

Target-enrichment sequencing yields valuable genomic data for challenging-to-culture bacteria of public health importance

Tristan P. W. Dennis¹, Barbara K. Mable¹, Brian Brunelle², Alison Devault², Ryan W. Carter¹, Clare L. Ling^{3,4}, Blandina T. Mmbaga⁵, Jo E. B. Halliday¹, Katarina Oravcova^{1†} and Taya L. Forde^{1*,†}

Abstract

Genomic data contribute invaluable information to the epidemiological investigation of pathogens of public health importance. However, whole-genome sequencing (WGS) of bacteria typically relies on culture, which represents a major hurdle for generating such data for a wide range of species for which culture is challenging. In this study, we assessed the use of culture-free target-enrichment sequencing as a method for generating genomic data for two bacterial species: (1) *Bacillus anthracis*, which causes anthrax in both people and animals and whose culture requires high-level containment facilities; and (2) *Mycoplasma amphoriforme*, a fastidious emerging human respiratory pathogen. We obtained high-quality genomic data for both species directly from clinical samples, with sufficient coverage (>15×) for confident variant calling over at least 80% of the baited genomes for over two thirds of the samples tested. Higher qPCR cycle threshold (*Ct*) values (indicative of lower pathogen concentrations in the samples), pooling libraries prior to capture, and lower captured library concentration were all statistically associated with lower capture efficiency. The *Ct* value had the highest predictive value, explaining 52% of the variation in capture efficiency. Samples with *Ct* values ≤30 were over six times more likely to achieve the threshold coverage than those with a *Ct* > 30. We conclude that target-enrichment sequencing provides a valuable alternative to standard WGS following bacterial culture and creates opportunities for an improved understanding of the epidemiology and evolution of many clinically important pathogens for which culture is challenging.

DATA SUMMARY

The authors confirm all supporting data, code and protocols have been provided within the article or through supplementary data files. Scripts used in this study can be accessed on GitHub at <https://github.com/tristanpwdennis/bactocap>. All sequence data generated during this study have been deposited in the European Nucleotide Archive (ENA) Sequence Read Archive (SRA) under project accession numbers PRJEB46822 (*B. anthracis*) and PRJEB50216 (*M. amphoriforme*). Accession numbers for individual samples, along with metadata, laboratory parameters and sequence quality metrics, are available at the University of Glasgow's data repository, Enlighten, at <http://dx.doi.org/10.5525/gla.researchdata.1249>.

INTRODUCTION

Genomic data can provide critical information towards understanding bacterial epidemiology and evolution, and thereby help to guide public health policy and infectious disease control [1]. Whole-genome sequence (WGS) data contribute to determining transmission dynamics at various scales [2], and are used to track the emergence and evolution of particular strains or traits [3].

Received 17 March 2022; Accepted 26 April 2022; Published 27 May 2022

Author affiliations: ¹Institute of Biodiversity, Animal Health & Comparative Medicine, College of Medical Veterinary and Life Sciences, University of Glasgow, Glasgow, UK; ²Daicel Arbor Biosciences, Ann Arbor, Michigan, USA; ³Shoklo Malaria Research Unit, Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Mae Sot, Thailand; ⁴Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK; ⁵Kilimanjaro Clinical Research Institute, Kilimanjaro Christian Medical Centre, Moshi, Tanzania.

***Correspondence:** Taya L. Forde, taya.forde@glasgow.ac.uk

Keywords: target-enrichment sequencing; RNA baits; bacteria; genomic; *Bacillus anthracis*; *Mycoplasma amphoriforme*.

Abbreviations: bp, base pairs; CL, containment level; Ct, cycle threshold; LMICs, low-and-middle-income countries; NPA, nasopharyngeal aspirates; NPS, nasopharyngeal swabs; WGS, whole-genome sequence.

†These authors contributed equally to this work

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Supplementary material is available with the online version of this article.

000836 © 2022 The Authors



This is an open-access article distributed under the terms of the Creative Commons Attribution License.

Impact Statement

Sequence data are increasingly relied upon to help understand the epidemiology, evolution and virulence of pathogens. However, for a variety of reasons, many bacterial species remain challenging to culture, precluding genomic data to be obtained using standard methods. Here, we apply a culture-free method – target-enrichment sequencing – to obtain such data from two bacterial species of public health importance: the bacterium that causes anthrax (*Bacillus anthracis*), and an emerging respiratory pathogen (*Mycoplasma amphoriforme*). This approach enabled high-coverage sequence data to be obtained for both species. The specific enrichment panels designed for these pathogens will facilitate their further molecular epidemiological study. More broadly, this study highlights the value of this methodology for genomic sequencing of bacteria in situations where they cannot be cultured.

