Internal control for process monitoring of clinical metagenomic next-generation sequencing of urine samples

Victoria A. Janes [a]#, Jennifer S. van der Laan [a], Sébastien Matamoros [a], Daniel R. Mende [a], Menno D. de Jong [a] and Constance Schultsz [a,b]

[a] Amsterdam UMC, University of Amsterdam, Medical Microbiology, Amsterdam, The Netherlands

[b] Amsterdam UMC, University of Amsterdam, Global Health - Amsterdam Institute for Global Health and Development (AIGHD), Amsterdam, The Netherlands

Running head: Internal control for clinical mNGS of urine sample

#Corresponding author: Victoria A. Janes, v.a.janes@amsterdamumc.nl

ABSTRACT

Background

Process control for clinical metagenomic next-generation sequencing (mNGS) is not yet widely applied, while technical sources of bias are plentiful. We present an easy-to-use internal control (IC) method focussing on technical process control applied to metagenomics in clinical diagnostics.

Methods

DNA of nine urine samples was sequenced in the absence and presence of *Thermus thermophilus* DNA as IC in incremental concentrations (0.5-2-5%). Between aliquots of each sample, we compared the IC relative abundance (RA), and after *in silico* subtraction of IC reads, the microbiota and the RA of pathogens. The optimal IC spike-in concentration was defined as the lowest concentration still detectable in all samples.

Results

The RA of IC correlated linearly with the spiked IC concentration (r^2=0.99). IC added in a concentration of 0.5% of total DNA concentration was detectable in all samples, regardless of human/bacterial composition and after *in silico* removal gave the smallest difference in RA of pathogens compared to the unspiked aliquot of the sample. The microbiota of sample aliquots sequenced in the presence and absence of IC was highly similar after *in silico* removal of IC reads (median BC-dissimilarity per sample of 0.059), provided samples had sufficient bacterial read counts.

Conclusion

T. thermophilus DNA at a percentage of 0.5% of the total DNA concentration can be applied for the process control of mNGS of urine samples. We demonstrated negligible alterations in sample microbial composition after *in silico* subtraction of IC sequence reads. This approach contributes toward implementation of mNGS in the clinical microbiology laboratory.

INTRODUCTION

Metagenomic next-generation sequencing (mNGS) holds potential as a rapid pathogen detection tool for clinical microbiology diagnostics(1). In mNGS, all DNA in a sample has an equal chance of being sequenced, including DNA from host and contaminant cells. This means the interpretation of mNGS results can be challenging. For example, common reagent contaminants were previously incorrectly identified as causative of infection(2, 3). Differences in sample type, starting DNA concentration of a sample, library preparation efficiency for GC-rich organisms, and sequencing depth can all potentially bias the mNGS readout(4–7). If mNGS is to be safely used for clinical diagnostics, potential sources of error or variation in library preparation and sequencing need to be considered and controlled for.

While blank extraction controls and mock community and/or positive control samples can be used to assess the level of contamination and library preparation efficiency at batch level, external controls are not suitable for process monitoring of individual samples that vary greatly in microbial and host composition, DNA concentration and possible inhibitors. To this end, an internal control (IC) consisting of low concentration exogenous DNA that is completely unrelated to potential pathogens and the host microbiota, can be spiked into extracted DNA from each sample and subsequently detected nested within the diagnostic test procedure. Detection of a constant relative abundance (RA) of IC across all samples would ensure the library preparation and sequencing to be technically successful, ruling out technical error in samples where no pathogen reads are detected.

The application of ICs as process control is standard practise in molecular diagnostic microbiology such as for pathogen-specific PCR(8), but is still lacking in reported diagnostic applications of mNGS(2, 9–14). Studies that do describe ICs in metagenomic analyses typically applied fixed amounts of short synthetic DNA aimed at quantification of microorganisms or quantification of competition between target DNA and host DNA in library preparation protocols that include a DNA amplification step(2, 6, 15–18). It is unclear if the type and fragment length of IC used for such protocols are directly applicable to PCR-free sequencing protocols, such as the protocol used for diagnostic mNGS of clinical urine samples(19). Spike-in quantification standards are often patented and expensive while simpler and cheaper spike-ins may suffice for technical process control. Most importantly, because the DNA concentration of clinical urine samples varies greatly, spiking samples with a fixed amount of IC will result in overrepresentation of IC in low concentration samples and underrepresentation of IC reads in high concentration samples making such a strategy unsuitable for process control.

We describe the design and validation of an IC procedure for the process control of PCR-free library preparation and mNGS of clinical urine samples using full-length *Thermus thermophilus* DNA in a concentration that was titrated according to the DNA concentration of the urine sample.

METHODS

Urine sample collection

As part of a larger study of diagnostic mNGS of urinary tract infections (UTIs), we collected consecutive urine samples that were sent to the clinical microbiology laboratory of the Academic Medical Center, Amsterdam UMC for routine diagnostic semi-quantitative culture(19). Personal data were anonymised and handled in compliance with the General Data Protection Regulation and medical-ethical guidelines of the Academic Medical Center Amsterdam for anonymised use of patient materials. DNA extraction was performed after in-house enzymatic lysis, using the automated NucliSENS easyMag platform (Biomérieux) according to manufacturer's instructions(19). These samples were sequenced without IC and the DNA was stored at -80°C until further use. Nine representative samples of this collection were selected based on variations in bacterial versus human read counts, culture results, and DNA concentration (table 1) and were sequenced again in the presence of incremental concentrations of IC. The selected samples represented the extremes, i.e. both culture positive and negative high human/low bacterial read samples (A, B, C), high bacterial/low human read samples (F, G), equal human/bacterial read samples (E), and culture negative samples with a low DNA concentration (H, I).

