bioRxiv preprint doi: https://doi.org/10.1101/247445; this version posted January 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1
2
3
4 # Impact of DNA polymerase choice on assessment of bacterial
5 communities by a *Legionella* genus-specific next-generation sequencing
6 approach

7
8
9 Rui P. A. Pereira[a,c] #, Jörg Peplies[b], Ingrid Brettar[a], Manfred G. Höfle[a]

10
11
12
13 [a] Department of Vaccinology and Applied Microbiology, Helmholtz Centre for
14 Infection Research, Inhoffenstrasse 7, 38124 Braunschweig, Germany

15 [b] Ribocon GmbH, Fahrenheitstrasse 1, 28359 Bremen, Germany

16 [c] current address: School of Life Sciences, University of Warwick, Coventry, CV4 7AL,
17 UK

18
19
20
21 Keywords: KAPA HiFi, HotStarTaq, Illumina MiSeq, alpha-diversity, beta-diversity

22
23
24
25
26 #Address correspondence Rui P. A. Pereira; email: r.abreu-pereira@warwick.ac.uk; tel:
27 +44 (0) 24 765 24550

28
29

bioRxiv preprint doi: https://doi.org/10.1101/247445; this version posted January 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

30 **ABSTRACT**

31

32 The library preparation step is a major source of bias in NGS-based studies. Several PCR-
33 related factors might negatively influence the application of NGS tools in environmental
34 studies and diagnostics. Among the most understudied factors are DNA polymerases. In
35 our study, we evaluated the effect of DNA polymerase type on the characterisation of
36 bacterial communities, more precisely *Legionella*, using a genus-specific NGS approach.
37 The assay with proof-reading high fidelity KAPA HiFi showed better amplification yield
38 than the one with widely used non-proofreading HotStarTaq. *Legionella* community
39 richness metrics were significantly overestimated with HotStarTaq. However, the choice
40 of DNA polymerase did not significantly change the community profiling and
41 composition. These results substantiate the use of proof-reading high fidelity DNA
42 polymerases in NGS assays and highlight the need of considering the impact of different
43 DNA polymerases in comparative studies and future guidelines for NGS-based diagnostic
44 tools.

45

bioRxiv preprint doi: https://doi.org/10.1101/247445; this version posted January 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

46 **INTRODUCTION**

47

48 Since its arrival, next-generation sequencing (NGS) has undergone fast and
49 continuous progress in terms of technology, allowing enhanced performance, throughput
50 and accuracy in a shorter time [1]. NGS methodologies have been widely applied in
51 environmental microbiology to provide a broader characterisation of bacterial community
52 structure, a better comprehension of the influence of environmental and anthropogenic
53 drivers in both the emergence and persistence of bacterial pathogens; and for the detection
54 and characterisation of clinically and environmentally relevant organisms [2–4]. One of
55 the most environmentally and clinically important water-based bacterial taxa worth
56 studying is *Legionell*a. *Legionella* species are persistently and abundantly found in a
57 highly diverse set of aquatic environments and more than 25 species have been associated
58 with disease [5]. Legionellosis outbreaks are frequently reported and have been mostly
59 linked to hot water settings and foremost cooling towers, suggesting the need of a
60 continuous environmental surveillance. NGS assays have already been demonstrated to
61 have the capacity to provide a detailed understanding of the spatio-temporal dynamics
62 and diversity of *Legionella* species and be used as reliable and precise monitoring tools
63 of pathogenic species, such as *L. pneumophila* [6].

64 NGS has the potential to be an upgrade to the current accepted methodologies,
65 and consequently be translated into active frameworks for routine surveillance and
66 investigation of man-made freshwater systems. However, thorough *in vitro* and *in situ*
67 validation studies are needed to evaluate if NGS-based methodologies are precise and
68 reliable methodologies that are worth to be widely implemented and applied. Guidelines
69 have already been presented and implemented for the use of NGS in clinical diagnostics
70 and genetics screening [7–9]. Yet, no recommendations, guidelines or regulatory

bioRxiv preprint doi: https://doi.org/10.1101/247445; this version posted January 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

71 approaches currently exist for validation and establishment of the assays and
72 interpretation of the results in environmental settings. Among the factors to consider in
73 the validation of a NGS assay are the potential bias introduced by library preparation and
74 sequencing steps, which can affect not only sensitivity and specificity of the molecular
75 assay developed but also inter-laboratory comparison studies. The potential of NGS as an
76 accurate and sensitive tool for environmental research and diagnostics can be affected by
77 the choice of the primers, targeted taxonomic group, PCR thermo-cycling conditions,
78 template concentration and DNA polymerase [10–13].

