Benchmark

BioTechniques

Comparison of target enrichment strategies for ancient pathogen DNA

Anja Furtwängler*,[1,2], Judith Neukamm[1,3,4], Lisa Böhme[5], Ella Reiter[1], Melanie Vollstedt[5], Natasha Arora[6], Pushpendra Singh[7], Stewart T Cole[8], Sascha Knauf[9,10], Sébastien Calvignac-Spencer[11], Ben Krause-Kyora[5,12], Johannes Krause[1,2,12], Verena J Schuenemann,‡,[1,2,3] & Alexander Herbig***,‡,[1,12]**

*[1]Institute for Archaeological Sciences, Archaeo- & Palaeogenetics, University of Tübingen, 72070 Tübingen, Germany; [2]Senckenberg Centre for Human Evolution & Palaeoenvironment, University of Tübingen, 72070 Tübingen, Germany; [3]Institute of Evolutionary Medicine, University of Zurich, 8057 Zurich, Switzerland; [4]Institute for Bioinformatics & Medical Informatics, University of Tübingen, 72070 Tübingen, Germany; [5]Institute of Clinical Molecular Biology, Kiel University, 24105 Kiel, Germany; [6]Zurich Institute of Forensic Medicine, University of Zurich, 8057 Zurich, Switzerland; [7]Indian Council of Medical Research National Institute of Research in Tribal Health, Jabalpur 482003, MP, India; [8]Institut Pasteur, 75015 Paris, France; [9]Deutsches Primatenzentrum GmbH, Leibniz-Institute for Primate Research, 37077 Goettingen, Germany; [10]Department for Animal Sciences, Georg-August-University, 37075 Goettingen, Germany; [11]Robert Koch Institut, 13353 Berlin, Germany; [12]Max Planck Institute for the Science of Human History, 07745 Jena, Germany ; *Author for correspondence: anja.furtwaengler@uni-tuebingen.de; **Author for correspondence: verena.schuenemann@iem.uzh.ch; ***Author for correspondence: herbig@shh.mpg.de; ‡Authors jointly supervised this study*

BioTechniques 69: 455–459 (December 2020) 10.2144/btn-2020-0100

First draft submitted: 10 July 2020; Accepted for publication: 10 September 2020; Published online: 30 October 2020

ABSTRACT

In ancient DNA research, the degraded nature of the samples generally results in poor yields of highly fragmented DNA; targeted DNA enrichment is thus required to maximize research outcomes. The three commonly used methods – array-based hybridization capture and in-solution capture using either RNA or DNA baits – have different characteristics that may influence the capture efficiency, specificity and reproducibility. Here we compare their performance in enriching pathogen DNA of *Mycobacterium leprae* and *Treponema pallidum* from 11 ancient and 19 modern samples. We find that in-solution approaches are the most effective method in ancient and modern samples of both pathogens and that RNA baits usually perform better than DNA baits.

METHOD SUMMARY

We compared three targeted DNA enrichment strategies used in ancient DNA research for the specific enrichment of pathogen DNA regarding their efficiency, specificity and reproducibility for ancient and modern *Mycobacterium leprae* and *Treponema pallidum* samples. The three methods – array-based capture and in-solution capture with RNA and DNA baits – were all tested in three independent replicates.

KEYWORDS:

ancient DNA ● high-throughput sequencing ● hybridization capture ● *Mycobacterium leprae* ● pathogen DNA ● target enrichment ● *Treponema pallidum*

The field of ancient DNA (aDNA), which studies DNA retrieved from paleontological and archaeological material, was revolutionized by the invention of high-throughput sequencing. In combination with high-throughput sequencing, the development of targeted DNA enrichment protocols has made a crucial contribution in advancing aDNA research during the last decade.

