1 **Comparison of library preparation and sequencing depths for direct sequencing of** 2 ***Bordetella pertussis* positive samples**

3

4 Winkie Fong^{1,3,4}, Keenan Pey^{1,3,4}, Rebecca Rockett^{1,3}, Rosemarie Sadsad^{1,2,3,4}, Vitali
5 Sintchenko^{1,3,4}, and Verlaine Timms^{1,3}

6

7 ¹Centre for Infectious Diseases and Microbiology – Public Health, Westmead Hospital,
8 Westmead, NSW, Australia

9 ²Centre for Infectious Diseases and Microbiology Laboratory Services, Westmead Hospital,
10 Westmead, NSW, Australia

11 ³Westmead Clinical School, The University of Sydney, Westmead, NSW, Australia

12 ⁴Marie Bashir Institute for Infectious Diseases and Biosecurity, The University of Sydney,
13 Camperdown, NSW, Australia

14

15 Corresponding author: Winkie Fong
16 Email: wfon4473@sydney.edu.au

17

18 **Abstract**

19 Whooping cough, or pertussis, is a highly transmissible respiratory infection caused by
20 *Bordetella pertussis*. Due to the high burden of pertussis, vaccine programmes were
21 introduced internationally and in Australia since the 1950s. This has resulted in a significant
22 decrease of pertussis infections. However, since the 1990s the number of pertussis
23 notifications has increased considerably. Currently circulating *B. pertussis* strains differ in
24 vaccine antigen composition compared to strains that circulated in the pre-vaccination era.
25 These genetic differences are thought to contribute, in part, to the re-emergence of pertussis
26 in Australia and around the world. Whole genome sequencing (WGS) can resolve minute

27 differences in circulating strains and provides unparalleled resolution of vaccine antigens.
 28 This high-resolution snapshot can provide clues that enable more targeted public health
 29 interventions. However, pertussis is primarily diagnosed with culture-independent diagnostic
 30 assays which offer fast turnaround result times and reduced laboratory costs, eliminating the
 31 need to culture isolates. Current WGS methods require a cultured isolate, resulting in an
 32 absence of *B. pertussis* genome sequences in the post vaccination era. This scarcity has, in
 33 turn, limited understanding of currently circulating strains and respective vaccine antigen
 34 compositions.

35

36 Recent advancements of WGS technologies have allowed direct sequencing of clinical
 37 specimens without the need for a cultured isolate. However, recovering reliable sequence
 38 data from clinical samples of low bacterial load infections such as *B. pertussis* is a pressing
 39 challenge. We sought to increase the yield of *B. pertussis* sequences direct from a clinical
 40 sample by evaluating widely available WGS library preparation methods.

41

42 We report that the Illumina DNA prep library preparation kit combined with deep sequencing
 43 allowed the detection of important surveillance information such as allelic variations in the *B.*
 44 *pertussis* vaccine antigens. Further, our method generates high coverage over the 23S
 45 ribosomal RNA of *B. pertussis* enabling macrolide resistance to be easily determined.
 46 Overall, this method can improve surveillance of *B. pertussis*, by monitoring changes in
 47 vaccine antigens, detecting antimicrobial resistance and guiding Public Health control
 48 interventions.

49

50 **Introduction**

51 *Bordetella pertussis*, the primary causative agent of whooping cough, is a highly contagious
 52 respiratory pathogen that causes mild symptoms in adults but severe infection with a high
 53 mortality rate in infants. Global vaccination efforts have significantly reduced pertussis
 54 mortality, however, pertussis remains endemic globally, even in countries with high vaccine
 55 coverage¹. The cause of the recent resurgence is hypothesised to include waning host
 56 immunity² and vaccine escape³. Investigation of the persistence of pertussis has been
 57 challenged by the lack of *B. pertussis* isolates. The introduction of highly sensitive and fast
 58 PCR-based diagnostics has rendered culture of *B. pertussis* largely redundant, and limited the
 59 ability to perform phenotypic strain typing and antibiotic susceptibility testing⁴. The re-
 60 emergence of pertussis, along with the emergence of macrolide resistant *B. pertussis*,
 61 demonstrates how critical it is that we continue to monitor *B. pertussis* strains, despite a
 62 decline in the number of culture isolates.