79 In this study, our aim was to evaluate the impact of the choice of the DNA
80 polymerase have on assessment of microbial diversity, more specifically on the alpha-
81 and beta-diversity of *Legionella* community in freshwater systems, with a genus-specific
82 16S rDNA-based NGS approach.

83

84 **MATERIALS AND METHODS**

85

86 To study the effect of the DNA polymerase on *Legionella* libraries community
87 diversity metrics and structure, 8 freshwater samples, i.e. 2 cold drinking water (A and
88 B), 2 hot drinking water (C and D) and 4 cooling tower water (E to H) samples, were
89 collected at the campus of the Helmholtz Centre for Infection Research (HZI), in
90 Braunschweig, Germany, and processed as previously described [6,14,15]. These
91 freshwater samples were then analysed by a genus-specific NGS approach using two
92 optimised and independent assays with two hot-start DNA polymerases, HotStarTaq
93 (Qiagen, Hilden, Germany) and proofreading KAPA HiFi (KAPA Biosystems,
94 Wilmington, MA, USA). A 16S rRNA gene fragment, with a length of 421 bp, comprising
95 the V3-V4 hypervariable regions was amplified with *Legionella* genus-specific primer

bioRxiv preprint doi: https://doi.org/10.1101/247445; this version posted January 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

96 pair Lgsp17F 5´-GGCCTACCAAGGCGACGATCG-3'/ Lgsp28R 5´-
97 CACCGGAAATTCCACTACCCTCTC-3' [16]. Detailed descriptions of the used
98 Illumina MiSeq-based approach with both DNA polymerases as well as of the 16S rDNA
99 data processing and taxonomic classification by the bioinformatics pipeline of SILVA
100 project are given elsewhere [17,18].

101

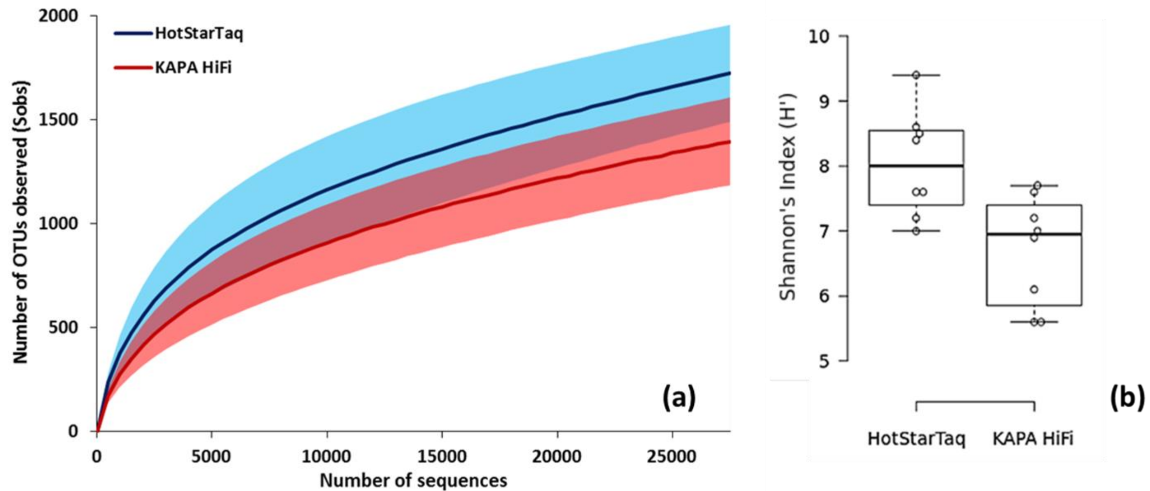
102 **RESULTS**

103

104 Before evaluating the effect of the DNA polymerase on community diversity
105 metrics, we compared the yield of the two optimised assays by equivolume pooling of all
106 sample libraries and sequencing in a single Illumina MiSeq 250 bp paired-end run. The
107 data revealed a statistically significant difference between the sample sets concerning the
108 number of retrieved 16S rRNA *Legionella* gene sequences per sample, after quality
109 filtering and taxonomic assignment (Welch's t-test, $P<0.05$). The mean number of
110 *Legionella* 16S rRNA gene sequences per sample in the sample set amplified by KAPA
111 HiFi was significantly higher (52,385 ± 12,907) than in the one amplified by HotStarTaq
112 (37,343 ± 8,995). Of the 8 samples, 7 had a higher number of sequences when amplified
113 with KAPA HiFi, representing, on average, an increase of 45% when compared with
114 HotStarTaq. This discrepancy does not seem to be an outcome of poor sequence quality
115 as the sample sets presented a similar percentage of rejected sequences during quality
116 filtering (1%). These results suggest that the optimised assay with KAPA HiFi
117 polymerase provided a higher *Legionella* amplification yield than the one with
118 HotStarTaq polymerase, regardless of the observed inferior specificity to *Legionella*
119 genus (86% ± 11% against 97% ± 2%). Nonetheless, rarefaction analyses revealed