As DNA decays over time, aDNA is usually only present in trace amounts of highly fragmented sequences [1–3]. Detecting endogenous pathogen aDNA from archaeological material is additionally compounded by the larger amount of background DNA from the environment, including soil microorganisms. The background of host DNA in ancient remains is an additional obstacle to obtaining ancient pathogen DNA. Shotgun sequencing of libraries from aDNA extracts to sufficient genomic coverage is therefore cost-intensive [4]. To circumvent this problem, specific regions of interest – such as bacterial chromosomes, mammalian mitochondrial genomes or regions with single-nucleotide polymorphisms – are often target-enriched before sequencing [4]. Aside from its application in aDNA sequencing, targeted DNA enrichment is also useful to retrieve pathogen DNA from clinical samples, particularly for infectious agents that are found in low quantities in the host organism and are difficult to culture, as is the case for *Mycobacterium leprae* and *Treponema pallidum*. Removal of background DNA prior to sequencing increases the yield of pathogen DNA and thus allows information to be obtained that is valuable for epidemiologists investigating outbreaks.

For the enrichment of entire bacterial and mammalian chromosomes, three methods are currently available, which are all based on hybridization capture: DNA microarrays (here represented by SureSelect from Agilent Technologies), in-solution capture with DNA baits (represented by SureSelect from Agilent Technologies, according to Fu and colleagues) and in-solution capture with RNA baits (here represented by myBaits® from Arbor Biosciences) [5,6].

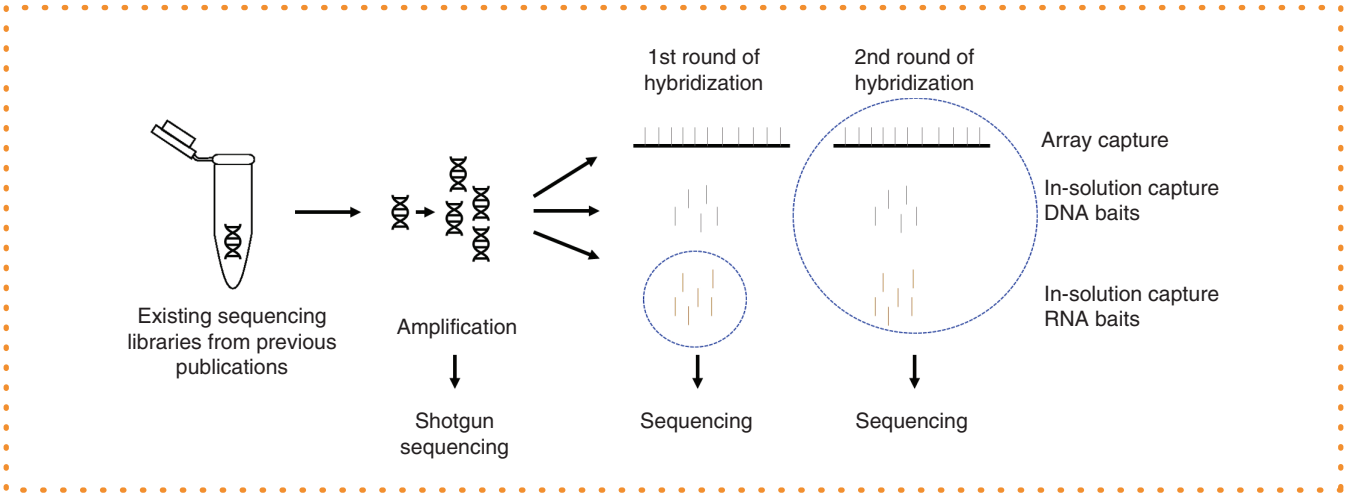

Figure 1. Schematic representation of the workflow. For all samples, the three different enrichment protocols were tested in three independent replicates. Blue circles indicate the libraries that were sequenced at each particular step.

In the case of the DNA array-based method, up to 1 million artificial DNA baits are printed on the surface of a glass slide [7]. Additionally, there is the possibility to perform in-solution capture with baits cleaved from the glass slides and used right away or immortalized in DNA bait libraries [6]. The second in-solution approach uses up to 100,000 artificial RNA baits. These three approaches rely on the hybridization of target fragments to the complementary sequence of the baits (immobilized or in solution), which can be levered to wash background DNA away.

To our knowledge there has been, to date, no statistical comparison of the performance of all three methods. So far only microarrays and the in-solution capture with DNA baits were compared for *Salmonella enterica* and no replicates for statistical assessment were produced [8].