63
 64 Whole genome sequencing (WGS) of bacteria with epidemic potential provides highly
 65 nuanced data^{5, 6} used in outbreak control where it can identify clusters and potential
 66 transmission chains⁶. Read depth, defined as the number of reads over a genome position, and
 67 genome coverage, the percentage of read coverage across the genome, are important
 68 parameters in direct sequencing protocols as they provide assurance of the accuracy and
 69 quality of the consensus sequence. When this high-resolution technology is applied directly
 70 to a clinical sample the overwhelming majority of sequences obtained are of human origin. In
 71 a respiratory sample, the number of human cells poses the greatest obstacle to resolving
 72 sufficient sequencing reads for a low-load pathogen, such as *B. pertussis*⁶. Studies have
 73 previously shown that pre-treatment of nasopharyngeal aspirates (NPA) can reduce human
 74 DNA levels, and increase target pathogen DNA and therefore pathogen specific sequencing
 75 reads^{4, 7}. However, despite the increased yield of *B. pertussis* from a saponin treated sample,

microbial genome coverage and read depth was still insufficient in providing informative molecular typing data⁴. One way to improve sequence coverage is increase the number of total reads per samples, known as deep sequencing. Deep sequencing aims to increase the read depth and the accuracy of detecting mutations over positions in genes of interest by generating more unique reads across the genome⁸. However, read depth and coverage can also be enhanced by selecting the appropriate library preparation method⁹.

82

Library preparation kits play a critical role in coverage yield, as the method of fragmentation used can introduce significant biases. For *B. pertussis*, G+C bias is a major limiting factor in transposase and PCR-based library preparation methods as it results in a loss of reads and low coverage over high G+C content areas of the genome¹⁰⁻¹³. Considering *B. pertussis* has an average G+C content of 67%, library preparation methods (e.g., Nextera XT) can bias library construction to human DNA present and not produce libraries as efficiently from *B. pertussis* DNA present in a clinical sample¹⁰⁻¹³. Moreover, in contrast to mechanical fragmentation-based library preparation methods, enzymatic fragmentation contains slightly greater insertion biases due to sequence preferences of the transposase¹⁴⁻¹⁶. While there is little impact of these insertion biases on rapid, parallel sequencing of isolates, the effect of this bias on sequencing of a high G+C content pathogen directly from clinical samples cannot be ignored and requires further investigation^{12, 17}.

95

This study examines the influence of library preparation on culture-independent sequencing of *B. pertussis*. We compare library preparation kits from Illumina, the most widely used platform in Public Health Microbiology laboratories and assess the read depth and genome coverage of a low-load, high GC pathogen. Further, by comparing libraries at increased

sequencing depths we investigate whether accurate genomic information relevant for public health control of *B. pertussis* can be recovered directly from a clinical sample.

Methods

Nasopharyngeal aspirates (NPA) were collected from the Centre of Infectious Diseases and Microbiology – Laboratory Services (CIDMLS), NSW Health Pathology, Westmead Hospital, Sydney. These NPA specimens were submitted for clinical testing of other infectious diseases and were negative for *B. pertussis*. The samples were pooled and 200 µL was taken for baseline extractions with the Qiagen DNeasy Blood and Tissue (QIAGEN, Germany) DNA Extraction Kit (further described below) to confirm the NPA was *B. pertussis* negative by PCR. The pooled NPA were then spiked with 0.5 McFarland suspension of *B. pertussis* (ATCC®9797™ 18323) corresponding to approximately 1.4×10^6 CFU/mL (BP10-2), 1.4×10^5 CFU/mL (BP10-3), and 1.4×10^4 CFU/mL (BP10-4), then stored at -20°C. The specific dilutions were chosen as C_T cycles by *IS481* real-time PCR (rtPCR) demonstrate these are within the normal range of clinical samples¹⁸. Further clinical NPA (n=2) that were positive for *B. pertussis* and collected by CIDMLS with the highest C_T cycle were included for standard and deep sequencing. Details of *IS481* PCR are presented in the Supplementary Material.

Extraction and DNA Quality Control

The spiked and clinical samples were treated with 0.025% Saponin and TurboDNase as previously described^{4, 7}. The spiked samples were then extracted with the Qiagen DNeasy Blood and Tissue (QIAGEN, Germany) DNA Extraction Kit for “Purification of Total DNA from Animal Tissues (Spin-Column Protocol)” with modifications. The modifications included lysis of the NPA with Proteinase K at 56°C for 1.5 hours. The DNA extracts were

split into four aliquots for use in rtPCR targeting *IS48I* and *ERV3* as previously described⁴,
and for use in library preparation method comparison (Supplementary Material).