bioRxiv preprint doi: https://doi.org/10.1101/247445; this version posted January 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

120 Operational Taxonomic Units (OTU) values levelling off markedly for all samples (data
121 not shown).
122



123
124 **Fig. 1. Alpha-diversity assessment of *Legionella* communities in 8 water samples**
125 **with HotStarTaq and KAPA HiFi DNA polymerases. (a)** Rarefaction curves showing
126 progress of number of OTUs observed with increasing number of sequences analysed.
127 Number of OTUs observed were calculated after normalisation to the sample with the
128 lowest number of sequences. Lines indicate mean values with corresponding coloured
129 shaded standard deviation for HotStarTaq DNA polymerase (blue) and KAPA HiFi DNA
130 polymerase (red). (**b**) Boxplots of Shannon's diversity index (*H'*) of 8 water samples
131 amplified with both DNA polymerases. Bars within boxes indicate the median value of
132 diversity. Whiskers extend to data points that are less than 1.5 x the interquartile range
133 away from $1^{st}/3^{rd}$ quartile.
134

135 Alpha-diversity metrics, i.e. observed OTU richness (OTU_{obs}) and Shannon's
136 diversity index (*H'*), were calculated, using the software package Explicet [19], for the
137 *Legionella* community of the 8 water samples amplified with both KAPA HiFi and
138 HotStarTaq DNA polymerases, after clustering of NGS reads into OTUs using an identity
139 criterion of 98%. (**Fig. 1**). The use of HotStarTaq, comparatively to KAPA HiFi, showed
140 a steeper slope and exhibited a higher *Legionella* richness with a statistically significant
141 increase of the number of OTUs observed (Welch's t-test, $p<0.05$) regardless of the

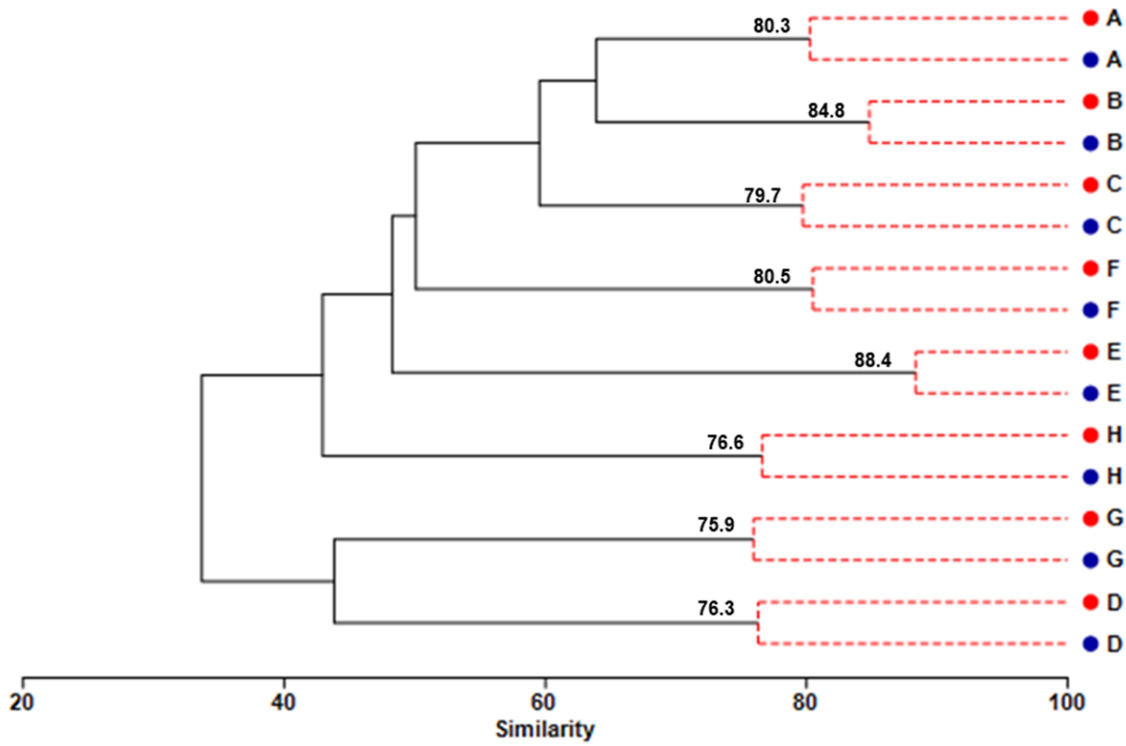
bioRxiv preprint doi: https://doi.org/10.1101/247445; this version posted January 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

