreno-Mayar, J. V., Raghavan, M., Rasmussen, M., Velasco, M. S., Schroeder, H., Schubert, M., Seguin-Orlando, A., Wales, N., Gilbert, M. T. P., Willerslev, E., and Orlando, L. (2015). Ancient genomics. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 370(1660):20130387.

Durai, D. A. and Schulz, M. H. (2016). Informed kmer selection for de novo transcriptome assembly. *Bioinformatics*, page btw217.

Earl, D., Bradnam, K., St. John, J., Darling, A., Lin, D., Fass, J., Yu, H. O. K., Buffalo, V., Zerbino, D. R., Diekhans, M., Nguyen, N., Ariyaratne, P. N., Sung, W. K., Ning, Z., Haimel, M., Simpson, J. T., Fonseca, N. A., Birol, I., Docking, T. R., Ho, I. Y., Rokhsar, D. S., Chikhi, R., Lavenier, D., Chapuis, G., Naquin, D., Maillet, N., Schatz, M. C., Kelley, D. R., Phillippy, A. M., Koren, S., Yang, S. P., Wu, W., Chou, W. C., Srivastava, A., Shaw, T. I., Ruby, J. G., Skewes-Cox, P., Betegon, M., Dimon, M. T., Solovyev, V., Seledtsov, I., Kosarev, P., Vorobyev, D., Ramirez-Gonzalez, R., Leggett, R., MacLean, D., Xia, F., Luo, R., Li, Z., Xie, Y., Liu, B., Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F. J., Sharpe, T., Hall, G., Kersey, P. J., Durbin, R., Jackman, S. D., Chapman, J. A., Huang, X., DeRisi, J. L., Caccamo, M., Li, Y., Jaffe, D. B., Green, R. E., Haussler, D., Korf, I., and Paten, B. (2011). Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research*, 21(12):2224–2241.

Eisen, R. J., Petersen, J. M., Higgins, C. L., Wong, D., Levy, C. E., Mead, P. S., Schriefer, M. E., Griffith, K. S., Gage, K. L., and Ben Beard, C. (2008). Persistence of Yersinia pestis in soil under natural conditions. *Emerging Infectious Diseases*, 14(6):941–943.

Ferragina, P. and Manzini, G. (2000). Indexing Compressed Text. *Journal of the ACM*, 52(4):552–581.

Ginolhac, A., Rasmussen, M., Gilbert, M. T. P., Willerslev, E., and Orlando, L. (2011). mapDamage: Testing for damage patterns in ancient DNA sequences. *Bioinformatics*, 27(15):2153–2155.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., Sharpe, T., Hall, G., Shea, T. P., Sykes, S., Berlin, A. M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E. S., and Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, 108(4):1513–8.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., Leproust, E. M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Gabriel, S., Jaffe, D. B., Lander, E. S., and Nusbaum, C. (2009). Solution Hybrid Selection with Ultra-long Oligonucleotides for Massively Parallel Targeted Sequencing. *Nature biotechnology*, 27(2):182–189.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M. N., Smith, S. W., Middle, C. M., Rodesch, M. J., Albert, T. J., Hannon, G. J., and McCombie, W. R. (2007). Genome-wide in situ exon capture for selective resequencing. *Nature genetics*, 39(12):1522–7.

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(3):284–293.

Khan, A. A., un Nabi, S. R., and Iqbal, J. (2013). Surface estimation of a pedestrian walk for outdoor use of power wheelchair based robot. *Life Science Journal*, 10(3):1697–1704.

Knapp, M. and Hofreiter, M. (2010). Next generation sequencing of ancient DNA: Requirements, strategies and perspectives. *Genes*, 1(2):227–243.

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Sudmant, P. H., Schraiber, J. G., Castellano, S., Kirsanow, K., Economou, C., Bollongino, R., Fu, Q., Bos, K., Nordenfelt, S., de Filippo, C., Prüfer, K., Sawyer, S., Posth, C., Haak, W., Hallgren, F., Fornander, E., Ayodo, G., Babiker, H. a., Balanovska, E., Balanovsky, O., Ben-Ami, H., Bene, J., Berrada, F., Brisighelli, F., Busby, G. B., Cali, F., Churnosov, M., Cole, D. E., Damba, L., Delsate, D., van Driem, G., Dryomov, S., Fedorova, S. a., Francken, M., Gallego Romero, I., Gubina, M., Guinet, J.-M., Hammer, M., Henn, B., Helvig, T., Hodoglugil, U., Jha, A. R., Kittles, R., Khusnutdinova, E., Kivisild, T., Kučinskas, V., Khusainova, R., Kushniarevich, A., Laredj, L., Litvinov, S., Mahley, R. W., Melegh, B., Metspalu, E., Mountain, J., Nyambo, T., Osipova, L., Parik, J., Platonov, F., Posukh, O. L., Romano, V., Rudan, I., Ruizbakiev, R., Sahakyan, H., Salas, A., Starikovskaya, E. B., Tarekegn, A., Toncheva, D., Turdikulova, S., Uktveryte, I., Utevska, O., Voevoda, M., Wahl, J., Zalloua, P., Yepiskoposyan, L., Zemunik, T., Cooper, A., Capelli, C., Thomas, M. G., Tishkoff, S. a., Singh, L., Thangaraj, K., Villems, R., Comas, D., Sukernik, R., Metspalu, M., Meyer, M., Eichler, E. E., Burger, J., Slatkin, M., Pääbo, S., Kelso, J., Reich, D., and Krause, J. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518):409–413.