Library Preparation and Sequencing

The three Illumina (Illumina, USA) library preparation kits selected were; NexteraXT DNA Library Preparation Kit v2.5, TruSeq DNA Nano Library Preparation Kit and the Illumina DNA prep Library Preparation Kit. Libraries were constructed following the manufacturer's protocol. The two clinical NPA samples were prepared with Illumina DNA prep only. Fragment size and quality was assessed by the Agilent TapeStation 4200 using the High-Sensitivity DNA ScreenTape Assay (Agilent, USA) on TruSeq input DNA pre- and post-sonication, and Nextera XT, Illumina DNA prep and TruSeq post-library construction. Library quantification was assessed using the KAPA (Roche Diagnostics, Switzerland) assay on all libraries for normalisation and final pool concentration. Illumina DNA prep libraries were sequenced on the Illumina NextSeq 500 instrument in conjunction with other routine samples that generate between 3-5 million reads per sample. TruSeq and Illumina DNA prep libraries were sequenced together on the Illumina MiniSeq High-Throughput platform. These were considered as sequenced on a standard level. In addition, Illumina DNA prep libraries were sequenced again on the Illumina NextSeq 500 with a High-Throughput flow cell for deep sequencing to generate around 80 million reads per sample.

Analysis

Quality Control

Analysis began with FastQC¹⁹(v 0.11.3) to assess the sequencing quality of raw reads, followed by trimming the ends of poor-quality reads with Trimmomatic (v0.36)²⁰ using optimised parameters (Leading:3 Trailing:3 SlidingWindow:4:20 Minlen:36).

150

151 Mapping

152 Trimmed reads were then mapped to the Human genome GRCh38.p12
 153 (GCA_000001405.27) by Burrows-Wheelers Aligner (BWA v0.7.12) using default settings²¹.
 154 Three strains of *B. pertussis* were utilised as mapping reference genomes – Tohama I
 155 reference genome (NC_002929.2), ATCC®9797™ 18323 whole genome (NC_018518.1) as
 156 this was the strain used to spike the NPA, or B1917 (NZ_CP009751.1) as this is a
 157 representative strain of currently circulating *B. pertussis*²².

158

159 Gene and Genome Counts

160 Percentage of each representative organism (Human and *B. pertussis*) was determined by the
 161 number of reads mapped to the respective genomes divided by the number of reads of the
 162 total sample.

163

$$\% \text{ of } B. \text{ pertussis reads} = \frac{\text{No. of } B. \text{ pertussis reads}}{\text{No. of Total reads}}$$

164

165 Manual visualisations of coverage across genes of interest such as the 23S ribosomal RNA,
 166 vaccine antigen genes: *ptx*, *prn*, and *fhaB* encoding regions were performed using Qiagen
 167 CLC Genomics Workbench 12 (Qiagen, Germany). The list of all genes of interest has been
 168 provided in Supplementary Material.

169

170 HTSeq (v1.1.2)²³ with default ‘union’ settings was used to count the number of unique reads
 171 mapped over these genes of interest, mapped reads over multiple genes were discarded.
 172 Samtools depth (v1.9)²⁴ and BEDtools intersectBed and genomecov (v2.25.0)²⁵ were used to

173 calculate the average read depth, gene and genome coverage and plotted with ggplot2
 174 (v3.3.0)²⁶.

175

176 **Results**

177 The first question this study aimed to address was whether the method of library preparation
 178 had any impact on the sequencing efficiency of a high G+C, low load microbial pathogen
 179 such as *B. pertussis*.

Table 1: Standard sequencing comparison of library preparation methods by average total number of reads across duplicates, percentage of human and *B. pertussis* reads, average coverage depth of the *B. pertussis* Tohama I reference genome and total coverage of both *B. pertussis* Tohama I and ATCC®9797™ 18323 genomes.