WGS is routinely implemented for the characterization of a wide range of bacteria, but typically involves working from cultured isolates to yield DNA of sufficient quantity and quality. This reliance on culture as a first step represents a major hurdle for generating genomic data for many bacteria of public health importance. For instance, for bacteria with high biosafety requirements, suitable laboratories may not be available. Similar obstacles exist for fastidious bacteria (i.e. due to their complex nutritional requirements and slow growth), rendering their culture for genetic analysis difficult or impossible [4]. Moreover, while sequencing can be performed directly on extracts from clinical samples (e.g. bodily fluids and tissues), pathogen DNA is often present only in low quantity, representing only a fraction of the overall DNA in a sample [5], thus complicating WGS-based analyses.

Hybridization capture followed by sequencing, or ‘target-enrichment sequencing’, is one approach that may be able to overcome this obstacle and enable bacterial genomic sequence data to be obtained more effectively from clinical samples [6, 7]. In this culture-free method, libraries are prepared from DNA extracted from the sample, after which specific fragments are selectively enriched prior to sequencing. This is accomplished using custom-designed probes, or ‘baits’ – typically RNA oligonucleotides labelled with biotin that hybridize to the DNA of interest and are then captured by streptavidin-coated magnetic beads, allowing non-target DNA to be removed through wash steps [8]. Baits are able to selectively capture a wide range of fragment sizes [8, 9] and also allow for polymorphisms between their sequence and that of the enriched fragment; depending on the stringency of capture, this can be as low as 70% sequence similarity [8]. When compared to sequencing directly from samples, target-enrichment sequencing has been shown to substantially increase the proportion of on-target reads and sequencing depth [4]. For several bacterial species, this has enabled the reconstruction of whole-genome sequences from relatively small amounts of target DNA from clinical samples, including sputum, vaginal swabs and cerebrospinal fluid, as well as from arthropod vectors [4, 10–12]. This approach thus has the potential to facilitate in-depth investigations, including phylogenomics, into otherwise difficult-to-study pathogens.

In this study, we evaluated target-enrichment sequencing as a means of generating genomic data directly from samples containing one of two different difficult-to-culture bacteria: *Bacillus anthracis* and *Mycoplasma amphoriforme*. *Bacillus anthracis* is a Gram-positive bacterium that causes anthrax – a neglected bacterial zoonosis that is a threat to livestock and human health in low-and-middle-income countries (LMICs) [13]. The high virulence of *B. anthracis* makes it hazardous to culture and therefore requires containment level (CL) 2+ or 3 facilities [14]. Limited high-containment laboratory capacity in LMICs thus represents a barrier to the generation of sequence data for *B. anthracis* and understanding of its diversity and transmission in endemic settings. Moreover, concerns over the use of *B. anthracis* as a biological weapon make the shipment of samples for culture elsewhere challenging [15]. Although the true clinical significance of *M. amphoriforme* is yet to be elucidated, it has been associated with respiratory tract infections in immunocompetent and immunocompromised patients [16–21]. While there is great potential to use comparative genomics to gain insights into its pathogenicity and epidemiology, *M. amphoriforme* requires specialized culture media and conditions; several weeks are required to generate bacterial cultures with suitable cell densities for DNA extraction and not all PCR positive samples yield a positive culture despite high estimated bacterial loads [16, 18].

The primary objective of this study was to assess the feasibility of generating sequence data for *B. anthracis* and *M. amphoriforme* at coverage suitable for variant calling over the majority of the genome. We also sought to determine the upper pathogen-specific qPCR cycle threshold (*Ct*) value limit to provide sufficient coverage for accurate variant calling, and the influence of different variables (i.e. *Ct*, whether libraries were pooled prior to bait capture, and concentration of the captured libraries) on capture efficiency (i.e. the proportion of on-target reads based on reference mapping).

METHODS

Custom myBaits sets were designed for *B. anthracis* and *M. amphoriforme* in partnership with Daicel Arbor Biosciences (Ann Arbor, MI, USA). RNA baits of ~80 base pairs (bp) were designed to achieve 50% overlap (tiling) across targeted regions (i.e. at least two baits covering any given position). For *B. anthracis*, baits were designed to tile the core chromosomal genome. Publicly