142 sample (**Fig. 1(a)**). The sample set when amplified by HotStarTaq had an averaged
143 observed richness of 2389 ± 851 OTUs. However, when amplification performed with
144 KAPA HiFi, the sample set showed an averaged observed richness of 1846 ± 469 OTUs,
145 representing a decrease of 23% to the values of *Legionella* observed richness shown with
146 HotStarTaq. Also, the number of singletons and doubletons were reduced by an average
147 of 20% and 22%, respectively, with KAPA HiFi. Similar findings were observed when
148 separately amplifying technical replicates (n=3) of a single strain (*L. pneumophila* ATCC
149 33152^{T}), with a total reduction of 38% in singletons and doubletons detected when using
150 KAPA HiFi (data not shown).

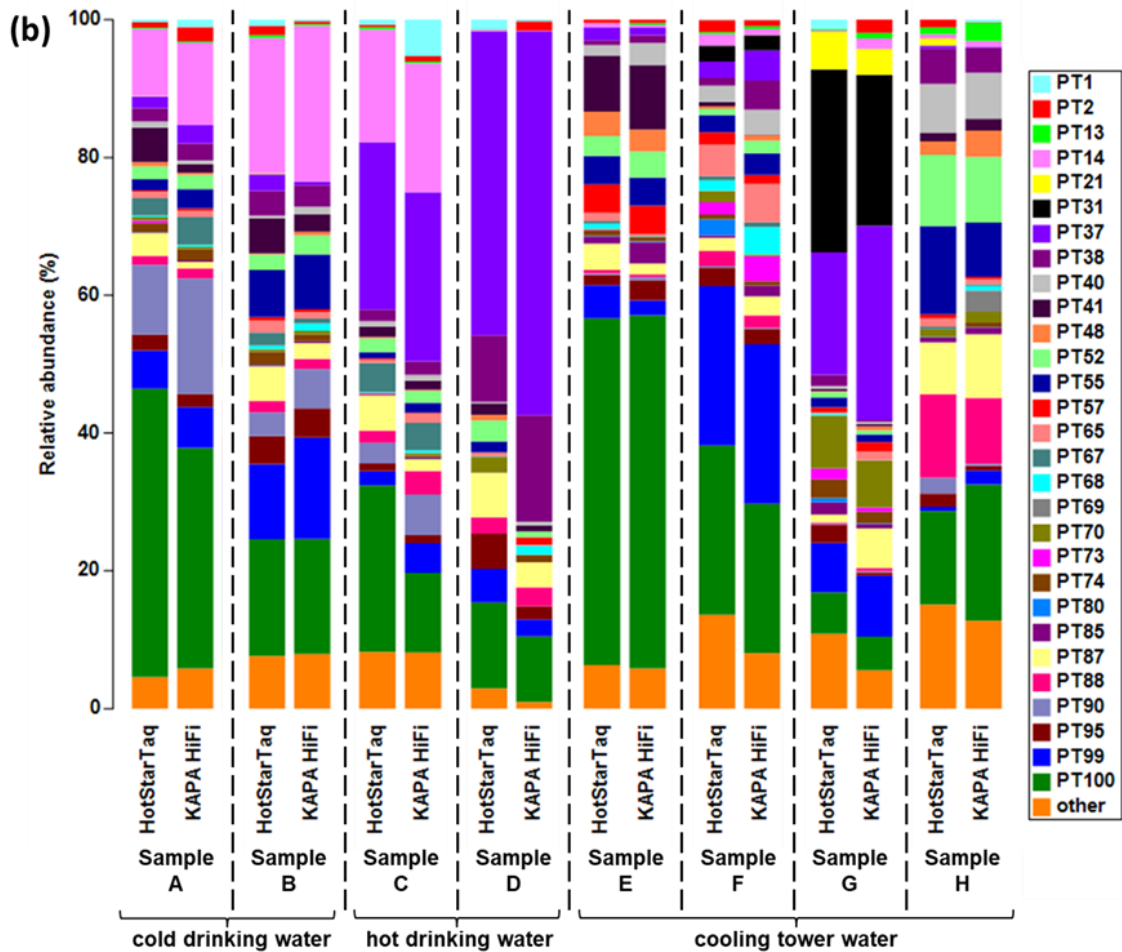
151 Moreover, the comparison of the two sample sets showed a statistically significant
152 difference in the Shannon´s diversity index (Welch's t-test, $P<0.05$) with higher values
153 observed in samples amplified with HotStarTaq (**Fig. 1(b)**). Shannon´s diversity index on
154 the samples amplified with HotStarTaq ranged from 7.06 to 9.94 (mean: 8.12 ± 0.97)
155 whereas on the samples amplified with KAPA HiFi the index values ranged from 5.64 to
156 7.99 (mean: 6.86 ± 0.88).