Library Preparation Method	CFU/mL	Post-depletion <i>IS481</i> PCR (C _T)	Post-depletion <i>ERV3</i> PCR (C _T)	Total Reads	% of Human reads	% of <i>B. pertussis</i> reads	Average Coverage Depth	Coverage of the <i>B. pertussis</i> Tohama I Genome (%)	Coverage of the <i>B. pertussis</i> ATCC Genome (%)
Nextera XT									
BP10-2	1.4 x 10 ⁶	23.53	36.48	1,465,653	40.9%	0.44%	1.36	57.02%	56.11%
BP10-3	1.4 x 10 ⁵	28.72	36.35	3,446,052	36.7%	1.29%	0.48	3.77%	3.77%
BP10-4	1.4 x 10 ⁴	32.75	36.09	1,690,120	55.8%	0.30%	0.18	0.77%	0.78%
Illumina DNA prep									
BP10-2	1.4 x 10 ⁶	Same as above		5,207,437	51.0%	4.25%	4.25	89.16%	93.28%
BP10-3	1.4 x 10 ⁵	Same as above		7,635,405	36.4%	0.89%	0.89	13.31%	13.78%
BP10-4	1.4 x 10 ⁴	Same as above		3,095,938	57.8%	0.37%	0.01	1.26%	1.27%
TruSeq									
BP10-2	1.4 x 10 ⁶	Same as above		Not Sequenced					
BP10-3	1.4 x 10 ⁵	Same as above		Not Sequenced					
BP10-4	1.4 x 10 ⁴	Same as above		514,032	67.5%	0.33%	0.34	0.33%	0.34%

NexteraXT DNA Library Preparation using standard sequencing

As expected, the percentage of *B. pertussis* reads identified decreased as CFU decreased (Table 1). Genome mapping to both *B. pertussis* (Tohama I and ATCC®9797™ 18323) genomes show only $56.5 \pm 0.5\%$ of the genomes are covered by these reads. Deeper analysis into coverage over the vaccine antigen and 23S ribosomal RNA showed poor coverage across the vaccine antigen genes, with BP10-2-1 failing (zero reads) to present reads to 10 out of 14 genes and BP10-2-2 failing 3 out of 14. BP10-3-1 contained reads to all genes, 170 reads mapped to *fhaB* which accounted for 85.6% of the gene. BP10-3-2, BP10-4-1 and BP10-4-2 failed to produce reads for 10, 10 and 11 genes out of 14, respectively.

Illumina DNA Library Preparation using standard sequencing

Similar to Nextera XT, the percentage of *B. pertussis* reads recovered decreased with decreasing CFU. However, unlike Nextera XT, genome mapping of BP10-2 showed Illumina DNA prep reads covered $91.2 \pm 2.1\%$ of both Tohama I and ATCC®9797™ 18323 genomes, thus, increasing whole genome coverage by 34.7%. Only one sample of BP10-4 was sequenced due to BP10-4-2 failing to pass library QC checkpoints with low library concentrations at the correct size of 300-600 bp. Compared to Nextera XT, Illumina DNA prep resulted in a more even distribution of reads across the vaccine antigens and 23S ribosomal RNA (Figure 1 and Figure 3). BP10-2 Illumina DNA prep at a standard sequencing level was able to provide 4.1 X coverage across all 14 vaccine antigen genes. BP10-3 and BP10-4 produced zero coverage for 6 and 10 genes respectively out of the 14. To test whether the library preparation result could be enhanced by deep sequencing, we sequenced these libraries (BP10-2, BP10-3 and BP10-4) at a concentration estimated to generate 80 million reads.

TruSeq DNA Nano Library Preparation using standard sequencing

Most samples of the TruSeq library preparation showed promising library concentration (236 – 3,350 pg/μL), however the majority of the library fragments were only 92-106 bp in size, which were too small to be sequenced. Only BP10-4-2 had a library peak at 490 bp with a concentration of 37.6 pg/μL, hence only this sample was sequenced. BP10-4-2 produced 127,330 reads with 426 reads to *B. pertussis*, these reads provided a depth of 0.34 X across 0.33% of the genome. TruSeq at standard sequencing level of BP10-4-2 failed to yield any reads to genes of interest (Supplementary Material)

Figure 1: Comparison of Standard Sequencing (SS) gene coverage map over the sequences of the vaccine antigens – *cyaA*, *fhaB*, *fim2*, *fim3*, *prn*, *ptxA* and *ptxP*. The figure also compares the Illumina DNA prep (FX) (Red) and Nextera XT (XT) (Teal) kits at a SS level. The TruSeq (TN) kit was not presented in this figure as no reads were obtained across these regions. Coverage was capped at a maximum of 40x coverage.

212 **Comparisons of Illumina DNA prep at Standard and Deep Sequencing Levels**

213 Based on the results above, all Illumina DNA prep libraries were re-sequenced on a High-
 214 Throughput Flow cell on the NextSeq 500 platform, delivering a total of 493,937,514 reads to
 215 5 samples – BP10-2 (n=2), BP10-3 (n=2) and BP10-4 (n=1). Average read depth and
 216 coverage across the genome are presented in

Table 2, and library quantification C_T cycles are presented in the Supplementary Material.