available *B. anthracis* genomes available at the time of bait design ($n=52$; Table S1, available in the online version of this article) – including genomes from all three major genetic subgroups (clades A, B and C) [22] – were used to generate a core-genome alignment in Parsnp [23]. This alignment was 4.66 MB (89% of the 5.23 MB Ames Ancestor reference chromosomal genome, NC_007530.2 [24]). Plasmid (PXO1 and PXO2) sequences were not included in bait design because the primary goal was to use the data for phylogenetic inference, for which accessory genome content is unsuited. For *M. amphoriforme*, bait sequences were designed to cover the pangenome, which was constructed using all *M. amphoriforme* whole-genome sequences available at the time of the experimental design, and included the A39 reference genome NCTC 11740 (accession number HG937516; 1029022 bp) and 18 additional *M. amphoriforme* genomes [19]. Pairwise comparison of these genomes was then done using Artemis Comparison Tool [25], concatenating any regions of difference to the reference genome of strain A39, resulting in a pangenome of 1064467 bp. For both bacterial species, baits that targeted loci with high copy number in the genome were omitted (e.g. rRNA, tRNA), as were simple repeats, which were soft-masked using RepeatMasker [26]. Additionally, *in silico* analysis was performed to identify and remove any baits that could potentially cross-hybridize with host or non-target bacteria (File S1, Table S2). The final bait set for *B. anthracis* comprised 148811 baits and covered 4637856 bp (88%) of the 5227419 Ames Ancestor chromosomal reference sequence. For *M. amphoriforme*, the final bait set comprised 24 444 baits. This set covered 1002426 bp (97%) of the 1029022 bp A39 reference genome, or 94% of the 1064467 bp pangenome. Final bait set sequences for both case studies (as .bed annotations) are provided on GitHub [27], and are also available through the Daicel Arbor Biosciences Community Panels webpage.

For *B. anthracis*, DNA extracts ($n=93$) from animal carcasses sampled in northern Tanzania between 2016 and 2018 were analysed. All sample extracts had tested positive by qPCR for three regions of the *B. anthracis* genome (*cap*, *lef* and *PL3* [28]), with one exception where *cap* failed to amplify. Taking the highest of the three *Ct* values per sample, the *Ct* values ranged from 19 to 37 (median=25). For *M. amphoriforme*, all DNA extractions from nasopharyngeal swabs (NPS, $n=52$) and nasopharyngeal aspirates (NPA, $n=4$) that had tested positive with a *M. amphoriforme*-specific qPCR targeting the *udg* gene [18] during two surveillance studies were included. The NPS originated from a longitudinal *Streptococcus pneumoniae* carriage study on the Thailand–Myanmar border [29], and were collected from 11 participants, each providing between three and six consecutive swabs between 2007–2010. The NPA originated from influenza virus PCR surveillance also on the Thailand–Myanmar border in 2016 and 2017. The *Ct* values for *M. amphoriforme* samples ranged from 20 to 40 (median=29). Sample-specific metadata for both bacterial species are available as Supplementary Tables A and B at the University of Glasgow’s Enlighten data repository [30].

All DNA extracts were quantified using Qubit dsDNA high sensitivity Assay Kits (Thermo Fisher Scientific). All initial libraries (referred to as ‘pre-capture libraries’) were prepared using the NEBNext Ultra II FS DNA Library Prep Kit for Illumina with NEBNext Sample Purification Beads (New England BioLabs). This protocol includes a short PCR amplification (five–seven cycles) to add sample-specific barcodes and to increase the total DNA concentration prior to bait capture. We varied the number of cycles across library preparation batches to achieve the recommended minimum DNA input for bait capture (100 ng), whilst aiming to use the minimum number of cycles required (details in File S1).

For bait capture, we followed the myBaits Hybridization Capture for Targeted NGS protocol (Daicel Arbor Biosciences, Manual v4.01). For each bait capture, 7 μ l of pre-capture library was used as input. Most libraries were captured individually. For a subset, libraries were pooled before bait hybridization with the goal of reducing the cost of baits per sample; this was done for 34 *B. anthracis* libraries that spanned a range of *Ct* values, and eight *M. amphoriforme* libraries prepared from samples with higher *Ct* values. As per the myBaits protocol, we aimed to include four libraries per pool; however, in some cases where certain initial library preparations were unsuccessful, only three pre-capture libraries were pooled. For each pool, approximately equal input concentrations of each pre-capture library were included to yield a total input of 125 ng in 7 μ l.

Following bait hybridization, the number of cycles in the PCR amplification step – included in the myBaits protocol to yield at least 1.5 ng μ l⁻¹ DNA for downstream sequencing – was varied among preparation batches; between 12 and 15 cycles were used for *B. anthracis* and 12 or 13 cycles for *M. amphoriforme* samples. Post-amplification clean-up was done by column-based purification using Zymo DNA Clean and Concentrator –5 kit (Cambridge Bioscience, UK), eluting in 12 μ l of 0.1 \times TE buffer. We refer to the final product as ‘captured libraries’.

Captured libraries were quality checked using a Bioanalyzer (Agilent) and sequenced at Glasgow Polyomics on an Illumina NextSeq instrument, generating 75 bp paired-end reads. Given differences in genome size, approximately 4 and 1 million reads were targeted for *B. anthracis* and *M. amphoriforme*, respectively. Raw reads were trimmed with fastp 0.20.1 with default settings [31] and aligned to the *B. anthracis* Ames Ancestor reference genome or the *M. amphoriforme* pangenome, respectively, with bwa-mem v0.7.17-r1188 [32], after which PCR duplicates were removed in GATK [33]. We defined a ‘successful’ sequencing outcome as a sample achieving >15 \times coverage (the cut-off for confidently calling a genotype in a haploid organism [34]) over 80% of the baited genome (i.e. the parts of the core *B. anthracis* genome or *M. amphoriforme* pangenome against which baits were designed). Full analysis scripts are available on Github [27].