Li, D., Liu, C. M., Luo, R., Sadakane, K., and Lam, T. W. (2014). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10):1674–1676.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv*, 00(00):3.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J., and Wang, J. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, 20(2):265–272.

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B., and Fan, W. (2012). Comparison of the two major classes of assembly algorithms: Overlap-layout-consensus and de-bruijn-graph. *Briefings in Functional Genomics*, 11(1):25–37.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y. Y. Y. Y. Y. Y. Y. Y. Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y. Y. Y. Y. Y. Y. Y. Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D. W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J. J., Lam, T.-w., and Wang, J. J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1):18.

Manber, U. and Myers, G. (1993). Suffix Arrays: A New Method for On-Line String Searches. *SIAM Journal on Computing*, 22(5):935–948.

Maricic, T., Whitten, M., and Pääbo, S. (2010). Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS ONE*, 5(11):9–13.

Mendum, T. A., Schuenemann, V. J., Roffey, S., Taylor, G. M., Wu, H., Singh, P., Tucker, K., Hinds, J., Cole, S. T., Kierzek, A. M., Nieselt, K., Krause, J., and Stewart, G. R. (2014). Mycobacterium leprae genomes from a British medieval leprosy hospital: towards understanding an ancient epidemic. *BMC genomics*, 15(1):270.

Myers, E. W. (2005). The fragment assembly string graph. *Bioinformatics*, 21(SUPPL. 2):79–85.

Okonechnikov, K., Conesa, A., and García-Alcalde, F. (2015). Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, 32(2):292–294.

Peltzer, A., Jäger, G., Herbig, A., Seitz, A., Kniep, C., Krause, J., and Nieselt, K. (2016). EAGER: Efficient Ancient Genome Reconstruction. *Genome Biology*, 17(1):60.

Peng, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2010). IDBA - A practical iterative De Bruijn graph De Novo assembler. In *Annual International Conference on Research in Computational Molecular Biology*, volume 6044, pages 426–440. Springer.

Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J. S., Albrechtsen, A., and Others (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, 463(7282):757–762.

Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., and Pääbo, S. (2012). Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS ONE*, 7(3).

Schatz, M. C., Delcher, A. L., and Salzberg, S. L. (2010). Assembly of Large Genomes using Cloud Computing. *Genome research*, 20(9):1165–1173.

Schubert, M., Ermini, L., Der Sarkissian, C., Jónsson, H., Ginolhac, A., Schaefer, R., Martin, M. D., Fernández, R., Kircher, M., McCue, M., Willerslev, E., and Orlando, L. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nature Protocols*, 9(5):1056–1082.

Schuenemann, V. J., Singh, P., Mendum, T. A., Krause-Kyora, B., Jager, G., Bos, K. I., Herbig, A., Economou, C., Benjak, A., Busso, P., Nebel, A., Boldsen, J. L., Kjellstrom, A., Wu, H., Stewart, G. R., Taylor, G. M., Bauer, P., Lee, O. Y.-C., Wu, H. H. T., Minnikin, D. E., Besra, G. S., Tucker, K., Roffey, S., Sow, S. O., Cole, S. T., Nieselt, K., and Krause, J. (2013). Genome-Wide Comparison of Medieval and Modern Mycobacterium leprae. *Science*, 341(6142):179–183.

Shapiro, B. and Hofreiter, M. (2014). A paleogenomic perspective on evolution and gene function: new insights from ancient DNA. *Science*, 343(January):1236573.

Simpson, J. T. and Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. *Genome Research*, 22(3):549–556.

Skoglund, P., Northoff, B. H., Shunkov, M. V., Derevianko, A. P., Pääbo, S., Krause, J., and Jakobsson, M. (2014). Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of Sciences*, 111(6):2229–2234.

Utturkar, S. M., Klingeman, D. M., Land, M. L., Schadt, C. W., Doktycz, M. J., Pelletier, D. A., and Brown, S. D. (2014). Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics*, 30(19):2709–2716.

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.