IC procedure

We designed an IC with comparable GC-content and genome length to the common Gram-negative uropathogen *Escherichia coli* (~5.1Mb; GC% 50.6; NCBI:taxid 562). Further, the IC should not be a human pathogen or part of the human microbiota or have genomic similarity to these. The full-length bacterial genome of the extremophile *Thermus thermophilus* met these criteria and was previously successfully used as IC in analyses of environmental river microbiota(20).

To allow for detection of IC DNA at a low constant RA across all samples, we titrated the IC concentration relative to the sample DNA concentration. Therefore, DNA was first extracted from urine and quantified before addition of IC DNA at an appropriate concentration relative to the total DNA concentration.

DNA concentrations of stored DNA aliquots (-80 C) of each sample were measured using the Qubit HS dsDNA quantification kit (Thermo Fisher Scientific). The IC DNA concentrations were titrated to a final concentration of 0.5%, 2%, and 5% of the total DNA concentration of the sample. Not all concentrations were tested in each sample if limited amounts of DNA were available. IC DNA and sample DNA were mixed to a final DNA concentration of 1 ng/µl in a volume of 100µl (supplementary table 1). To assess variation between sequencing runs, two aliquots of the same sample were sequenced without IC in both runs (B and B1 + 0% IC sample, Table 1). In addition, B1+0% IC sample enabled us to see which proportion of reads were derived through contamination as this was the only sample sequenced without IC on the 2nd run. A sample only containing IC (*T. thermophilus* DNA suspended in TE buffer 1x, 1ng/µl) was sequenced to confirm that *T. thermophilus* reads were not misidentified as other taxa.

Library preparation and sequencing

Sample aliquots were sheared into 200bp fragments using the S220 Focused-ultrasonicator (Covaris). The Ion Xpress™ Plus Fragment Library Kit (Thermo Fisher Science) was used for PCR-free, library preparation according to manufacturer's specifications. Sequencing was performed on the Ion Torrent Proton and PGM (Thermo Fisher Scientific). This platform was used because of its sequencing speed which is required for clinical diagnostics. The platform was set to produce 2 million reads per sample for run 1 (10 samples sequenced as part of 55 samples in total) and 3 million reads per sample for run 2, both 200bp single-end (SE).

Sequence analysis

We used FastQC to check the read quality(21), Trimmomatic V0.38 to remove low quality reads(22), and BBMap to quantify and remove human reads by aligning to the GRCh38 Human Genome obtained from Genbank(23, 24). The IC and bacterial reads were identified using Kraken2 against the MiniKraken2_v1_8GB database, downloaded on the 2nd July 2019(25). The RA of IC was defined as the *Thermus* read count at genus level divided by the total number of reads per sample (including human and unclassified reads). We assessed the RA of the IC

per sample and per spiked IC DNA concentration to identify the lowest IC DNA concentration that was still detectable across all samples. We assessed the RA of IC in the B1+0% sample to assess the proportion of IC reads derived through cross-contamination. Next, all reads mapping to the genus *Thermus* were removed *in silico* before we used Bracken to calculate the RA of all classified bacterial species by Bayesian re-estimation of sequence reads(26). Bacterial species were annotated as urinary pathogens based on clinical microbiology common practice as well as the results of a scoping review of the literature (Supplementary table 2).

For each sample, the RA of urinary pathogens was compared between aliquots sequenced in the presence and absence of IC. In case of a polymicrobial composition, the cumulative RA of pathogens was used. The cumulative RA of pathogens was compared between sample aliquots sequenced in the presence and absence of IC after *in silico* subtraction of IC reads. To assess whether sequencing samples in the presence of IC and subsequent *in silico* subtraction of IC reads affected the sample microbiota, the RA of each detected species was compared between aliquots by calculating the Bray-Curtis dissimilarity (BC-dissimilarity). The optimal spike-in concentration of IC DNA was defined as the lowest still detectible IC concentration with minimum impact on the microbiota and cumulative RA of pathogens.

RESULTS

Sequencing results

We sequenced 28 aliquots of DNA obtained from 9 urine samples, spiked with incremental IC concentrations, in addition to one sample containing 100% IC DNA. The mean read count per sample aliquot was 1,486,600 (SD=252,004) for the unspiked samples of the first run, and 4,093,176 reads (SD = 615,999) for the spiked samples sequenced in the second run (table 1). The read distributions per sample are depicted in figure 1. In the samples with high human read count (samples A-D, H and I), the mean human read count varied between 1,2-1,4 million reads and the mean bacterial read count varied between 2,442-244,708 per sample. For the samples with high bacterial read count (samples E and F) the mean human read count per sample was 44,517 and 138,446 versus 3 and 3,4 million bacterial reads respectively. Resequencing the sample with equal bacterial/human read count (sample G) again produced an equal distribution of human and bacterial reads (figure 1). Two aliquots of

131 sample B were sequenced without IC on both runs. The RA of bacterial species of the two aliquots correlated strongly (BC-dissimilarity 0.061),

132 indicating results from the two sequencing runs were comparable.