157 To assess beta-diversity, we characterised the relative abundance and composition
158 of the *Legionella* phylotypes in the freshwater samples amplified with each DNA
159 polymerase (HotStarTaq and KAPA HiFi) and calculated the Bray-Curtis similarity,
160 using PRIMER (Version 7.0.7) [20], to compare the level of community structure
161 similarity between different samples and between the same sample amplified with the two
162 different DNA polymerases (**Fig. 2**). The term *Legionella* phylotype is used according to
163 what has been previously described [17]. Briefly, NGS sequences were assigned to a
164 species when sequence identity ≥ 97%. The term phylotype englobes the acknowledged
165 *Legionella* species as well as other defined sequence clusters with sequence identity
166 <97% to a known species.

bioRxiv preprint doi: https://doi.org/10.1101/247445; this version posted January 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



167



168

bioRxiv preprint doi: https://doi.org/10.1101/247445; this version posted January 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

169 **Fig. 2. Effect of DNA polymerases on assessment of *Legionella* communities**
170 **profiling. (a)** Dendogram showing group-average hierarchical clustering of *Legionella*
171 communities amplified in 8 water samples with both HotStarTaq (blue circle) and KAPA
172 HiFi (red circle) DNA polymerases, using Bray-Curtis similarity**.** SIMPROF test was
173 performed and red dashed lines in cluster analysis represent groups of samples that do
174 not significantly differ in their *Legionella* community (P>0.05). Values on branches
175 indicate Bray-Curtis similarity index. Sample A, cold drinking water March 2009; sample
176 B, cold drinking water April 2009; sample C, hot water March 2009; sample D, hot water
177 January 2014; sample E, cooling tower water April 2013; sample F, cooling tower water
178 June 2013; sample G, cooling tower water August 2013; sample H, cooling tower water
179 January 2015. **(b)** Bar graph representing relative abundances of *Legionella* phylotypes
180 in 8 water samples amplified with both HotStarTaq and KAPA HiFi DNA polymerase.
181 Phylotypes labelled from 1 to 52 correspond to described *Legionella* species. PT1, L.
182 adelaidensis; PT2, *L. anisa*; PT13, *L. drozanskii*; PT14, *L. dumofii*; PT21, *L. gratiana*;
183 PT31, *L. maceachernii*; PT37, *L. pneumophila*; PT38, *L. quateirensis*; PT40, *L.*
184 *rowbothamii*; PT41, *L. rubrilucens*; PT48, *L. tucsonensis*; PT52, *L. worsleiensis*.
185 Different freshwater samples are separated by a dashed line.
186

187 Based on SIMPROF analysis (**Fig. 2(a)**) [21], we observed that the difference
188 introduced in the *Legionella* community structure by the use of the two different enzymes
189 was not statistically significant in the sample set used (SIMPROF, P>0.05).
190 Nonetheless, shifts in the *Legionella* community of every sample studied were
191 observed (**Fig. 2(b)**), with the similarity levels ranging from 75.9 (Sample G) to 88.4
192 (Sample E), with an averaged value of 80.3 ± 4.4. Clear clusters for every sample,
193 independently of the type of water, were formed and unambiguously separated the 8
194 different water samples tested. Furthermore, analysis of the composition of the *Legionella*
195 community in each sample amplified with the two DNA polymerases, revealed a
196 substantial overlap of *Legionella* phylotypes. When the composition of the communities
197 amplified by the two enzymes were compared, 7 out of 8 samples shared more than 85%

bioRxiv preprint doi: https://doi.org/10.1101/247445; this version posted January 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

