Parallel detection of ancient pathogens via array-based DNA capture

Kirsten I. Bos1,†, Günter Jäger2,†, Verena J. Schuenemann1, Åshild J. Vågene1, Maria A. Spyrou1, Alexander Herbig1,2, Kay Nieselt2 and Johannes Krause1,3,4

1Department of Archaeological Sciences, 2Centre for Bioinformatics Tübingen (ZBIT), and 3Senckenberg Centre for Human Evolution and Palaeoenvironment, University of Tübingen, Tübingen 72070, Germany
4Max Planck Institute for History and Sciences, Jena 07745, Germany

DNA capture coupled with next generation sequencing is highly suitable for the study of ancient pathogens. Screening for pathogens can, however, be meticulous when assays are restricted to the enrichment of single organisms, which is common practice. Here, we report on an array-based DNA capture screening technique for the parallel detection of nearly 100 pathogens that could have potentially left behind molecular signatures in preserved ancient tissues. We demonstrate the sensitivity of our method through evaluation of its performance with a library known to harbour ancient Mycobacterium leprae DNA. This rapid and economical technique will be highly useful for the identification of historical diseases that are difficult to characterize based on archaeological information alone.

1. Introduction

Next generation sequencing as applied to ancient DNA has improved our understanding of the evolutionary history of pathogenic bacteria and the phylogeographic distribution of their associated diseases [1–6]. Thus far, most genomic analyses have relied on molecular capture techniques designed to target DNA from individual organisms suspected to have left molecular traces in preserved tissues. Candidate biological specimens have been selected based on either historical documentation [1,4,6], a burial context that links them to a past pandemic [5], or diagnostic pathological changes that occur in bone as a result of prolonged infection [2,3]. Such exclusive evidence that links an individual to a specific disease, however, is rare in the archaeological record. Many modern diseases for which genetic information obtained from ancient tissues would be highly valuable are not associated with major pandemics of the past; thus, data to implicate a suspected pathogen in a given skeletal assemblage is often lacking. In addition, physical changes in bone occur in a minority of diseases and form only after prolonged active infection. The vast majority of illnesses leave no physical evidence on the skeleton [7], and even when present, morphological changes often lack the specificity needed for reliable disease classification [8].

Close on the horizon are investigations that will consider co-circulating infectious diseases in the past. Synergistic interactions between seemingly unrelated pathogens that influence transmission and evasion of host immune responses have been demonstrated for several contemporary infections [9], suggesting that facilitative relationships may play a role in pathogen adaptation and evolution. Parallel information of co-infections in historical populations will be highly valuable to investigate the changing landscape of infectious disease through time. In the absence of obvious genetic evidence for differences in virulence between ancient and modern organisms investigated thus far, these themes have been raised potentially to account for the presumed higher rates of infection and greater disease severity in the past [1].

As such, more efficient and more comprehensive screening techniques are needed to identify preserved molecular signals of pathogens in ancient tissues. DNA capture microarrays have shown phenomenal success for pathogen DNA retrieval from highly complex ancient DNA extracts [1–5]; hence we have adapted this technology to screen for not just one, but multiple pathogens at the same time.
Here, we describe an ancient pathogen screening array (APSA) designed for the parallel DNA detection of almost 100 different pathogens that might be found in ancient remains including bacteria, DNA viruses, protozoa, and multicellular organisms. Capture can be performed with multiple libraries in parallel, making this both a highly powerful and economical method for pathogen screening of archaeological material. Our enrichment technique offers advantages over alternative fluorescence-based approaches [10] as captured reads can be subsequently authenticated by evaluation of DNA damage patterns [11] and in some cases by provisional phylogenetic placement. In addition, adjustments in mapping stringency can be applied in downstream analyses to diminish the threat of false-positives. Here, we demonstrate the value of our tool by evaluating its performance with an ancient sample known to harbour preserved *Mycobacterium leprae* DNA.

2. Capture array design

We designed probes for an Agilent 1-million feature array to accommodate unique regions in 92 different pathogens. Candidate pathogens were chosen to represent those that could reasonably be present in archival samples including bone, dental pulp, mummified tissue, or specimens from soft tissue anatomical collections. RNA viruses were omitted in our design owing to their predicted poor preservation in ancient tissues [12]. As most ancient DNA extracts are dominated by molecules derived from environmental sources that can share genomic motifs with pathogenic organisms, care was taken to ensure high specificity in the capture regions. A summary of the array design workflow is shown in figure 1. Selection of capture regions was performed based on NCBI genome sequences.