Sample	Culture result	DNA concentration (ng/µl)	% spiked IC DNA	Total no. of reads	No. of IC reads	No. of unclassified reads	No. of bacterial reads	No. of human reads
A	*S. aureus* $>10^4$ CFU/ml	12,8	0	1800197	0	54004	686	1745507
			0.5	2834871	2598	3687	2761	2825825
			2	2906634	13076	4006	11769	2877783
			5	3476412	36171	7003	2124	3431114
B	*P. aeruginosa* $>10^4$ CFU/ml, *E. cloacae* $>10^4$ CFU/ml	17,68	0	1217911	0	31239	3697	1182975
			0	4230673	42	5609	16060	4208962
			0.5	3669184	6844	5283	13648	3643409
			2	3146673	22956	5870	12104	3105743
			5	3142635	58897	9333	11404	3063001
C	Commensal flora 10^4 CFU/ml	10,68	0	1222137	0	39799	16904	1165434
			0.5	4060689	8404	10986	45521	3995778
			5	3525448	62633	14092	38950	3409773
D	Commensal flora $>10^4$ CFU/ml	12,26	0	1018979	0	111370	94437	813172
			2	2469179	32993	225690	244708	1965788
			5	3107498	83231	271172	298285	2454810
E	*E. coli* $>10^4$ CFU/ml	11,68	0	1276427	2	45820	1222817	7788
			0.5	3283575	8834	143238	3081986	49517
			5	3891596	117764	176733	3534129	62970
F	*E. coli* $>10^5$ CFU/ml, K. *pneumoniae* $>10^5$ CFU/ml	22,00	0	1529193	0	57574	1423548	48071
			0.5	3935399	3266	181040	3612043	139050
			2	4386781	12872	201483	4013053	159373
			5	3572208	29667	162892	3241806	137843
G	*P. mirabilis* $>10^5$ CFU/ml, *M. morganii* $>10^5$ CFU/ml,*P. aeruginosa* $<10^4$ CFU/ml	9,78	0	1427110	0	56789	649441	720880
			5	2860345	54974	71919	1285320	1448132
H	Commensal flora 10^4 CFU/ml	8,2	0	1395035	0	17319	15325	1362391
			5	3628947	76747	29465	96711	3426024
I	Commensal flora 10^4 CFU/ml	7,94	0	1624302	0	108709	12326	1503267
			2	2780217	29602	117812	23174	2609629
IC	*T. thermophilus*	1,00	100	2680871	2425350	199119	10558	45844

Table 1. Characteristics of the sequenced urine samples. The total number of reads, IC, unclassified, bacterial and human reads per sample aliquot are stated. Semi-quantitative culture results are given in colony forming units per ml of urine (CFU/ml).