198 of the phylotypes (data not shown). Sample D showed the lowest percentage of shared
199 phylotypes (60.1%).
200

201 **DISCUSSION**

202

203 Currently, NGS technologies are widely applied in the characterisation of
204 communities, discovery of novel microorganisms and pathogen detection [22]. For NGS
205 assays to succeed on these fields, a high amplification efficiency during library
206 preparation and target enrichment is of critical importance to downstream sequencing,
207 especially when targeting specific low-abundant taxa in complex environmental samples.
208 The amplification efficiency is influenced by several PCR-related factors, including the
209 choice of DNA polymerase. Our results, despite a better performance of KAPA HiFi,
210 confirmed that the two different polymerases had a very good amplification performance
211 allowing a subsequent appropriate sequencing depth of the target. This was in a certain
212 amount expected due to the hot-start nature of the two enzymes. Plus, our findings
213 highlight that the choice of DNA polymerase, as well as optimisation of the cycling
214 parameters in PCR-based NGS assays, is important to find an appropriate balance
215 between non-specific amplification and yield, in order to maximise detection/sensitivity.
216 In addition, several studies have reported the impact of different amplification and
217 sequencing-related factors on high-throughput studies of microbial communities and how
218 these might distort diversity estimates and community profiling [23–25]. Yet, only a few
219 studies have evaluated the effects of DNA polymerases on the study of communities [26–
220 29].
221 Past NGS studies have revealed increasing richness values with increasing
222 sequencing depth [17,30]. Yet, our results showed that the freshwater samples amplified

bioRxiv preprint doi: https://doi.org/10.1101/247445; this version posted January 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

223 by HotStarTaq DNA polymerase, despite a lower sequencing depth, had significantly
224 higher OTU-based alpha-diversity metrics than KAPA HiFi DNA polymerase, indicating
225 that alpha-diversity measurements are enzyme-dependent, as previously suggested [26–
226 29].
227 A recent study by Pereira et al. [17], using the same approach, revealed that when
228 comparing the sequence accuracy and quality of the generated NGS *Legionella* libraries,
229 KAPA HiFi had a significantly higher fidelity, with a mean error rate of 0.38% against
230 0.45% and 65% of error-free sequences against 39%. All data considered, the significant
231 increase in bacterial community richness and diversity, plus the significant higher number
232 of singletons and doubletons observed after amplification of the same 16S rRNA region
233 with HotStarTaq DNA polymerase, is almost certainly linked to a higher generation of
234 erroneous sequences with an enzyme without proofreading activity. This leads to an
235 increased number of OTUs and consequently an artificial overestimation of the richness
236 and diversity estimators [23,31,32]. Similar inflated community richness with
237 polymerases presenting lower fidelity was observed in bacterial libraries by Wu et al. [26]
238 and fungal libraries by Oliver et al. [28]. Yet, conflicting reports by Ahn et al. [27] and
239 Qiu et al. [29] have detected an overestimation of richness with a high-fidelity enzyme
240 instead, attributing it to a higher frequency of chimeric sequences. In our study, as we are
241 targeting a specific taxa, low diversity libraries are generated, not allowing to precisely
242 quantify the potential amount of chimeric sequences in the water samples. However,
243 analysis of *Legionella* mock communities with the same NGS approach has revealed a
244 small representation of these spurious sequences with KAPA HiFi [17].
245 When comparing beta-diversity metrics after amplification with both DNA
246 polymerases, the impact of the enzymes on the *Legionella* community composition and
247 structure (beta-diversity) was not significant for the samples studied, contrarily to what

bioRxiv preprint doi: https://doi.org/10.1101/247445; this version posted January 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

248 was indicated by Wu et al. [26], when comparing the DNA polymerases TAKARA ExTaq
249 with PfuUltra II Fusion HS. Though HotStarTaq and KAPAHiFi have different
250 properties, kinetics and slightly different polymerase cocktails were used, the two
251 enzymes seem to have an alike amplification profile of the different *Legionella*
252 phylotypes. However, differences in Bray-Curtis similarity between the samples
253 amplified by two distinct DNA polymerases were observed (mean BC: 80.3 ± 4.4). These
254 discrepancies on *Legionella* community profiling are slightly higher to the ones
255 previously observed between technical replicates (mean BC: 86.8 ± 4.9) [17].
256 In summary, our data highlights the potential advantageous effects of the use of
257 proof-reading high-fidelity enzymes such as KAPA HiFi on NGS library preparation
258 methodologies. As well, this study emphasises the influence the choice of DNA
259 polymerase has on the characterisation of microbial communities, especially in alpha-
260 diversity metrics, and the critical importance of taking this factor in consideration in
261 comparative studies and in the future use of these high-throughput technologies for
262 pathogen detection and quantification.
263