![Workflow followed in probe design for the ancient pathogen screening array (APSA).](image)

Probes 60 bp in length were generated at 6 bp tiling for each genomic region considered suitable for capture design. Oligos were subsequently filtered for: (i) complexity by removing probes with repetitive elements, (ii) removal of all probes that contained characters other than A, C, T or G, (iii) self-complementarity using *mfold* v. 3.5 [14], and (iv) temperature compatibility retaining only those with melting temperatures ($T_m$) between 60 and 85°C. A total of 974 016 probes are permitted on an Agilent 1-million feature array, and efforts were made to provide equal representation of all pathogens, regardless of genome size. For pathogens with fewer than 30 000 bases remaining after the filtering steps, all candidate probes were included on the array. The remaining probes were distributed evenly between organisms with greater than 30 000 candidates. A list of all pathogens and probe numbers is provided in the electronic supplementary material, table S1.

3. DNA capture

We used a library generated from a leprosy sample (SK8) we reported on in an earlier publication where *M. leprae* DNA was successfully obtained using DNA array-based capture [2]. Libraries were generated by ligating adaptors and unique index identifiers to both ends of the molecules following established protocols [15,16]. Three negative extraction controls were included, as were three library blanks generated by adding water in place of template. Libraries were subsequently amplified [15] to obtain 0.7 μg of library SK8 and 0.9 μg of each negative control, which were serially enriched over two identical capture arrays in parallel (SK8 on one array and the negative controls on another), as described elsewhere [1,17]. Enriched products for SK8 were sequenced on an Illumina MiSeq platform with 2 × 150 + 8 + 6 cycles using the MiSeq reagent kit v. 2 and the manufacturer’s protocol for multiplex sequencing. For higher resolution identification of false-positives, negative controls were subsequently sequenced on an Illumina HiSeq 2500 following the manufacturer’s protocol.

4. Read processing and analysis

Paired-end reads were merged as described elsewhere [2]. Reads that could not be merged were incorporated as single reads, which were subject to an additional quality filtering step that trimmed all reads from the 3'-end until a Phred
score of greater than 20 was obtained for each nucleotide in the sequence. Reads shorter than 30 bp were removed and the remaining reads were mapped to the full genomes of organisms used in APSA probe design with BWA [18] using the aln/samse algorithm with the –n parameter set to 0.1 for increased mapping specificity. To eliminate artefacts, all reads with mapping quality scores below Q37 were removed. Duplicate removal was performed using an in-house modification to the samtools rmdup function [18], which allows duplicate fragments to be removed from both merged and unmerged pools. Reads were subsequently filtered based on their overlap of the template regions, with retention of only those reads that overlapped an array probe sequence by a minimum of 10 bp. Mapping results based on these parameters are reported in the electronic supplementary material, table S2. To reduce the bias of greater probe representation for certain organisms, we further normalized the reads by dividing the number of mapped reads by the number of oligos on the array for the corresponding organism. Normalized data are also reported in the electronic supplementary material, table S2.

The normalized read count tables were visualized using the R programming environment [19]. Figure 2 shows a scatter plot of the normalized read counts and the distribution of reads mapping after duplicate removal to the different pathogens on our array.

5. Analysis of negative controls

Each of our negative controls consisted of between 71 766 and 279 577 shotgun reads and between 44 761 and 991 460 reads after ASPA capture. Mapping results are presented in the electronic supplementary material, table S2. Duplicate removal statistics are not reported because no reads were removed after applying this filter. The majority of APSA-enriched organisms had no detectable reads in our shotgun data at our current sequencing depth; hence enrichment efficiencies could not be calculated. Predictably, greater manipulation of the extraction blanks contributed to an increased number of APSA-captured reads compared with the library blanks. Of those organisms for which enrichment efficiencies could be calculated, the highest enrichment was obtained for Streptococcus pneumoniae (electronic supplementary material, table S2). While this bacterium can reside in the nasopharynx of asymptomatic carriers [20], this enrichment derived from only three unique fragments (electronic supplementary material, table S2); hence potential in-lab contaminants are present in very low quantities. Additional human commensal and soil-dwelling bacterial species such as Veillonella parvula, E. coli, Burkholderia cepacia and Peptostreptococcus magnus received enrichments in the extraction blanks, which indicates that analyses of commensal organisms that can become pathogenic in compromised individuals may be challenging with our DNA capture approach owing to their ubiquity in the environment. In addition, enrichments stemming from up to 25 (non-unique) reads were observed for Trypanosoma cruzi, though normalization reveals that this seemingly high capture is merely an artefact of its higher representation on the array due to its larger genome.

Our library blanks had far fewer mapped reads, with notable enrichments only of Trypanosoma brucei in one blank and V. parvula in another. In all, the low background in our negative controls suggests that our capture approach, coupled with our strict mapping parameters, yields a low number of false-positives.