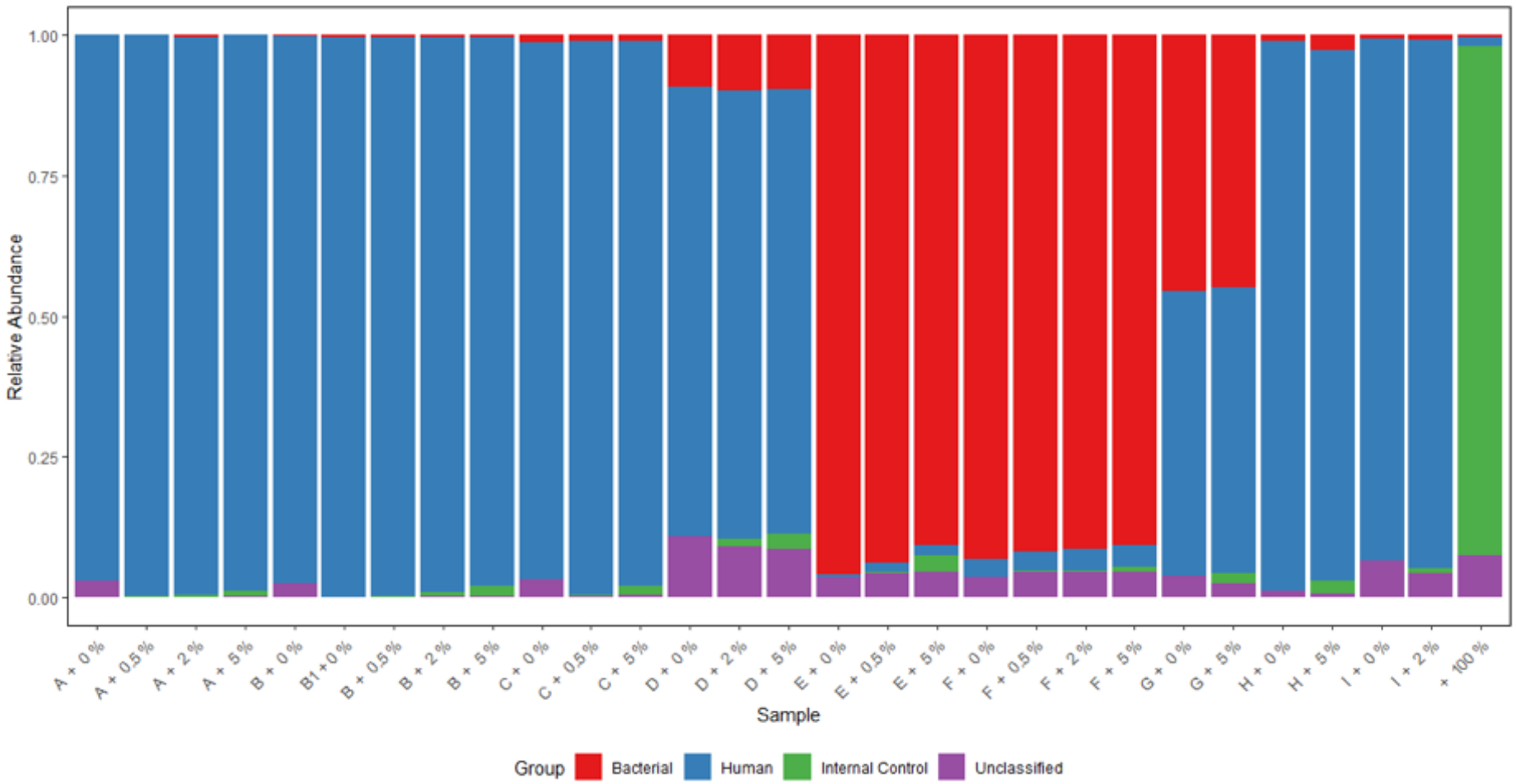


Figure 1. Read distributions. Shown is the total read distribution of each sequenced aliquot, indicating the RA of bacterial (red), human (blue), IC (green) and unclassified (purple) reads. Following the sample name (A – IC) is the spiked IC concentration (%).

Detection of IC

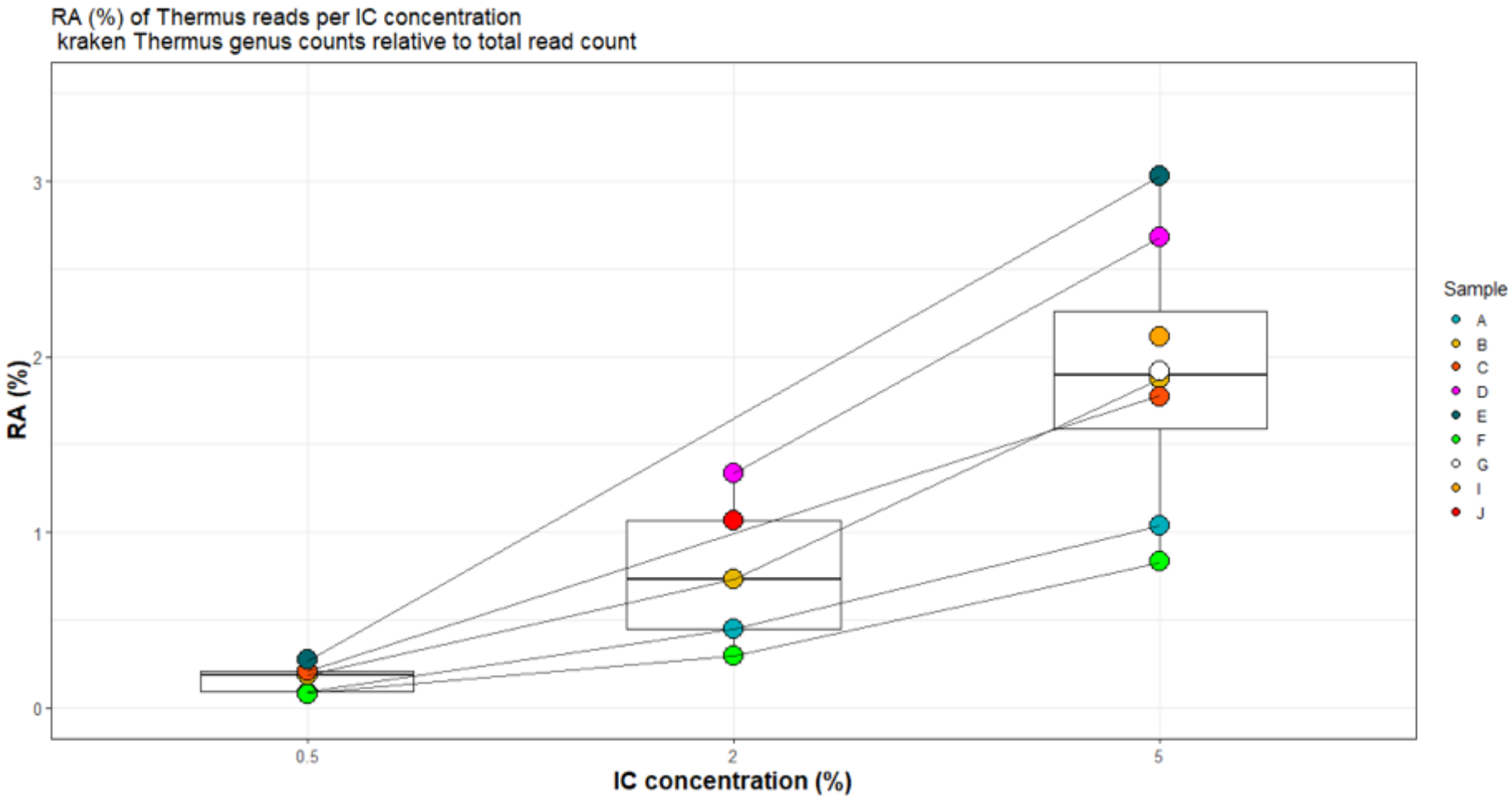
The resequencing of an aliquot of sample B without IC on the 2nd run (amidst samples sequenced in the presence of IC) not only allowed assessment of between-run variability but was also used to assess the proportion of IC reads derived through contamination. Out of 4,230,673 reads in this aliquot (aliquot name B1+0%), 42 reads (0.001%) identified as *Thermus* at genus level of which 39 identified as *T. thermophilus*.

To assess whether *T. thermophilus* reads mapped against other bacterial species, a sample containing 100% IC DNA was sequenced. Of all 2,635,027 reads, 2,425,350 reads (92%) mapped to the *Thermus* genus and of these 1,885,710 (71,6% of total) were classified as *T.*

148 *thermophilus*. 2,544 (0.096%) reads mapped to other bacterial species, predominantly to *E. coli* and *Klebsiella pneumoniae*, 8,014 reads (0.3%)

149 mapped to the domain Bacteria, 45,844 reads identified as human (1.7%), and 199,119 (7,5%) reads remained unclassified.