The number of reads in deep sequencing was amplified 16-18 times and accordingly improved average read depth across the genome by 5-16-fold. In the Illumina DNA prep deep sequencing level, BP10-2 reads provided full 100% coverage of all 14 genes with 50-90X depth, however BP10-3 and BP10-4 had partial coverage only.

Comparisons of standard and deep sequencing across the vaccine antigen genes of all Illumina DNA prep libraries are presented in Figure 2. Given that a mutation (A $\rightarrow$ G) present in position 2037 of *B. pertussis* Tohama I has been accepted as indicative of macrolide resistance in *B. pertussis*, coverage of this region was investigated (Figure 3). All kits provided coverage over position 2037 ranging from 66 X coverage in BP10-2-1 standard level Nextera XT to 2795 X coverage in BP10-2-1 deep sequencing Illumina DNA prep.

Table 2: Summary of Illumina DNA prep deep sequencing average reads between duplicate samples.

	CFU/m L	Post- depletion <i>IS481</i> PCR (C _T)	Post- depletion <i>ERV3</i> PCR (C _T)	KAPA PCR (C _T)	Total Reads	% of human DNA reads	% of <i>B.</i> <i>pertussis</i> DNA reads	Average read depth	Coverage across <i>B.</i> <i>pertussis</i> Tohama I (%)
BP10-2	1.4 x 10 ⁶	23.53	36.48	15.44	82,375,087	53.44%	2.71%	67.93	95.22
BP10-3	1.4 x 10 ⁵	28.72	36.35	14.32	138,165,091	49.76%	2.05%	16.07	4.68
BP10-4	1.4 x 10 ⁴	32.73	36.09	16.32	52,857,158	60.18%	0.45%	3.86	1.81
BP12	N/A	25.71	-	19.04	48,289,160	27.84%	2.20%	20.00	98.82
BP16	N/A	22.30	38.56	15.96	6,773,770	31.18%	4.14%	9.70	95.68

169

170

171 **Figure 2.** Gene coverage map for genes coding for vaccine antigens. Sequencing at Standard
172 Sequencing (Blue) and Deep Sequencing (Red) levels of Illumina DNA prep (FX) library are
173 shown. Coverage in this figure was capped at 125, only small regions of the genes contained
174 reads exceeding 400X coverage.

Figure 3. Coverage of the 23S rRNA gene of *B. pertussis*, comparing library preparation kits - Illumina DNA prep (FX) and Nextera XT (XT), and Deep Sequencing (DS) and Standard Sequencing (SS) depths, with varying *B. pertussis* spiked CFU concentrations (BP10-2 1.4×10^6 CFU/mL and BP10-3 1.4×10^5 CFU/mL duplicates).

Trials on Clinical NPA specimens

Standard sequencing using Illumina DNA prep of the two clinical samples BP12 and BP16, yielded 3,443,046 and 514,032 total reads, with 70,960 (2.06%) and 24,141 (4.69%) *B. pertussis* reads, respectively. These reads produced an average coverage depth of 2.54 X for BP12 and 1.57 X for BP16. Furthermore, BP12 reads covered 50.2% of the B1917 (NZ_CP009751.1) genome, while BP16 covered 42.5%. When these libraries were sequenced at a deeper level the genome coverage increased by +48.7% and +53.2% for BP12 and BP16 respectively

Table 2 and Figure 4. Analysis of the genes of interest showed variability in recovery with most genes having more than 92% coverage, with an average depth of 32.91 X and 10.98 X respectively (Supplementary Material).

Figure 4: Whole genome coverage map of the positive strand in BP12 and BP16 at Standard Sequencing (SS) (Blue and Pink) and Deep Sequencing (DS) Levels (Purple and Green), mapped to B1917 (NZ_CP009751.1). Panel A represents BP12 and Panel B represents BP16. Scale of both samples have been capped at 50 X depth.

Discussion

In the era of culture-independent diagnostic testing, the primary consequence is the loss of routine laboratory culture. Without cultured isolates, many gold-standard phenotypic typing systems have been lost. Therefore, further laboratory classification of currently circulating strains has become unavailable and this in turn affects public health outbreak control strategies. The aim of this study was to compare commercially available library preparation methods and sequencing depths to enhance our ability to directly sequence *Bordetella pertussis*-positive nasopharyngeal aspirates. Comparisons were made based on sequencing read depths, average genome coverage and the ability to accurately extract sequencing data from regions of interest such as the vaccine antigens.