Here we present results from the enrichment of modern and ancient samples containing pathogen DNA, using the three aforementioned approaches. All samples had previously tested positive but had also shown low amounts of target DNA for *M. leprae* or *T. pallidum* (Supplementary Table 1).

The different enrichment concepts tested were chosen to represent the methods as they are applied in ongoing research; thus they differ not only in the technology used (DNA vs RNA baits, immobilized vs in-solution) but also in the design of elements such as bait length and number of unique baits, which might have an effect on the performance.

We used eight ancient samples positive for *M. leprae* and six modern libraries from leprosy patients that were shown to contain *M. leprae* DNA (Supplementary Note 1). Genetic data from the ancient and modern *M. leprae* samples have been published previously [9,10]. Samples with less than 0.6% endogenous bacterial DNA were selected.

Modern *T. pallidum* samples (n = 13) have been published previously [11,12]. Three ancient samples of *T. pallidum* were used [13]. The portion of endogenous DNA for the selected *T. pallidum* samples was below 0.01% for ancient and modern samples.

Starting from existing sequencing libraries, all three methods were applied with three independent replicates each (see Figure 1 and Supplementary Note 1 for a detailed description of the methods; the newly generated data are available at the Sequence Read Archive under the BioProject PRJNA645054). Following the manufacturer's suggestion for libraries with low yields of target DNA, we performed two successive rounds of hybridization for all methods. To investigate the effectiveness of this procedure, we compared results from the first and second round for the in-solution capture with RNA baits. We then evaluated differences in efficiency, reproducibility and specificity across the three approaches by calculating mean coverage, standard deviation of the mean coverage, enrichment factor (calculated by dividing the percentage of target DNA after enrichment by the percentage of target DNA in the shotgun data) and the percentage of the genome covered fivefold or more after normalizing the data of each bacterial species to the same number of raw reads (Supplementary Tables 2, 3 & 5 & Supplementary Figures 1 & 2).

For most ancient samples (8 out of 11), the highest mean coverage (Figure 2A) was reached with the RNA bait in-solution capture (Supplementary Notes 2 & 3 & Supplementary Tables 1 & 2). On average the RNA bait capture resulted in a 1.5- and 20.0-times higher mean coverage than the DNA bait or the array capture, respectively. As illustrated in Figure 2B, the highest enrichment factor was obtained in the RNA bait capture of ancient *T. pallidum* DNA (all three samples) and *M. leprae* (four samples showed best results for the RNA bait, three for the DNA bait and one for the array), with values between 2- and 150-times higher compared with the other two approaches. An in-solution approach seems, therefore, to be advantageous for enriching ancient pathogen DNA.

A similar pattern can be observed in the data of the modern *M. leprae* and *T. pallidum* samples (Figure 2A & B), further highlighting the performance of the in-solution approach in general and RNA baits in particular.