264 **FUNDING INFORMATION**
265 This work was supported by the Deutsche Forschungsgemeinschaft (DFG-project No.
266 HO-930/5-1/2) and the EU project AQUAVALENS (No. 311846).
267

268 **ACKNOWLEDGEMENTS**
269 We gratefully acknowledge Josefin Koch, Marina Pecellin, René Lesnik and Verena
270 Maiberg for their work in environmental sampling and DNA extraction.

271 **CONFLICTS OF INTEREST**
272 The authors declare no conflict of interest.

bioRxiv preprint doi: https://doi.org/10.1101/247445; this version posted January 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

273

274 **REFERENCES**

275

276 1. **van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C.** Ten years of next-generation
277 sequencing technology. *Trends Genet*. 2014;30(9):418–426.

278 2. **Tan B, Ng C, Nshimyimana JP, Loh LL, Gin KY-H *et al.*** Next-generation
279 sequencing (NGS) for assessment of microbial water quality: current progress,
280 challenges, and future opportunities. *Front Microbiol*. 2015;6:1027.

281 3. **Wang H, Bédard E, Prévost M, Camper AK, Hill VR *et al.*** Methodological
282 approaches for monitoring opportunistic pathogens in premise plumbing: A review.
283 *Water Res*. 2017;117:68–86.

284 4. **Forbes JD, Knox NC, Ronholm J, Pagotto F, Reimer A.** Metagenomics: The next
285 culture-independent game changer. *Front Microbiol*. 2017;8:1069.

286 5. **Fields BS, Benson RF, Besser RE.** *Legionella* and Legionnaires' disease: 25 years
287 of investigation. *Clin Microbiol Rev*. 2002;15(3):506–526.

288 6. **Pereira RPA, Peplies J, Höfle MG, Brettar I.** Bacterial community dynamics in a
289 cooling tower with emphasis on pathogenic bacteria and *Legionella* species using
290 universal and genus-specific deep sequencing. *Water Res*. 2017;122:363–376.

291 7. **Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP *et al.*** Assuring the
292 quality of next-generation sequencing in clinical laboratory practice. *Nat*
293 *Biotechnol*. 2012;30(11):1033–1036.

294 8. **Matthijs G, Souche E, Alders M, Corveleyn A, Eck S *et al.*** Guidelines for
295 diagnostic next-generation sequencing. *Eur J Hum Genet*. 2016;24(1):2–5.

296 9. **Hardwick SA, Deveson IW, Mercer TR.** Reference standards for next-generation
297 sequencing. *Nat Rev Genet*. 2017;18(8):473–484.

298 10. **Eckert KA, Kunkel TA.** DNA polymerase fidelity and the polymerase chain
299 reaction. *PCR Methods Appl*. 1991;1(1):17–24.

300 11. **Chandler DP, Fredrickson JK, Brockman FJ.** Effect of PCR template
301 concentration on the composition and distribution of total community 16S rDNA
302 clone libraries. *Mol Ecol*. 1997;6(5):475–482.

303 12. **Sipos R, Székely AJ, Palatinszky M, Révész S, Márialigeti K *et al.*** Effect of
304 primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-
305 targetting bacterial community analysis. *FEMS Microbiol Ecol*. 2007;60(2):341–
306 350.

307 13. **Brandariz-Fontes C, Camacho-Sanchez M, Vilà C, Vega-Pla JL, Rico C *et al.***
308 Effect of the enzyme and PCR conditions on the quality of high-throughput DNA
309 sequencing results. *Sci Rep*. 2015;5:8056.

bioRxiv preprint doi: https://doi.org/10.1101/247445; this version posted January 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

310 14. **Eichler S, Christen R, Höltje C, Westphal P, Bötel J *et al.*** Composition and
311 dynamics of bacterial communities of a drinking water supply system as assessed
312 by RNA- and DNA-based 16S rRNA gene fingerprinting. *Appl Environ Microbiol*.
313 2006;72(3):1858–1872.

314 15. **Henne K, Kahlisch L, Höfle MG, Brettar I.** Analysis of structure and composition
315 of bacterial core communities in mature drinking water biofilms and bulk water of
316 a citywide network in Germany. *Appl Environ Microbiol*. 2012;78(10):3530–3538.