208

209 This study demonstrated that different library preparation methods can impact the recovery of
210 important genomic information from culture-independent sequencing. Nextera XT libraries at
211 a standard sequencing depth produced uneven coverage across the genome, only accounting
212 for ~50% of the *B. pertussis* Tohama I genome, compared to more than 80% of the genome
213 when libraries are prepared with Illumina DNA prep. Overall, the Illumina DNA prep library
214 preparation kit generated better average coverage across the *B. pertussis* genome at both the
215 standard and deep sequencing depths. Sequencing at a deeper level only amplified depth and
216 increased confidence to detect SNPs. In line with what others have shown⁴, we found that
217 sequencing at a standard level with any library preparation method, was insufficient in
218 obtaining accurate molecular typing information.

219

220 We report variations in total reads between all Illumina DNA prep samples most likely due to
221 library normalisation and final library pooling. Interestingly, we noted that in this study the
222 library concentrations ranged between 3 C_T cycles from each other and these C_T cycles
223 reflected how many sequencing reads were dedicated to each sample. This variation is
224 generally acceptable at a standard sequencing scale, hence further investigations into the
225 effect of library concentration and normalisation would need to be performed to optimise
226 deep sequencing protocols and ensure an even distribution of sequencing resources for all
227 samples.

228

229 The TruSeq DNA Nano library kit was chosen based on its fragmentation method –
230 mechanical instead of enzymatic to reduce the inherent G+C bias of the transposase.
231 However, post-library preparation QC determined very poor concentrations of library
232 fragments in the appropriate range (400-500bp). TruSeq libraries potentially failed due to low

starting concentration of DNA given the sample had been host DNA depleted. Fragment size analysis was performed before and after sonication on BP10-2 samples, however low input DNA concentration made it difficult to observe fragment size peaks efficiently. Sequencing of one of the TruSeq libraries (BP10-4-2) proceeded due to the presence of a peak in the 400-500bp region. TruSeq libraries were constructed with other samples with higher starting DNA concentration, which sequenced successfully, hence library preparation methods and reagents were not the problem. Therefore, TruSeq was not an appropriate kit for low DNA concentrations as selection of sonication parameters is difficult to optimise for host depleted clinical samples with low concentration of DNA.

As expected, decreasing *B. pertussis* CFU in spiked samples were able to demonstrate that changes in bacterial load can influence the amount of sequencing reads recovered from a sample. Based on routine rtPCR results, clinical *IS481*-positive specimens have an average C_T cycle of 30.92, suggesting a CFU load of 1.4×10^4 (CFU/mL) equivalent to sample BP10-4, is a typical *B. pertussis* load for clinical NPA samples. As such, the higher number of *B. pertussis* cells in BP10-2 of 1.4×10^6 CFU is an uncommon occurrence in a clinical sample. The two clinical cases enrolled in this study had a very high initial C_T cycle of 12.15 (BP12) and 17.97 (BP16), and their total read counts of 48,289,160 and 6,773,770, respectively, demonstrate the importance of quality clinical sample collection ensuring high load of target bacterial DNA prior to sequencing to allow sufficient genome coverage.

To achieve adequate depth and coverage of the *B. pertussis* genome in a low bacterial load NPA sample would require further enrichment of the clinical sample and/or library. However, our results indicate that a *B. pertussis* positive sample with a CFU equivalent to or more than 1.4×10^6 CFU would, without enrichment, generate coverage and depth sufficient

for genomic surveillance analyses. By combining host DNA depletion protocols (Saponin), an appropriate library preparation method (Illumina DNA prep) and applying a deep sequencing approach (>50 million reads) with a library quantification result of more than 15 C_T cycles, should result in 20X coverage at least, across the regions of interest if not the whole genome of *B. pertussis*. While sequencing at a deeper level costs more compared to sequencing a pure isolate, it is now an option when there are few isolates to sequence in this era of PCR diagnostics.

This study was limited by the number of replicates that could be performed as a result of cost-limitations; hence results are subjected to sample bias. Numbers of NPA samples are also slowly declining as preferences to nasopharyngeal swabs are more popular. The volume of liquid in NPA are also very limited due to its use in previous diagnostic protocols and therefore replicates of clinical samples could not be performed. Trialling the protocol on more common respiratory sample types such as nasopharyngeal and throat swabs would also expand its use.