A binomial generalized linear mixed model (GLMM) was performed in R [35] with *glmer* in the *lme4* library [36] to investigate the association between the capture efficiency – defined as the proportion of mapped reads per sample – and the explanatory variables: *Ct* values from pathogen-specific qPCR performed on original sample DNA extracts; bacterial species (i.e. *B. anthracis*

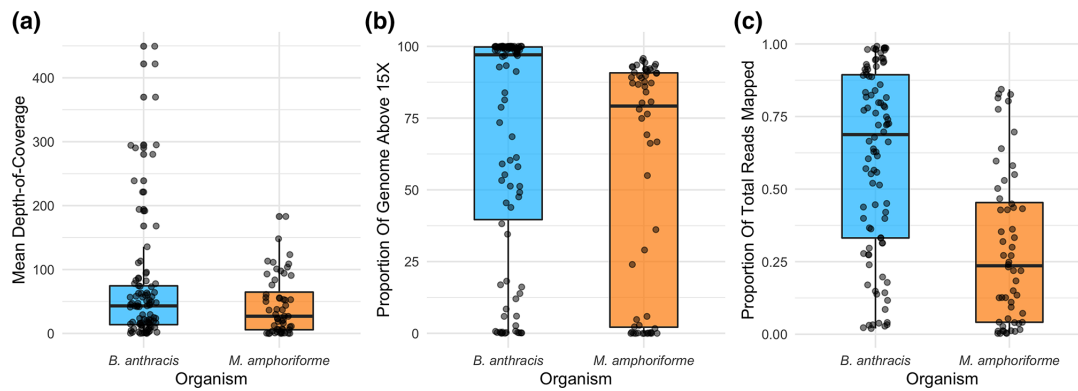


Fig. 1. Target-enrichment sequencing outcomes. (a) Mean depth-of-coverage over baited regions of the reference genomes, per-sample for *B. anthracis* ($n = 93$; blue) and *M. amphoriforme* ($n = 56$; orange). Both pooled and unpooled samples are included. (b) Proportion of baited genome covered to $>15\times$ per-sample; and (c) proportion of reads mapped to baited regions. Points indicate individual samples, box and whisker plots indicate maximum, minimum, upper-lower quartiles and median values.

or *M. amphoriforme*); whether or not libraries were pooled; and captured library concentration ($\text{ng } \mu\text{l}^{-1}$). All models included an observation level random effect to account for overdispersion [37]. Pairwise interactions between bacterial species and the other variables were also considered. Model simplification was evaluated based on comparison of AIC, pseudo-R2 analysis was performed in the *MuMIn* library [38], and the final model was plotted in *sjPlot* [39].

RESULTS

For *B. anthracis*, the median number of total reads generated per sample was 6413796 (range: 539623–32708157). The mean depth of coverage per sample across the baited regions of the Ames Ancestor reference genome ranged from $0.2\times$ to $449.4\times$ (median= $43.2\times$) (Fig. 1a). Based on our previously defined threshold of $>15\times$ coverage over $>80\%$ of the baited genome, 55 of 93 *B. anthracis* samples (59.1%) were considered successfully sequenced: 40/59 (68%) of unpooled and 15/34 (44%) pooled samples. The proportion of the baited genome covered to $15\times$ depth ranged from 0–100% per sample (median= 97.3% ; Fig. 1b). The proportion of mapped reads per sample ranged from 2.3–99.2% (median= 69.7%) (Fig. 1c). This latter metric is independent of the per-sample sequencing depth.

For the *M. amphoriforme*-positive samples, the total number of reads generated per sample ranged from 1082454 to 4247924 (median= 2989985). The mean depth of coverage per sample across the baited regions of the *M. amphoriforme* pangenome ranged from $0.1\times$ to $183.1\times$ (median= $27.0\times$) (Fig. 1a). Twenty-eight of 56 samples (50%) achieved $>15\times$ coverage over 80% of the baited *M. amphoriforme* pangenome: 28/48 (58%) of unpooled and 0/8 (0%) pooled. The proportion of the pangenome covered to this depth ranged from 0–95.8% (median= 79.2% ; Fig. 1b). Mapped reads represented between 0.3 and 84.4% of the total reads per sample (median= 23.6%) (Fig. 1c). Summary statistics of the sequencing metrics for both datasets are provided in Table S3. Sample-specific sequencing results and accession numbers for both bacterial species are available in the Supplementary Tables on Enlighten [30].