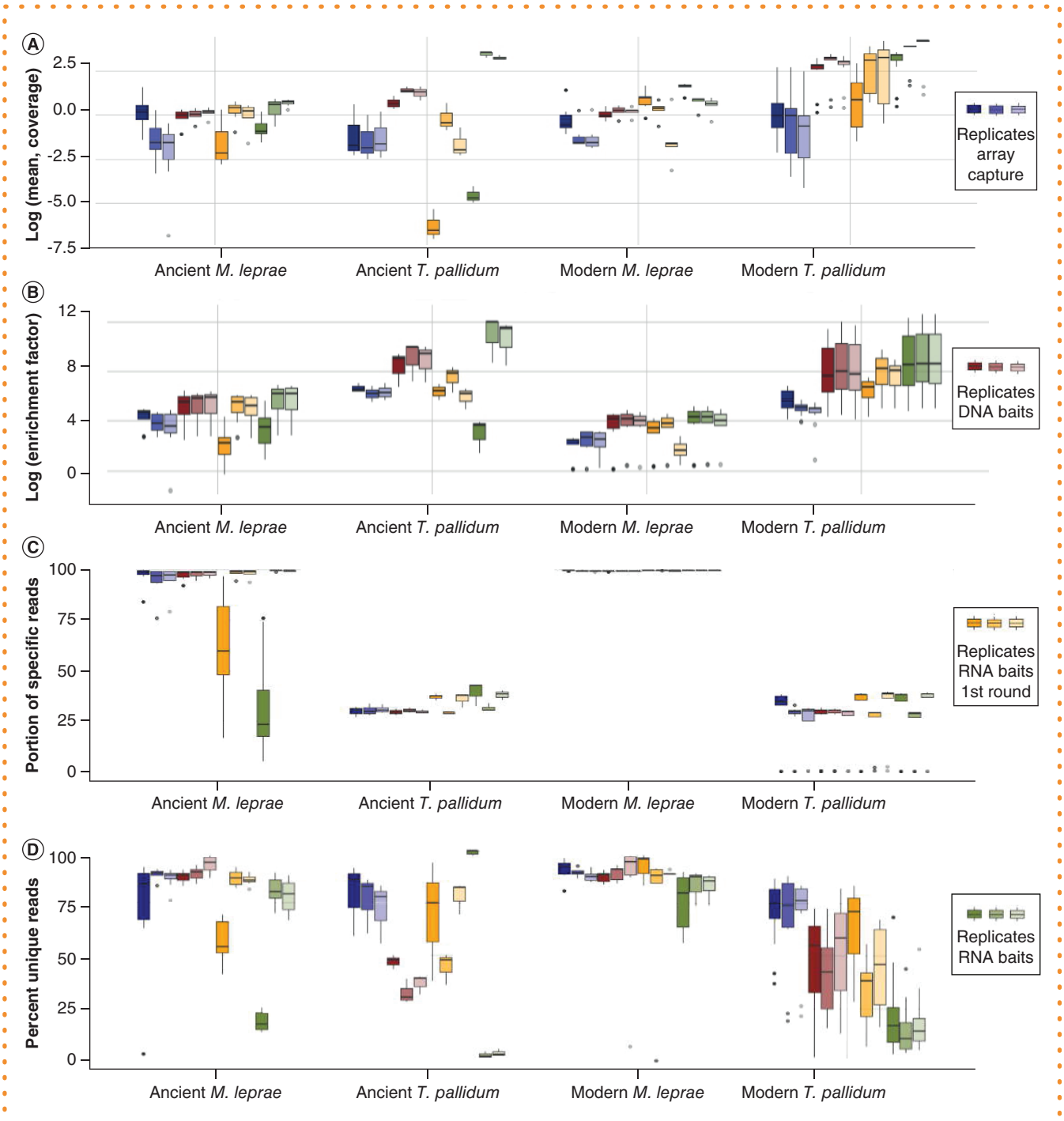

Figure 2. Differences between the three tested protocols in ancient and modern *M. leprae* and *T. pallidum* samples. **(A)** Log-transformed values of the mean coverage. **(B)** log-transformed values of the enrichment factor calculated by dividing the percentage of endogenous DNA by the percentage of endogenous DNA after shotgun sequencing. **(C)** The proportion of specific reads corresponding to *M. leprae* and *T. pallidum* compared to other mycobacterial and treponemal reads, respectively. D) Percentage of unique reads calculated by the number of unique reads divided by the total number of sequences mapped to represent library complexity in *M. leprae* and *T. pallidum* samples.

In-solution capture with DNA baits was used with robot assistance in this study, whereas the in-solution capture with RNA baits was performed in two different labs. Unsurprisingly, the DNA bait capture showed the smallest differences (2- to 50-fold lower) between the replicates and the RNA bait capture showed the largest. These large differences might be due to the use of different versions of

the kit, which was updated by the manufacturer during the course of the project. The DNA array capture showed intermediate results. Consistent conditions are therefore crucial for reproducibility.

Another important feature of targeted enrichment is specificity. We estimated the specificity of the three tested methods by comparing the number of reads specific to either *M. leprae* or *T. pallidum* in comparison with general mycobacterial or treponemal reads, respectively (Figure 2C); differences between the two pathogens could be observed. In the ancient and modern *T. pallidum* samples, the RNA bait capture consistently showed the highest proportion (up to 1.5-times higher) of specific reads. The same trend was observed for the libraries prepared using recent samples from leprosy patients (i.e., modern samples of *M. leprae*). However, the DNA bait capture was more specific for ancient *M. leprae* samples. The highest percentages of specific reads were not necessarily found in samples with high percentages of endogenous DNA in the shotgun data before enrichment.