Future research should focus on improving sequencing depth and coverage over regions of interest, particularly genes coding for vaccine antigens and other molecular typing targets. Despite the expense, further deep sequencing trials on larger number of clinical NPA are required to validate the protocol and determine information yield across sample types.

In conclusion, our study demonstrated the feasibility of direct sequencing of *B. pertussis* from clinical NPA specimens with a high bacterial load and the recognition of potentially actionable targets in *B. pertussis* genome. This can be achieved through the combination of optimised sample and library preparation followed by deep sequencing.

283

284 **Author Statements**

285 The study was conceptualised by WF, VT, and VS. Laboratory work was performed by WF,
286 RR and VT. Bioinformatics and genome analysis was executed by WF, KP and RS. The
287 manuscript was written by WF and reviewed and edited by KP, RR, RS, VT and VS.

288

289 The authors declare no conflict of interest

290

291 Reads mapping to *Bordetella pertussis* ATCC9797 18323 for spiked nasopharyngeal
292 aspirates and BP1917 for clinical nasopharyngeal aspirates have been uploaded to SRA under
293 BioProject: PRJNA694997

294

295 This study as supported financially by the Centre for Infectious Diseases and Microbiology –
296 Public Health Post-graduate scholarship

297

298 Consent for the images in this article has been provided by the authors.

299

300 This work was supported by the Prevention Research Support Program, funded by the New
301 South Wales Ministry of Health. Special thanks to Illumina for providing a complimentary
302 Illumina DNA prep Library Preparation Kit. The authors are grateful to the staff of Centre for
303 Infectious Diseases and Microbiology Laboratory Services, NSW Health Pathology for their
304 technical assistance and expertise. Computational analysis was performed on the University
305 of Sydney High Performance Cluster, with the assistance of the Sydney Informatics Hub. The
306 authors would like to acknowledge the Microbial Genomics Reference Laboratory, Centre for

Infectious Disease and Microbiology –Public Health, Westmead Hospital, for their assistance with genome sequencing and bioinformatics analysis.

Nasopharyngeal aspirates were collected by the Centre for Infectious Diseases and Microbiology Laboratory services under the Western Sydney Local Health District Research Ethics and Governance committee. Project identifier: 2019/PID02294

Consent was not obtained from patients, as these NPA were left-over from previous diagnostic testing, and otherwise discarded. These NPA samples were pooled to prevent any identification during sequencing, and no record of identifiable data was collected.

References

1. Crowcroft NS, Stein C, Duclos P, Birmingham M. How best to estimate the global burden of pertussis? The Lancet Infectious Diseases. 2003;3(7):413-8. doi:10.1016/s1473-3099(03)00669-8
2. Schwartz KL, Kwong JC, Deeks SL, Campitelli MA, Jamieson FB, Marchand-Austin A, et al. Effectiveness of pertussis vaccination and duration of immunity. CMAJ. 2016;188(16):E399-E406. doi:10.1503/cmaj.160193
3. Bart MJ, van Gent M, van der Heide HG, Boekhorst J, Hermans P, Parkhill J, et al. Comparative genomics of prevaccination and modern Bordetella pertussis strains. BMC Genomics. 2010;11:627. doi:10.1186/1471-2164-11-627
4. Fong W, Rockett R, Timms V, Sintchenko V. Optimization of sample preparation for culture-independent sequencing of Bordetella pertussis. Microb Genom. 2020. doi:10.1099/mgen.0.000332
5. Yang J, Yang F, Ren L, Xiong Z, Wu Z, Dong J, et al. Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. J Clin Microbiol. 2011;49(10):3463-9. doi:10.1128/JCM.00273-11
6. Bachmann NL, Rockett RJ, Timms VJ, Sintchenko V. Advances in Clinical Sample Preparation for Identification and Characterization of Bacterial Pathogens Using Metagenomics. Front Public Health. 2018;6:363. doi:10.3389/fpubh.2018.00363
7. Hasan MR, Rawat A, Tang P, Jithesh PV, Thomas E, Tan R, et al. Depletion of Human DNA in Spiked Clinical Specimens for Improvement of Sensitivity of Pathogen Detection by Next-Generation Sequencing. J Clin Microbiol. 2016;54(4):919-27. doi:10.1128/JCM.03050-15
8. Depledge DP, Kundu S, Jensen NJ, Gray ER, Jones M, Steinberg S, et al. Deep sequencing of viral genomes provides insight into the evolution and pathogenesis of varicella