The final model to evaluate variables influencing capture efficiency included the C_t value, pooling libraries before capture (yes/no), and captured library concentration variables, with an observation level random effect (Table S4, File S1). The model indicated that lower C_t is significantly associated with increased capture efficiency (Fig. 2). Captured library concentration was significantly positively associated with capture efficiency (Fig. S1), while pooling libraries prior to bait capture was negatively associated with capture efficiency (Fig. S2). The model accounted for 55% of the variation in capture efficiency. We found that C_t explained 52% of the variation in our data. In contrast, captured library concentration explained 13%, and pooling only 1%.

Of ten *B. anthracis* samples with $C_t >30$, only one of these was successfully sequenced (with a C_t of 31) vs. 65% of samples (54/83) with $C_t \leq 30$. Similarly for *M. amphoriforme*, 25 of 33 samples with $C_t \leq 30$ (76%) were successful, compared with 3 of 23 with $C_t >30$ (13%). The relationship between C_t and mean depth of coverage is shown in Fig. S3.

DISCUSSION

We have shown that broad, deep sequencing coverage can be obtained for *M. amphoriforme* and *B. anthracis* directly from samples using target-enrichment sequencing. These species are representative of the breadth of typical bacterial genome

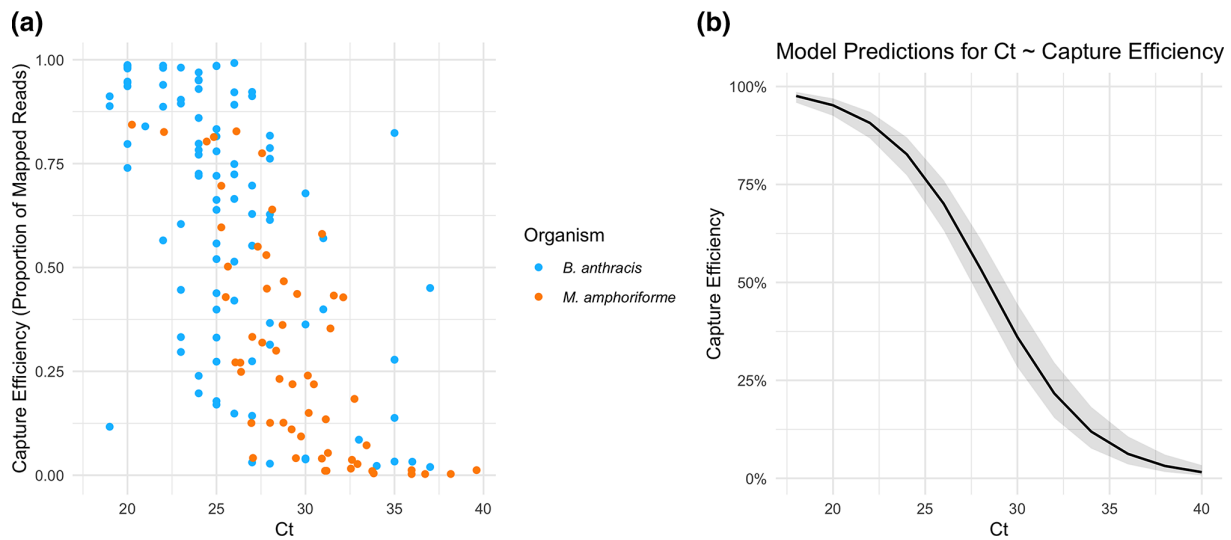


Fig. 2. Relationship between Ct value and capture efficiency. (a) Each point represents an individual sample, with bacterial species distinguished by colour. (b) Binomial GLMM prediction of the relationship between Ct value and capture efficiency for both species. Shading represents a 95% prediction interval.

sizes (~1.0 MB and ~5.2 MB, respectively). For both species, we were able to recover a substantial portion of the genome at a sufficient depth of coverage for haploid variant calling where otherwise such data would have been extremely challenging or impossible to obtain. Since total sequencing read amounts are independent of enrichment performance, it is possible that even more of the study samples could have reached our coverage threshold ($>15\times$ across 80% of the baited genome) had we performed additional sequencing to generate more reads per sample. Target-enrichment sequencing will be transformative for the molecular epidemiological study of these two species of public health importance, and more generally provides a valuable alternative when standard WGS following bacterial culture is impracticable [11, 12, 40, 41], so long as a reference sequence is available to facilitate bait design. Culture-free methods such as targeted-sequence capture expand the opportunities for the genomic study of several Hazard Group 2+/3 pathogens, such as *B. anthracis*, within LMIC countries where they remain a threat to human and animal health. Whereas high-containment facilities are often lacking, sequencing platforms are increasingly available worldwide. Targeted-sequence capture also promises to facilitate our understanding of the genomic epidemiology of *M. amorphiforme*, which has only been isolated from patients on a few occasions due to its requirement for specific culture conditions and slow growth [19]. It has particular promise for studying pathogenic and antimicrobial resistance determinants, and thus aid timely clinical management of respiratory tract infections caused by this bacterium [17, 19, 21]. The developed bait sets and methods described in this study will provide a valuable resource for other researchers wishing to pursue the genomic study of these bacteria.