For ancient and modern samples, in-solution approaches are highly recommended due to their high efficiency, reproducibility and specificity.

Two rounds of hybridization are routinely performed in aDNA research; this is expected to improve enrichment but may also reduce library complexity in terms of proportions of unique reads. To formally investigate the effect of the second round of capture, we also sequenced libraries that were only enriched with one round of hybridization with the RNA baits and compared the results with those of the second round of hybridization. The second round of hybridization resulted in an increase in the enrichment factor for ancient and modern *M. leprae* samples (with an average of 2× increase) as well as for *T. pallidum* samples (with an average of 17× increase), demonstrating the utility of a second round of hybridization capture (Supplementary Table 5). On the other hand, when comparing the library complexity (Figure 2D, Supplementary Figure 3 & Supplementary Notes 2 & 3), we found a substantial loss of complexity after the second round of hybridization in all modern and ancient samples. This loss was reflected in the higher percentage of unique reads in all the reads mapped after the first round. Therefore if the portion of endogenous DNA in the initial sample is high, it may be worth considering whether a single round of capture combined with deeper sequencing is sufficient or even advantageous.

The three protocols also differ in terms of cost and effort. The most cost-intensive is the array capture approach (~€673 per sample), which requires additional equipment that is not usually necessary with the other approaches. By contrast, the in-solution capture with DNA baits is cheaper (~€56 per sample) once the baits are cleaved from the glass slide, but the version that can be used for immortalization of the baits by transforming them into a library is not freely available. At ~€109 per sample, the cost of in-solution capture with RNA baits is more comparable to the DNA bait capture than to the array; this method also needs the least amount of additional equipment and reagents (Supplementary Table 7).

After a detailed comparison of the three tested methods, it can be concluded that for ancient and modern pathogen samples, RNA bait capture with two rounds of hybridization seems to be the most suitable method. The generally high performance of the in-solution approach (especially the one with RNA baits) for both bacterial species suggests that the findings are highly representative and comparable performance is also expected for a variety of other bacterial/microbial organisms.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.future-science.com/doi/suppl/10.2144/btn-2020-0100

Author contributions

VJ Schuenemann, A Herbig and J Krause conceived of the study. B Krause-Kyora and S Calvignac-Spencer provided RNA baits and sequencing libraries. N Arora, P Singh, ST Cole and S Knauf provided sequencing libraries. A Furtwängler, L Böhme, E Reiter and M Vollstedt performed the laboratory work. A Furtwängler and J Neukamm performed the data analysis. A Furtwängler and A Herbig conducted the statistical analysis. A Furtwängler designed the figures. A Furtwängler, VJ Schuenemann and A Herbig wrote the manuscript with input from all authors. All authors reviewed the manuscript.

Acknowledgments

The authors thank all our colleagues providing samples for our study: S Inskip (University of Cambridge, UK); H Donoghue (University College, London, UK); R Barquera (Max Planck Institute for the Science of Human History, Germany); M Taylor, T Mendum and G Stewart (University of Surrey, UK); S Roffey and P Marter (The Magdalen Hill Archaeological Research Project, Winchester, UK); K Tucker (Deutsches Archäologisches Institut, Berlin, Germany); F Leendertz (Robert Koch Insitute, Berlin, Germany); R Wittig (Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany); A Kjellström and C Economou (University of Stockholm, Sweden); P Velemínský (National Museum, Czech Republic); A Marcsik, E Molnár and G Pálfi (University of Szeged, Hungary); V Mariotti and MG Belcastro (University of Bologna, Italy; Aix-Marseille Université, France); A Riga (University of Florence, Italy); JL Boldsen (University of Southern Denmark, Denmark); and C Avanzi (Colorado State University, USA). The authors would also like to thank the laboratory team of the Max Planck Institute for the Science of Human History in Jena for extensive support with capture experiments and sequencing and the developers of nf-core/eager, especially A Peltzer and J Fellows Yates.