317 16. **Kahlisch L, Henne K, Draheim J, Brettar I, Höfle MG.** High-resolution in situ
318 genotyping of *Legionella pneumophila* populations in drinking water by multiple-
319 locus variable-number tandem-repeat analysis using environmental DNA. *Appl*
320 *Environ Microbiol*. 2010;76(18):6186–6195.

321 17. **Pereira RPA, Peplies J, Brettar I, Höfle MG.** Development of a genus-specific
322 next generation sequencing approach for sensitive and quantitative determination of
323 the *Legionella* microbiome in freshwater systems. *BMC Microbiol*. 2017;17:79.

324 18. **Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W *et al.*** SILVA: a
325 comprehensive online resource for quality checked and aligned ribosomal RNA
326 sequence data compatible with ARB. *Nucleic Acids Res*. 2007;35(21):7188–7196.

327 19. **Robertson CE, Harris JK, Wagner BD, Granger D, Browne K *et al.*** Explicet:
328 graphical user interface software for metadata-driven management, analysis and
329 visualization of microbiome data. *Bioinformatics*. 2013;29(23):3100–3101.

330 20. **Clarke K, Warwick R.** Change in marine communities: an approach to statistical
331 analysis and interpretation. 3rd ed. Primer-E Ltd, Plymouth, UK; 2014.

332 21. **Clarke KR.** Non-parametric multivariate analyses of changes in community
333 structure. *Aust J Ecol*. 1993;18(1):117–143.

334 22. **Pallen MJ.** Diagnostic metagenomics: potential applications to bacterial, viral and
335 parasitic infections. *Parasitology*. 2014;141(14):1856–1862.

336 23. **Pinto AJ, Raskin L.** PCR biases distort bacterial and archaeal community structure
337 in pyrosequencing datasets. *PloS One*. 2012;7(8):e43093.

338 24. **Lee CK, Herbold CW, Polson SW, Wommack KE, Williamson SJ *et al.***
339 Groundtruthing next-gen sequencing for microbial ecology-biases and errors in
340 community structure estimates from PCR amplicon pyrosequencing. *PloS One*.
341 2012;7(9):e44224.

342 25. **Kennedy K, Hall MW, Lynch MDJ, Moreno-Hagelsieb G, Neufeld JD.**
343 Evaluating bias of Illumina-based bacterial 16S rRNA gene profiles. *Appl Environ*
344 *Microbiol*. 2014;80(18):5717–5722.

345 26. **Wu J-Y, Jiang X-T, Jiang Y-X, Lu S-Y, Zou F *et al.*** Effects of polymerase,
346 template dilution and cycle number on PCR based 16S rRNA diversity analysis
347 using the deep sequencing method. *BMC Microbiol*. 2010;10:255.

bioRxiv preprint doi: https://doi.org/10.1101/247445; this version posted January 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

348 27. **Ahn J-H, Kim B-Y, Song J, Weon H-Y.** Effects of PCR cycle number and DNA
349 polymerase type on the 16S rRNA gene pyrosequencing analysis of bacterial
350 communities. *J Microbiol*. 2012;50(6):1071–1074.

351 28. **Oliver AK, Brown SP, Callaham Jr. MA, Jumpponen A.** Polymerase matters:
352 non-proofreading enzymes inflate fungal community richness estimates by up to
353 15%. *Fungal Ecol*. 2015;15:86–9.

354 29. **Qiu X, Wu L, Huang H, McDonel PE, Palumbo AV *et al*.** Evaluation of PCR-
355 generated chimeras, mutations, and heteroduplexes with 16S rRNA Gene-Based
356 Cloning. *Appl Environ Microbiol*. 2001;67(2):880–7.

357 30. **Claesson MJ, Wang Q, O'Sullivan O, Greene-Diniz R, Cole JR *et al.*** Comparison
358 of two next-generation sequencing technologies for resolving highly complex
359 microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic*
360 *Acids Res*. 2010;38(22):e200.
361
362 31. **Kunin V, Engelbrektson A, Ochman H, Hugenholtz P**. Wrinkles in the rare
363 biosphere: pyrosequencing errors can lead to artificial inflation of diversity
364 estimates. *Environ Microbiol*. 2010;12(1):118–23.

365 32. **Reeder J, Knight R.** The 'rare biosphere': a reality check. *Nat Methods*.
366 2009;6(9):636–7.