Our results suggest that target-enrichment sequencing is most effective for samples with higher pathogen-specific DNA concentrations, as indicated by Ct value. Ct was strongly correlated with capture efficiency (i.e. a higher proportion of mapped reads, indicative of less off-target capture), as well as the success of sequencing to a sufficient depth for confident genotyping, a finding consistent with previous studies [11, 42]. At least two thirds of samples with Ct values ≤ 30 – indicative of higher starting pathogen concentrations in the samples – were successfully sequenced for both sample sets, whereas this success rate dropped to 10–13% for samples with $Ct > 30$. Based on previous studies implementing the *udg* gene as a PCR target for *M. amorphiforme* [18], a Ct of 30 would correspond to approximately 90–900 genome copies per PCR reaction, suggesting that target-enrichment sequencing will be most useful for pathogens that are present in moderate quantities in clinical samples. Generating genomic data for clinical samples with very low pathogen concentrations thus remains a challenge. While not assessed in this study, including a second round of bait capture (i.e. putting the enriched library through a second round of enrichment) has been suggested to improve sensitivity for such samples [43]. Although capture efficiency was also significantly associated with higher captured library concentration and with libraries not having been pooled prior to capture, the predictive value of these variables was low; Ct value alone could explain most of the variation in the data.

We found that pooling up to four libraries for capture still yielded the targeted coverage in nearly half of the pooled *B. anthracis* samples, thereby representing potential for even greater cost savings compared to sequencing without enrichment, by significantly reducing the per-sample bait capture reagent costs. The *M. amorphiforme* samples with libraries pooled in this study had a higher median Ct than unpooled samples (File S1), which likely accounts for the greater difference in capture

efficiency observed between pooled vs. unpooled libraries for the two bacterial species (Fig. S2). Previous studies have been more consistent in achieving successful coverage when pooling as many as 20 libraries for *Borellia burgdorferi* [10] and *Treponema pallidum* [40]; this success may be attributed in part to carefully selecting libraries for pooling within a narrow range of similar Ct values as performed by Beale *et al.* [40] and as recommended in later versions of the myBaits manual (v5); therefore higher levels of sample pooling for capture may be feasible for *B. anthracis* and *M. amphoriforme* as well.

The ability to generate culture-free genomic data opens significant opportunities for wider phylogenetic studies of these two bacterial species of public health concern, and highlights the potential value of target-enrichment sequencing for facilitating an improved understanding of the epidemiology and evolution of many clinically important pathogens for which genomic data have been limited.

Funding information

This work was supported by the Medical Research Council (MC_PC_16045). T.P.W.D. was supported by the European Research Council under the European Union's Horizon 2020 Research and Innovation Programme (grant agreement no. 852957). T.L.F. was supported by a Biotechnology and Biological Sciences Research Council Discovery Fellowship (BB/R012075/1) and by an Academy of Medical Sciences Springboard award, with contribution from the Wellcome Trust and Global Challenges Research Fund (SBF005\1023). The Shoklo Malaria Research Unit is part of the Wellcome Trust Mahidol University Oxford Tropical Medicine Research Unit, which is funded by the Wellcome Trust 220211. The funders had no role in study design, data collection and analysis, or preparation of the manuscript.

Acknowledgements

We are grateful to the members of the field team who collected the *B. anthracis* samples: Deogratus Mshanga, Sabore Ole Moko, Sironga Nanjicho and Kadogo Lerimba. We also thank Nichith Kollanandi Ratheesh and the staff at Kilimanjaro Clinical Research Institute for assistance in sample processing, particularly Alutu Masokoto. Thank you to Shoklo Malaria Research Unit, Thailand for the provision of *M. amphoriforme* positive samples, in particular Professor Paul Turner and Professor François Nosten, and to Wiphada Srinark for DNA extraction and PCR testing. We would like to thank Dr Mat Beale and Professor Nick Thomson from Wellcome Sanger Institute for their invaluable input for *M. amphoriforme* data analysis and constructive suggestions for the manuscript. We also thank Dr Paul Johnson and Professor Dan Haydon for their input on the statistical analyses.