150



151

152 **Figure 2 RA of IC per sample and per IC concentration.**

153

154 Next, we assessed the RA of IC reads per sample at IC DNA concentrations of 0.5, 2 and 5% of the total DNA concentration. The RA of IC

155 reads rose linearly with increasing concentrations of spiked IC DNA (Pearson's r^2=0.99) and was 10.24 times higher in the samples spiked with

156 IC DNA at 5% of the total DNA concentration compared to the samples spiked with IC DNA at 0.5% of the total DNA concentration (figure 2).

157 There was no significant difference in RA of IC between samples with a high or equal bacterial read count (E, F and G) compared to samples

158 with a low bacterial and high human read count (A-D, H and I) (Welch two sample t-test, p = 0.8).

IC DNA spiked into samples at concentration of 0.5% of the total DNA concentration was detectable in all samples (median RA 0.19%; IQR 0.12%; range 0.09-0.27%). The median RA of IC reads at this concentration was 187 times higher than the RA of the IC reads in the aliquot of sample B that was sequenced in the absence of IC on run 2, and thus clearly exceeded the RA of IC expected to be derived through cross-contamination.

After *in silico* subtraction of IC reads, the RA of the bacterial species reads was highly similar between sample aliquots sequenced in the presence and absence of IC, illustrated by the similar species distribution between aliquots of the same sample seen in figure 3 and by the median BC-dissimilarity per sample of 0.059 (IQR 0.04, range 0.008-0.61). The exception was sample A with a BC-dissimilarity of 0.61. The aliquots of this sample had a median bacterial read count of only 2442 (range 686-11,769), over 61 times lower than the median bacterial read count of all other samples (median 150,363; range 12,043 (sample B) – 3,426,924 (sample F)).

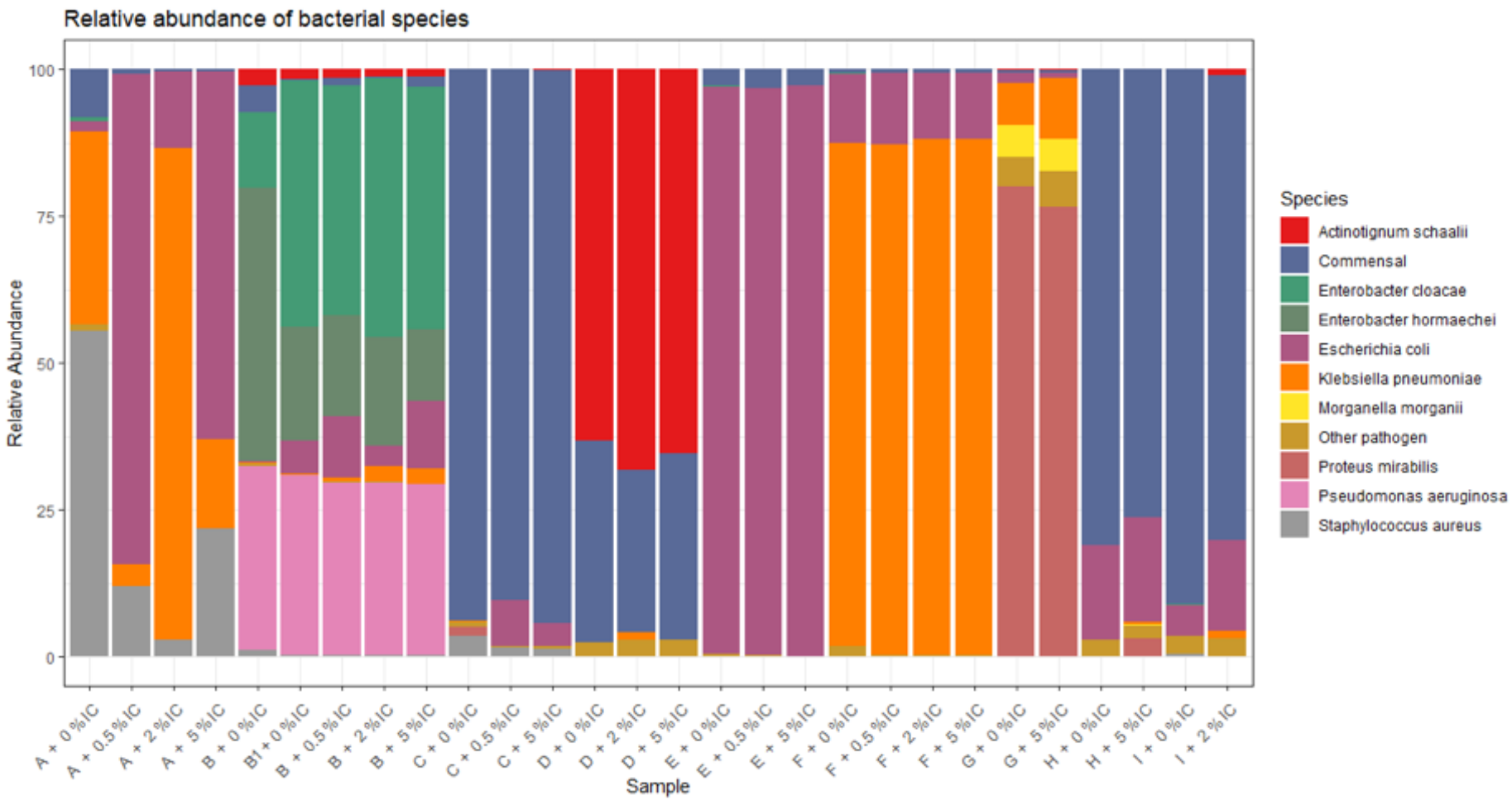


Figure 3. Relative abundance of bacterial pathogenic species and commensals reads per sample aliquot. Shown is the RA of the 9 most abundant bacterial pathogens classified after *in silico* removal of IC. Commensal bacterial species were grouped (ochre yellow).