Financial & competing interests disclosure

This work was supported by the University of Zurich's University Research Priority Program 'Evolution in Action: From Genomes to Ecosystems' (VJ Schuenemann), the Max Planck Society (J Krause and A Herbig), the Senckenberg Centre for Human Evolution and Palaeoenvironment at the University of Tübingen (VJ Schuenemann, J Krause and A Furtwängler). P Singh's research work is supported by the Indian Council of Medical Research, DBT India, R2STOP Canada and the Leprosy Research Initiative Netherlands. The manuscript has been approved by the Publication Screening Committee of ICMR-NIRTH, Jabalpur and assigned the number ICMR-NIRTH/PSC/44/2020. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Ethical conduct of research

For all samples used in this study, only existing sequencing libraries were used and no new material was collected. Statements about ethical approval and research permission can be found in the original publications (Supplementary Table 1). In this study only sequencing data of the two bacteria and no human data were generated.

Data sharing statement

All sequencing data generated will be deposited upon publication on the SRA under the BioProject PRJNA645054.

References

Papers of special note have been highlighted as: • of interest

1. Sawyer S, Krause J, Guschanski K, Savolainen V, Pääbo S. Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PloS ONE* 7(3), e34131 (2012).
• **Presents the special characteristics of ancient DNA.**
2. Allentoft ME, Collins M, Harker D *et al.* The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc. Biol. Sci.* 279(1748), 4724–4733 (2012).
• **Contains detailed explanation on DNA survival.**
3. Briggs AW, Stenzel U, Johnson PLF *et al.* Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl Acad. Sci. USA* 104(37), 14616–14621 (2007).
• **Presents special characteristics of ancient DNA.**
4. Krause J. From genes to genomes: what is new in ancient DNA? *Mitteilungen der Gesellschaft für Urgeschichte* 19, 11–33 (2010).
5. Spyrou MA, Bos KI, Herbig A, Krause J. Ancient pathogen genomics as an emerging tool for infectious disease research. *Nat. Rev. Genet.* 20(6), 323–340 (2019).
6. Fu Q, Meyer M, Gao X *et al.* DNA analysis of an early modern human from Tianyuan Cave, China. *Proc. Natl Acad. Sci. USA* 110(6), 2223–2227 (2013).
• **Presents the first description and usage of DNA bait capture.**
7. Vågene ÅJ, Herbig A, Campana MG *et al. Salmonella enterica* genomes from victims of a major sixteenth-century epidemic in Mexico. *Nat. Ecol. Evol.* 2(3), 520–528 (2018).
• **Comparison of DNA bait capture and array capture.**
8. Burbano HA, Hodges E, Green RE *et al.* Targeted investigation of the Neandertal genome by array-based sequence capture. *Science (NY)* 328(5979), 723–725 (2010).
• **Array capture used on Neanderthal DNA.**
9. Schuenemann VJ, Singh P, Mendum TA *et al.* Genome-wide comparison of medieval and modern *Mycobacterium leprae*. *Science (NY)* 341(6142), 179–183 (2013).
10. Schuenemann VJ, Avanzi C, Krause-Kyora B *et al.* Ancient genomes reveal a high diversity of Mycobacterium leprae in medieval Europe. *PLoS Pathog.* 14(5), e1006997 (2018).
11. Knauf S, Gogarten JF, Schuenemann VJ *et al.* Nonhuman primates across sub-Saharan Africa are infected with the yaws bacterium *Treponema pallidum* subsp. *pertenue*. *Emerging Microbes Infect.* 7(1), 157 (2018).
12. Arora N, Schuenemann VJ, Jäger G *et al.* Origin of modern syphilis and emergence of a pandemic *Treponema pallidum* cluster. *Nat. Microbiol.* 2, 16245 (2016).
13. Schuenemann VJ, Kumar Lankapalli A, Barquera R *et al.* Historic *Treponema pallidum* genomes from colonial Mexico retrieved from archaeological remains. *PLoS Negl. Trop. Dis.* 12(6), e0006447 (2018).