6. Analysis of the Mycobacterium leprae positive control library

Library SK8 was generated from a mediaeval bone from Winchester in the United Kingdom, and was included in a previous publication that explored genetic diversity in M. leprae [2]. This sample was confirmed to have preserved M. leprae DNA sufficient for genome assembly through array-based enrichment [2]. An analysis of multiple samples revealed that M. leprae DNA has exceptional preservation, likely owing to protection by its lipid-rich cell wall; for this reason, we chose a sample here (SK8) that contained fewer endogenous M. leprae molecules (8.4%) compared with other samples from that study (up to...
Figure 3. Damage plots generated using MapDamage: red indicates C to T transitions, blue indicates G to A transitions. Y-axis denotes percentage of sites containing a nucleotide change from the reference sequence, and x-axis denotes position along the DNA fragment [21]. (Online version in colour.)

40%) in order to mimic a more realistic situation whereby pathogen DNA has low representation in a highly complex DNA library. Shotgun reads generated from 20 μl of extract were obtained from a previous investigation and consisted of 3058 969 reads. Of these, 149 (0.0049%) mapped to genomic regions used in APSA probe design for all organisms on the array (electronic supplementary material, table S2). No reads were filtered out with our duplicate removal, indicating that all mapping reads are unique. Of these 149 reads, 148 mapped to the M. leprae probe template regions.

Our APSA enrichment from the above library generated 213 611 reads, of which 4774 (2.25%) mapped to our capture regions. A visual representation of the captured reads normalized by the number of hits per probe is shown in figure 2. By far the best mapping was to M. leprae which comprised 4756 (99.6%) of the total mapped reads, thus constituting a 460-fold enrichment and 1.16 hits per probe. Duplicate removal revealed that this enrichment derived from 3656 unique reads, constituting 0.89 unique reads per probe. To authenticate our captured fragments as ancient, damage plots were constructed using MapDamage [21], yielding a pattern expected of ancient mycobacterial DNA [2], with 10% C to T damage at the 5’-ends of the molecules (figure 3). To investigate possible phylogenetic signals, the inclusion of previously defined diagnostic M. leprae SNP positions [2] in our APSA probe set was investigated. The number of SNPs between M. leprae strains is low; however, the branch 2-specific SNP position 2 654 119 was present, and our APSA-captured reads indeed showed the expected derived character for sample SK8 at that position [2], albeit at only twofold coverage. While this coverage is too low to permit a robust phylogenetic placement, it provides further evidence that the captured reads are endogenous. The higher ninefold coverage of this SNP reported in our previous genomic work likely resulted from denser 3 bp tiling of the capture probes in comparison to the 6 bp tiling used here, and the richer library used for full genome assembly. While the non-enriched SK8 library had an appreciable amount of M. leprae DNA, mapping of the APSA-captured reads to the leprosy TN reference genome [22] revealed only 856 reads mapping to the 3.2 Mb region excluded in APSA probe design. Duplicate removal reduced this to only 607 reads. This indicates that the 3656 reads mapping to our M. leprae APSA probe regions derived from array enrichment as opposed to non-specific carry-over into the enriched fraction.

Single M. leprae molecules captured in extraction blank 1, and library blanks 2 and 3 indicate that background contamination, post-capture amplification artefact formation, and/or mismapping are very low for this pathogen in our capture regions. Additional molecules were captured for organisms that also had representation in the captured extraction and library blanks, though the levels identified in the SK8 library are lower than those in the negative controls. Aside from M. leprae, the only other organism unique to this sample was Burkholderia pseudomallei, a pathogenic organism responsible for melioidosis in humans; however, this organism can also survive in soil and water [23], and without a proper control sample for comparison, an environmental origin cannot be excluded.

The efficient capture for endogenous ancient M. leprae DNA, which we had previously determined to be present in this sample, provides a good indication that our APSA technique performs well when DNA matching our probes is present, even in small amounts relative to the background.

7. Conclusion

We have successfully demonstrated the efficiency of our capture screening technique through excellent performance with a known positive control, and low background from deep sequencing of negative controls. Reads were authenticated as ancient by both damage profile and phylogenetic placement. We look forward to exploring the uses of our screening approach in future investigations of infectious diseases in the past, and further evaluating its sensitivity in identifying candidate pathogenic organisms amidst proportionately high amounts of environmental contaminants.

Acknowledgements. We thank Michael Taylor and Stewart Cole for providing us access to the leprosy bone sample, and Pushpendra Singh for his participation in library preparation. We thank Alexander Peltzer for implementing the modification to the samtools rmdup function. We thank Anne Stone and Jane Buikstra for helpful comments on an earlier version of the manuscript.

Authors’ contributions. K.I.B. conceived of the investigation, designed the array, designed experiments, analysed data and drafted the manuscript. G.J. designed the array and analysed data. V.J.S. performed experiments. A.J.V. performed experiments. M.A.S. performed experiments. A.H. analysed data. K.N. conceived of the investigation, designed experiments and analysed data. All authors contributed to and approved the final version of the manuscript.
References