Since the main outcome measure for diagnostic mNGS of urine samples was the RA of bacterial pathogens, we calculated the difference in RA of bacterial pathogens between spiked and unspiked sample aliquots after *in silico* subtraction of IC reads (table 2). An IC DNA concentration of 0.5% gave the smallest difference in cumulative RA of bacterial pathogens between spiked and unspiked sample aliquots, and was still detected in all samples. The median difference in RA was 1.15% (table 2).

	Relative Abundance of IC (%)			
IC concentration (%)	Min	Max	Median	IQR
0.5	0.17	6.72	1.15	0.82
2.0	0.28	7.94	4.14	4.53
5.0	0.15	4.16	1.22	2.54

Table 2. The median difference in cumulative RA of pathogens between spiked and unspiked aliquots of a sample after *in silico* removal of IC reads. Results are shown per IC concentration.

DISCUSSION

In this study we successfully applied *T. thermophilus* DNA as IC for the process control of mNGS of clinical urine samples. DNA aliquots extracted from urine samples were sequenced in the absence and presence of incremental concentrations of IC DNA. IC DNA added at a concentration of 0.5% of the total sample DNA concentration was detectable in all samples, did not alter the microbial composition of the sample DNA or the RA of detected bacterial pathogens substantially after *in silico* removal of IC reads. The results presented here are directly relevant to clinical microbiology practise as they are derived from clinical urine samples, representing the full array of variation seen in microbial and host DNA composition and concentration.

Other studies described the use of fixed amounts of short fragment synthetic DNA as IC aimed at quantification or as a measure for pathogen detection sensitivity(2, 6, 15–17). For such uses of IC, the RA of IC in the readout is integral to the analysis, making this a different approach to using IC as process control, where it is desirable to detect a constant low RA of IC in each sample to ensure the sequencing process was technically successful. The DNA concentration of urine samples is highly variable, meaning spiking a fixed amount of IC, as done in these

other studies, can easily lead to over- or underrepresentation of IC depending on the sample DNA concentration. Moreover, short fragments of IC could be sheared to even shorter fragments during library preparation and could consequently be lost during size selection when targeting species that require longer fragmentation time than the IC(20). The spiking of samples with IC in a concentration relative to the total DNA concentration of that sample, ensured detection of IC regardless of the concentration of the sample and thus allowed for consistent quality control of each sample.

Many bacteria that cause UTI, such as *E. coli*, are commensal to the genito-urinary tract and are only considered causative of disease when present in concentrations above a certain threshold(27). This study was not designed to address quantification, but quantification of urine mNGS results is needed to align with current routine culture-based microbiology diagnostics whereby a diagnosis is established based on semi-quantitative cultures.

Two reads were initially identified as *T. thermophilus* in sample D+0% sequenced on run 1 in the absence of IC. No samples were spiked with IC on this run. We hypothesised that this finding was caused by misclassification of Taq-polymerase DNA, derived from *Thermus aquaticus*, which may have regions of genomic similarity to *T. thermophilus*. However, BLASTn identified these 2 reads as *E. coli* plasmids p94EC-5.

When assessing the clinical relevance of species present at very low relative abundance, the likelihood that these reads are present as a result of cross-contamination should be taken into consideration. Relative abundances of 0.05-2.78% have been described for droplet cross-contamination during sample preparation(28). Barcode hopping, the incorrect assignment of library molecules from the expected barcode to a different barcode in a multiplexed pool, was observed in biological mock community samples at a rate of 0.033% on the IonTorrent PGM platform which has similar technology to the IonTorrent Proton(29). We demonstrate that the inclusion of IC DNA allowed for the assessment of cross-contamination by sequencing replicates in the presence and absence of IC. The impact of cross-contamination and barcode hopping appeared minimal in our setting. We propose that such a sample sequenced in the absence of IC should be included in each library preparation and sequencing run, to assess cross-contamination and to determine the minimum threshold of the RA of IC to consider a sample successfully sequenced.

By resequencing biological replicates in the presence and absence of IC divided over 2 sequencing runs, we demonstrated negligible alterations in sample microbial and pathogen composition after *in silico* subtraction of IC reads for 8 out of 9 samples. The incongruent sample A consisted of >95% human reads, leaving only a median of 2,442 (range 686-11,769) classified bacterial reads per sequenced aliquot. The low bacterial read count can explain the poor correlation between sequenced aliquots and this data may help set guidelines for a minimum sequencing depth for clinical samples and stresses the need for removal of host cells prior to sequencing(28, 29). Despite the variation in RA of bacterial species in sample A, the IC was detected in all aliquots, indicating sequencing and library preparation processes had been technically successful.

Our study has some limitations. Not every sample could be sequenced with each IC concentration, because limited volumes of DNA were present for some samples. However the overall results were consistent and RA of IC correlated linearly with spiked IC concentrations, implying our results can be extrapolated to other concentrations. Even lower spike-in concentrations could be suitable but were not tested in sufficient numbers. One aliquot of sample B was spiked with 0.05% IC DNA, yielding a RA of 0.025% (747 reads; data not shown).