Author contributions

Conceptualization: B.K.M., J.E.B.H., K.O., T.L.F. Data curation: T.P.W.D., K.O., T.L.F. Formal analysis: T.P.W.D., K.O., T.L.F. Funding acquisition: J.E.B.H., B.K.M., K.O., T.L.F., B.B., A.D. Investigation: R.C., K.O., T.L.F. Methodology: B.B., A.D., B.K.M., J.E.B.H., T.P.W.D., K.O., T.L.F. Project administration: J.E.B.H., B.K.M., K.O., T.L.F. Resources: B.T.M., C.L.L. Software: T.P.W.D. Visualization: T.P.W.D. Writing – original draft: T.P.W.D., K.O., T.L.F. Writing – review and editing: T.P.W.D., B.K.M., B.B., A.D., C.L.L., J.E.B.H., K.O., T.L.F.

Conflicts of interest

A.D. and B.B. are employed by Daicel Arbor Biosciences, which sells the myBaits Custom hybridization capture kits utilized in this study. The other authors declare that there are no conflicts of interest.

Ethical statement

The *B. anthracis* study received ethical approval from the Kilimanjaro Christian Medical University College Ethics Review Committee (certificate No. 2050); the National Institute for Medical Research, Tanzania (NIMR/HQ/R.8a/Vol. IX/2660); Tanzanian Commission for Science and Technology (2016-95-NA-2016-45); and the College of Medical Veterinary and Life Sciences ethics committee at the University of Glasgow (200150152). Material and data transfer agreements were established as part of these approvals, and this study complied with Tanzania's national access measures for genetic material. Ethical approval related to the *M. amphoriforme* samples was granted in the original study by ethics committees of the Faculty of Tropical Medicine, Mahidol University, Thailand (MUTM-2009-306) and Oxford University, UK (OXTREC-031-06) [23].

References

- Brand A, Brand H, Schulte in den Bäumen T. The impact of genetics and genomics on public health. *Eur J Hum Genet* 2008;16:5–13.
- Kao RR, Haydon DT, Lycett SJ, Murcia PR. Supersize me: how whole-genome sequencing and big data are transforming epidemiology. *Trends Microbiol* 2014;22:282–291.
- Sajib MSI, Tanmoy AM, Hooda Y, Rahman H, Andrews JR, *et al.* Tracking the emergence of azithromycin resistance in multiple genotypes of typhoidal *Salmonella*. *mBio* 2021;12:e03481-20.
- Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, *et al.* Rapid whole-genome sequencing of *Mycobacterium tuberculosis* isolates directly from clinical samples. *J Clin Microbiol* 2015;53:2230–2237.
- Sinha M, Jupe J, Mack H, Coleman TP, Lawrence SM, *et al.* Emerging technologies for molecular diagnosis of sepsis. *Clin Microbiol Rev* 2018;31:e00089-17.
- Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, *et al.* Target-enrichment strategies for next-generation sequencing. *Nat Methods* 2010;7:111–118.
- Gaudin M, Desnues C. Hybrid capture-based next generation sequencing and its application to human infectious diseases. *Front Microbiol* 2018;9:2924.
- Gnrirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* 2009;27:182–189.
- Hodges E, Rooks M, Xuan Z, Bhattacharjee A, Benjamin Gordon D, *et al.* Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat Protoc* 2009;4:960–974.
- Carpi G, Walter KS, Bent SJ, Hoen AG, Diuk-Wasser M, *et al.* Whole genome capture of vector-borne pathogens from mixed DNA samples: a case study of *Borrelia burgdorferi*. *BMC Genomics* 2015;16:434.
- Christiansen MT, Brown AC, Kundu S, Tutill HJ, Williams R, *et al.* Whole-genome enrichment and sequencing of *Chlamydia trachomatis* directly from clinical samples. *BMC Infect Dis* 2014;14:591.
- Clark SA, Doyle R, Lucidarme J, Borrow N, Breuer J. Targeted DNA enrichment and whole genome sequencing of *Neisseria meningitidis* directly from clinical specimens. *Int J Med Microbiol* 2018;308:256–262.
- Carlson CJ, Kralacik IT, Ross N, Alexander KA, Hugh-Jones ME, *et al.* The global distribution of *Bacillus anthracis* and associated anthrax risk to humans, livestock and wildlife. *Nat Microbiol* 2019;4:1337–1343.
- WHO. *Anthrax in Humans and Animals*. 2008.