345 zoster virus and its vaccine in humans. *Mol Biol Evol.* 2014;31(2):397-409.
346 doi:10.1093/molbev/mst210

347 9. Rhodes J, Beale MA, Fisher MC. Illuminating choices for library prep: a comparison
348 of library preparation methods for whole genome sequencing of *Cryptococcus neoformans*
349 using Illumina HiSeq. *PLoS One.* 2014;9(11):e113501. doi:10.1371/journal.pone.0113501

350 10. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. Amplification-
351 free Illumina sequencing-library preparation facilitates improved mapping and assembly of
352 (G+C)-biased genomes. *Nat Methods.* 2009;6(4):291-5. doi:10.1038/nmeth.1311

353 11. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, et al. A large
354 genome center's improvements to the Illumina sequencing system. *Nat Methods.*
355 2008;5(12):1005-10. doi:10.1038/nmeth.1270

356 12. Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, et al. Rapid, low-
357 input, low-bias construction of shotgun fragment libraries by high-density in vitro
358 transposition. *Genome Biol.* 2010;11(12):R119. doi:10.1186/gb-2010-11-12-r119

359 13. Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, et al. Whole-
360 genome sequencing and variant discovery in *C. elegans*. *Nat Methods.* 2008;5(2):183-8.
361 doi:10.1038/nmeth.1179

362 14. Green B, Bouchier C, Fairhead C, Craig NL, Cormack BP. Insertion site preference of
363 Mu, Tn5, and Tn7 transposons. *Mob DNA.* 2012;3(1):3. doi:10.1186/1759-8753-3-3

364 15. Lan JH, Yin Y, Reed EF, Moua K, Thomas K, Zhang Q. Impact of three Illumina
365 library construction methods on GC bias and HLA genotype calling. *Hum Immunol.*
366 2015;76(2-3):166-75. doi:10.1016/j.humimm.2014.12.016

367 16. Kia A, Gloeckner C, Osothprapong T, Gormley N, Bomati E, Stephenson M, et al.
368 Improved genome sequencing using an engineered transposase. *BMC Biotechnol.*
369 2017;17(1):6. doi:10.1186/s12896-016-0326-1

370 17. Sato MP, Ogura Y, Nakamura K, Nishida R, Gotoh Y, Hayashi M, et al. Comparison
371 of the sequencing bias of currently available library preparation kits for Illumina sequencing
372 of bacterial genomes and metagenomes. *DNA Res.* 2019;26(5):391-8.
373 doi:10.1093/dnares/dsz017

374 18. Timms VJ, Fong W, Jeoffreys NJ, Sintchenko V. Evaluation of the BioGX BD-Max
375 PCR assay for detection of pathogenic *Bordetella*. *Pathology.* 2019;51(3):323-4.
376 doi:10.1016/j.pathol.2018.10.018

377 19. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data
378 2010 [Available from: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>]

379 20. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina
380 sequence data. *Bioinformatics.* 2014;30(15):2114-20. doi:10.1093/bioinformatics/btu170

381 21. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler
382 transform. *Bioinformatics.* 2009;25(14):1754-60. doi:10.1093/bioinformatics/btp324

383 22. Bart MJ, Zeddeman A, van der Heide HG, Heuvelman K, van Gent M, Mooi FR.
384 Complete Genome Sequences of *Bordetella pertussis* Isolates B1917 and B1920,
385 Representing Two Predominant Global Lineages. *Genome Announc.* 2014;2(6).
386 doi:10.1128/genomeA.01301-14

387 23. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-
388 throughput sequencing data. *Bioinformatics.* 2015;31(2):166-9.
389 doi:10.1093/bioinformatics/btu638

390 24. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
391 Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25(16):2078-9.
392 doi:10.1093/bioinformatics/btp352

393 25. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic
394 features. *Bioinformatics.* 2010;26(6):841-2. doi:10.1093/bioinformatics/btq033

- 395 26. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-
396 Verlag; 2016.
- 397 27. Bruinsma S, Burgess J, Schlingman D, Czyz A, Morrell N, Ballenger C, et al. Bead-
398 linked transposomes enable a normalization-free workflow for NGS library preparation.
399 BMC Genomics. 2018;19(1):722. doi:10.1186/s12864-018-5096-9
400