Our approach does not control for DNA extraction efficiency. To achieve this, samples could be spiked with bacterial cells prior to DNA extraction. However, this comes with additional pitfalls. Spiking samples with a quantified bacterial cell suspension in a concentration of 0.5% relative to the sample will not necessarily result in 0.5% IC DNA. The human genome is several orders of magnitude longer than bacterial genomes, meaning each human cell produces approximately 1000 times more DNA than each bacterial cell(6). Given the number of sequencing reads produced per sequencing run is finite, host DNA will easily outcompete bacterial DNA, leading to difficulty in interpreting the mNGS read out: in cases where no pathogen reads are detected, distinguishing between a true negative (at the given sequencing depth) and a technical fail would be impossible. The spiking of samples with IC DNA in a concentration relative to the total DNA concentration of that sample mitigated this problem and ensured detection of IC regardless of the concentration or composition of the sample and thus allowed for consistent process control of each sample.

In conclusion we developed an IC for the process control of mNGS of urine samples that at a percentage of 0.5% of the total DNA concentration was detectable in all samples, regardless of the sample composition or DNA concentration. By resequencing sample aliquots in the presence and absence of IC, we demonstrated negligible alterations in sample microbial and pathogen composition after *in silico* subtraction

IC. We showed how a lower detection threshold for detection of IC can be established taking contamination into account. This approach could be used as process control for diagnostic mNGS and contributes toward implementation of mNGS in the clinical microbiology laboratory.

Acknowledgements

This study was supported by the COMPARE Consortium, which received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 643476. The authors declare no conflict of interest.

REFERENCES

1. Chiu CY, Miller SA. 2019. Clinical metagenomics. Nat Rev Genet 20:341–355.

2. Wilson MR, O'Donovan BD, Gelfand JM, Sample HA, Chow FC, Betjemann JP, Shah MP, Richie MB, Gorman MP, Hajj-Ali RA, Calabrese LH, Zorn KC, Chow ED, Greenlee JE, Blum JH, Green G, Khan LM, Banerji D, Langelier C, Bryson-Cahn C, Harrington W, Lingappa JR, Shanbhag NM, Green AJ, Brew BJ, Soldatos A, Strnad L, Doernberg SB, Jay CA, Douglas V, Josephson SA, DeRisi JL. 2018. Chronic Meningitis Investigated via Metagenomic Next-Generation Sequencing. JAMA Neurol 75:947–955.

3. Hornung BVH, Zwittink RD, Kuijper EJ. 2019. Issues and current standards of controls in microbiome research. FEMS Microbiol Ecol 95.

4. Videnska P, Smerkova K, Zwinsova B, Popovici V, Micenkova L, Sedlar K, Budinska E. 2019. Stool sampling and DNA isolation kits affect DNA quality and bacterial composition following 16S rRNA gene sequencing using MiSeq Illumina platform. Sci Rep 9:13837.

5. Bowers RM, Clum A, Tice H, Lim J, Singh K, Ciobanu D, Ngan CY, Cheng J-F, Tringe SG, Woyke T. 2015. Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. BMC Genomics 16:856.

6. Schlaberg R, Chiu CY, Miller S, Procop GW, Weinstock G. 2017. Validation of Metagenomic Next-Generation Sequencing Tests for Universal Pathogen Detection. Arch Pathol Lab Med 141:776–786.

7. Ye SH, Siddle KJ, Park DJ, Sabeti PC. 2019. Benchmarking Metagenomics Tools for Taxonomic Classification. Cell 178:779–794.

8. Deer DM, Lampel KA, González-Escalona N. 2010. A versatile internal control for use as DNA in real-time PCR and as RNA in real-time reverse transcription PCR assays. Lett Appl Microbiol 50:366–372.

9. Mongkolrattanothai K, Naccache SN, Bender JM, Samayoa E, Pham E, Yu G, Dien Bard J, Miller S, Aldrovandi G, Chiu CY. 2017. Neurobrucellosis: Unexpected Answer From Metagenomic Next-Generation Sequencing. J Pediatric Infect Dis Soc2017/01/06. 6:393–398.

10. Hu Z, Weng X, Xu C, Lin Y, Cheng C, Wei H, Chen W. 2018. Metagenomic next-generation sequencing as a diagnostic tool for

toxoplasmic encephalitis. Ann Clin Microbiol Antimicrob 17:45.

11. Ivy MI, Thoendel MJ, Jeraldo PR, Greenwood-Quaintance KE, Hanssen AD, Abdel MP, Chia N, Yao JZ, Tande AJ, Mandrekar JN, Patel R. 2018. Direct Detection and Identification of Prosthetic Joint Infection Pathogens in Synovial Fluid by Metagenomic Shotgun Sequencing. J Clin Microbiol 56.

12. Jun C, Huan H, Wei F, Duozhi S, Chen L, Yang S, GuoFeng G, Hao W, Qian Z, LiQing W, HongLong W, Long H, Luyao C, Jin Z, Shela L, FeiYan W, Zhou Z. 2019. Detection of pathogens from resected heart valves of patients with infective endocarditis by next-generation sequencing. Int J Infect Dis.

13. Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G, Salamat SM, Somasekar S, Federman S, Miller S, Sokolic R, Garabedian E, Candotti F, Buckley RH, Reed KD, Meyer TL, Seroogy CM, Galloway R, Henderson SL, Gern JE, DeRisi JL, Chiu CY. 2014. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. N Engl J Med2014/06/04. 370:2408–2417.