15. D'Amelio E, Gentile B, Lista F, D'Amelio R. Historical evolution of human anthrax from occupational disease to potentially global threat as bioweapon. *Environ Int* 2015;85:133–146.
16. Pitcher DG, Windsor D, Windsor H, Bradbury JM, Yavari C, et al. *Mycoplasma amphoriforme* sp. nov., isolated from a patient with chronic bronchopneumonia. *Int J Syst Evol Microbiol* 2005;55:2589–2594.
17. Pereyre S, Renaudin H, Touati A, Charron A, Peuchant O, et al. Detection and susceptibility testing of *Mycoplasma amphoriforme* isolates from patients with respiratory tract infections. *Clin Microbiol Infect* 2010;16:1007–1009.
18. Ling CL, Oravcova K, Beattie TF, Creer DD, Dilworth P, et al. Tools for detection of *Mycoplasma amphoriforme*: a primary respiratory pathogen? *J Clin Microbiol* 2014;52:1177–1181.
19. Gillespie SH, Ling CL, Oravcova K, Pinheiro M, Wells L, et al. Genomic Investigations unmask *Mycoplasma amphoriforme*, a new respiratory pathogen. *Clin Infect Dis* 2015;60:381–388.
20. van Schaik ML, Patberg KW, Wallinga JG, Wolfhagen MJHM, Bruijnesteijn van Coppenraet LES. *Mycoplasma amphoriforme* vs *M. pneumoniae*: similarities and differences between patient characteristics in a regional hospital in the Netherlands. *J Med Microbiol* 2018;67:1348–1350.
21. Rehman SU, Day J, Afshar B, Rowlands RS, Billam H, et al. Molecular exploration for *Mycoplasma amphoriforme*, *Mycoplasma fermentans* and *Ureaplasma* spp. in patient samples previously investigated for *Mycoplasma pneumoniae* infection. *Clin Microbiol Infect* 2021;27:1697.
22. Van Ert MN, Easterday WR, Huynh LY, Okinaka RT, Hugh-Jones ME, et al. Global genetic population structure of *Bacillus anthracis*. *PLoS ONE* 2007;2:e461.
23. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol* 2014;15:524.
24. Ravel J, Jiang L, Stanley ST, Wilson MR, Decker RS, et al. The complete genome sequence of *Bacillus anthracis* Ames "Ancestor." *J Bacteriol* 2009;191:445–446.
25. Carver TJ, Rutherford KM, Berriman M, Rajandream M-A, Barrell BG, et al. ACT: the artemis comparison tool. *Bioinformatics* 2005;21:3422–3423.
26. Smit A, Hubley R, Green P. RepeatMasker Open-4.0; 2013. <http://www.repeatmasker.org>
27. Dennis TPW. Bactocap' Github repo; 2022. <https://github.com/tristanpwdennis/bactocap>
28. Aminu OR, Lembo T, Zadoks RN, Biek R, Lewis S, et al. Practical and effective diagnosis of animal anthrax in endemic low-resource settings. *PLoS Negl Trop Dis* 2020;14:e0008655.
29. Turner P, Turner C, Jankhot A, Helen N, Lee SJ, et al. A longitudinal study of *Streptococcus pneumoniae* carriage in A cohort of infants and their mothers on the Thailand-Myanmar border. *PLoS One* 2012;7:e38271.
30. Dennis TPW, Mable B, Brunelle B, Devault A, Carter R, et al. Supplementary Tables A & B; 2022. <http://dx.doi.org/10.5525/gla.researchdata.1249>
31. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;34:i884–i890.
32. Li H. *Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM*. Oxford University Press; arXiv:1303.3997,
33. Van der Auwera GA, O'Connor BD. *Genomics in the Cloud*. O'Reilly Media, Inc, 2020.
34. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008;456:53–59.
35. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021. <https://www.R-project.org/>
36. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw* 2015;67:1–48.
37. Harrison XA. Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ* 2014;2:e616.
38. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference*. 2nd ed. New York: Springer-Verlag; 2002.
39. Lüdtke D. sjPlot: Data Visualization for Statistics in Social Science; (n.d.). <https://CRAN.R-project.org/package=sjPlot>
40. Beale MA, Marks M, Sahi SK, Tantaló LC, Nori AV, et al. Genomic epidemiology of syphilis reveals independent emergence of macrolide resistance across multiple circulating lineages. *Nat Commun* 2019;10:3255.
41. Hadfield J, Harris SR, Seth-Smith HMB, Parmar S, Andersson P, et al. Comprehensive global genome dynamics of *Chlamydia trachomatis* show ancient diversification followed by contemporary mixing and recent lineage expansion. *Genome Res* 2017;27:1220–1229.
42. Enk J, Rouillard J-M, Poinar H. Quantitative PCR as a predictor of aligned ancient DNA read counts following targeted enrichment. *Biotechniques* 2013;55:300–309.
43. Furtwängler A, Neukamm J, Böhme L, Reiter E, Vollstedt M, et al. Comparison of target enrichment strategies for ancient pathogen DNA. *Biotechniques* 2020;69:455–459.

Five reasons to publish your next article with a Microbiology Society journal

1. The Microbiology Society is a not-for-profit organization.
2. We offer fast and rigorous peer review – average time to first decision is 4–6 weeks.
3. Our journals have a global readership with subscriptions held in research institutions around the world.
4. 80% of our authors rate our submission process as 'excellent' or 'very good'.
5. Your article will be published on an interactive journal platform with advanced metrics.

Find out more and submit your article at microbiologyresearch.org.