14. Hasman H, Saputra D, Sicheritz-Ponten T, Lund O, Svendsen CA, Frimodt-Møller N, Aarestrup FM. 2014. Rapid Whole-Genome Sequencing for Detection and Characterization of Microorganisms Directly from Clinical Samples. J Clin Microbiol 52:139 LP – 146.

15. Stämmler F, Gläsner J, Hiergeist A, Holler E, Weber D, Oefner PJ, Gessner A, Spang R. 2016. Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. Microbiome 4:28.

16. Tourlousse DM, Yoshiike S, Ohashi A, Matsukura S, Noda N, Sekiguchi Y. 2017. Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. Nucleic Acids Res 45:e23–e23.

17. Bal A, Pichon M, Picard C, Casalegno JS, Valette M, Schuffenecker I, Billard L, Vallet S, Vilchez G, Cheynet V, Oriol G, Trouillet-Assant S, Gillet Y, Lina B, Brengel-Pesce K, Morfin F, Josset L. 2018. Quality control implementation for universal characterization of DNA and RNA viruses in clinical respiratory samples using single metagenomic next-generation sequencing workflow. BMC Infect Dis 18:537.

18. Hardwick SA, Chen WY, Wong T, Kanakamedala BS, Deveson IW, Ongley SE, Santini NS, Marcellin E, Smith MA, Nielsen LK, Lovelock CE, Neilan BA, Mercer TR. 2018. Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis. Nat Commun 9:3096.

19. Janes VA, Matamoros S, Willemse N, Visser CE, de Wever B, Jakobs ME, Subramanian P, Hasan NA, Colwell RR, de Jong MD, Schultsz C. 2017. Metagenomic sequencing to replace semi-quantitative urine culture for detection of urinary tract infections: a proof of concept. bioRxiv.

20. Satinsky BM, Gifford SM, Crump BC, Moran MA. 2013. Use of Internal Standards for Quantitative Metatranscriptome and Metagenome Analysis. Methods Enzymol 531:237–250.

21. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data.

22. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

23. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. 2006. GenBank. Nucleic Acids Res 34:D16–D20.

24. Bushnell B. 2014. BBMap : A Fast , Accurate , Splice-Aware Aligner. LBNL Dep Energy Jt Genome Inst 3–5.

25. Wood DE, Salzberg SLS, Venter C, Remington K, Heidelberg J, Halpern A, Rusch D, Eisen J, Wu D, Paulsen I, Nelson K, Nelson W, Fouts D, Levy S, Knap A, Lomas M, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers Y-H, Smith H, Tyson G, Chapman J, Hugenholtz P, Allen E, Ram R, Richardson P, Solovyev V, Rubin E, Rokhsar D, Banfield J, Huttenhower C, Gevers D, Knight R, Abubucker S, Badger J, Chinwalla A, Creasy H, Earl A, FitzGerald M, Fulton R, Giglio M, Hallsworth-Pepin K, Lobos E, Madupu R, Magrini V, Martin J, Mitreva M, Muzny D, Sodergren E, Versalovic J, Wollam A, Worley K, Wortman J, Young S, Zeng Q, Aagaard K, Abolude O, Allen-Vercoe E, Alm E, Alvarado L, Altschul S, Gish W, Miller W, Myers E, Lipman D, Brady A, Salzberg SLS, Huson D, Auch A, Qi J, Schuster S, Brady A, Salzberg SLS, Rosen G, Garbarine E, Caseiro D, Polikar R, Sokhansanj B, Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M, Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C, Treangen T, Koren S, Sommer D, Liu B, Astrovskaya I, Ondov B, Darling A, Phillippy A, Pop M, Ames S, Hysom D, Gardner S, Lloyd G, Gokhale M, Allen J, Kindblom C, Davies J, Herzberg M, Svensäter G, Wickström C, Foweraker J, Cooke N, Hawkey P, Könönen E, Saarela M, Karjalainen J, Jousimies-Somer H, Alaluusua S, Asikainen S, Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden T, Zimin A, Marçais G, Puiu D, Roberts M, Salzberg SLS, Yorke J, Pruitt K, Tatusova T, Brown G, Maglott D, Marçais G, Kingsford C, Roberts M, Hayes W, Hunt B, Mount S, Yorke J, Magoc T, Pabinger S, Canzar S, Liu X, Su Q, Puiu D, Tallon L, Salzberg SLS, Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy A, Rigoutsos I, Salamov A, Korzeniewski F, Land M, Lapidus A, Grigoriev I, Richardson P, Hugenholtz P, Kyrpides N, Ondov B, Bergman N, Phillippy A. 2014.

321 Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:R46.

322 26. Lu J, Breitwieser FP, Thielen P, Salzberg SL. 2017. Bracken: estimating species abundance in metagenomics data. PeerJ Comput Sci
323 3:e104.

324 27. KASS EH. 1957. Bacteriuria and the Diagnosis of Infections of the Urinary Tract: With Observations on the Use of Methionine as a
325 Urinary Antiseptic. AMA Arch Intern Med 100:709–714.

326 28. Pereira-Marques J, Hout A, Ferreira RM, Weber M, Pinto-Ribeiro I, van Doorn L-J, Knetsch CW, Figueiredo C. 2019. Impact of Host
327 DNA and Sequencing Depth on the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis. Front
328 Microbiol 10:1277.

329 29. Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R, Jeanes C, Rae D, Grundy S, Turner DJ, Wain J, Leggett RM, Livermore
330 DM, O'Grady J. 2019. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. Nat Biotechnol
331 37:783–792.

332