Variant calling for cpn60 barcode sequence-based microbiome profiling

Sarah J. Vancuren, Janet E. Hill*

Department of Veterinary Microbiology, University of Saskatchewan

*To whom correspondence should be addressed: Janet.Hill@usask.ca

Abstract

Amplification and sequencing of conserved genetic barcodes such as the cpn60 gene is a common approach to determining the taxonomic composition of microbiomes. Exact sequence variant calling has been proposed as an alternative to previously established methods for aggregation of sequence reads into operational taxonomic units (OTU). We investigated the utility of variant calling for cpn60 barcode sequences and determined the minimum sequence length required to provide species level resolution. Sequence data from the 5´ region of the cpn60 barcode amplified from the human vaginal microbiome (n=25), infant stool (n=47), and a mock community were used to compare variant calling to de novo assembly of reads, and mapping to a reference sequence database. Results were compared in terms of number of OTU formed, resolution of OTU sequences, and overall community composition. Variant calling resulted in high resolution profiles that were consistent in composition to those generated with previous methods but with significant logistical advantages. Variant calling is rapid, achieves high resolution of taxa, and does not require reference sequence data. Our results further demonstrate that 150 bp from the 5´ end of the cpn60 barcode sequence is sufficient to provide species-level resolution of microbiota.

Keywords: cpn60, barcode, operational taxonomic unit, sequence variant, microbiome, metabarcoding

INTRODUCTION

Microbiome profiling is the process of determining which organisms are present in an environment, and their relative abundances. Profiling can be achieved by a "metabarcoding" approach involving targeted PCR and sequencing a genetic barcode: a conserved gene that is shared by many species and can be used to distinguish one from another (Gohl et al. 2016). Barcodes that have been demonstrated to meet the criteria established by the International Barcode of Life project (Hebert et al. 2003) include the cpn60 gene in bacteria (Links et al. 2012b), ITS in fungi (Schoch et al. 2012) and cytochrome c oxidase subunit 1 (COI) in animals (Hebert et al. 2003). Barcodes are important tools in distinguishing species when phenotypic differences are not conclusive; this is especially notable for prokaryotes. Barcode sequences must be universally conserved, so that a wide range of species can be distinguished by the sequence. The cpn60 gene encodes the 60 kDa chaperonin protein conserved in bacteria, eukaryotes and some archaea. A 549-567 bp region of this gene (the "universal target", UT) has been shown to provide greater resolving power between bacterial species than the widely used 16S rRNA gene sequences (Links et al. 2012b). The cpn60 UT sequence can be amplified with universal PCR primers and the chaperonin database, cpnDB (Vancuren and Hill 2019), provides a curated collection of chaperonin sequences that can be used to identify sequences. Sequencing of cpn60 UT amplicons from metagenomic samples has been used extensively in microbiome profiling (Hill et al. 2002; Dumonceaux et al. 2006a, 2006b; Desai et al. 2009; Mansfield et al. 2010; Chaban et al. 2013, 2014; Bondici et al. 2013; Albert et al. 2015; Freitas et al. 2017, 2018).

Metagenomic barcode amplicon sequence analysis relies on classifying sequence reads, usually involving formation of operational taxonomic units (OTU) or other methods to 'bin' or categorize reads. De novo assembly of OTU has been particularly useful for cpn60 barcode

sequences, allowing a high degree of taxonomic resolution of microbiota (Chaban et al. 2013, 2014; Costa et al. 2014; Links et al. 2014). Depending on the assembly method, mapping the raw reads on to the assembled OTU sequences can be subsequently performed to determine OTU abundances (Links et al. 2013a). More recently, reference mapping of cpn60 barcode sequence reads was established to speed up processing large amounts of data from microbial communities for which a comprehensive de novo OTU assembly is available, as well as to allow comparison of OTU across studies given a common reference set of OTU. In this procedure, raw reads are mapped on to a reference sequence database in order to "bin" reads into OTU and calculate their abundances (Links et al. 2013a; Albert et al. 2015; Freitas et al. 2017, 2018).

Formation of exact sequence variants ('variant calling') using a denoising algorithm such as DADA2 for Illumina sequence data, presents another alternative for formation of OTU, by distinguishing sequence errors from true variant positions (Callahan et al. 2016, 2017). Potential advantages to this approach include preservation of minor, but informative sequence differences among closely related OTU. Given the resolving power of the cpn60 barcode sequence, variant calling may offer OTU resolution similar to that achieved with de novo assembly, but more rapidly and with lower demand on computational resources. Another significant advantage to this approach is that since exact sequence variants are given a unique identifier based on the actual sequence, it is trivial to recognize identical OTU (variants) across multiple studies.

The objectives of the current study were to determine the length of cpn60 barcode sequence needed for species-level resolution, and to compare variant calling to previously established methods for cpn60 OTU generation in terms of logistics, and differences in apparent community composition. Results of this investigation will be applied in ongoing and future studies of microbiomes using the *cpn*60 barcode sequence.

METHODS

Mock community, vaginal microbiome and infant stool microbiome sequencing

A synthetic microbial community (the mixed vaginal panel, MVP) was generated by combining 20 cloned cpn60 barcode sequences as previously described (Dumonceaux et al. 2009). Plasmid DNA was extracted from *E. coli* with the EZ-10 Spin Column Plasmid DNA Miniprep Kit (Bio Basic Inc., Markham ON, Canada). Purified plasmid DNA was quantified, and diluted to 1 x 10^9 plasmid copies/µL with 10 mM Tris pH 8.5. Five µL of each plasmid solution was pooled to form the synthetic community. *cpn*60 UT PCR was performed in duplicate with a mixture 1:3 of *cpn*60 Illumina adapted primers M729/M280:M1612/M1613 (Table S1) as previously described (Hill et al. 2006). Sequence library preparation was conducted according to the Illumina MiSeq 16S Metagenomic Sequencing Library Preparation (Illumina Inc., San Diego, CA) protocol with some minor changes as optimized for the *cpn*60 UT: in PCR clean-up steps, 32.5 µL of magnetic beads were utilized. Libraries were sequenced with unrelated samples on the Illumina MiSeq plaform using a 500 cycle kit with 400 cycles for Read 1 (5´ end of the *cpn*60 UT) and 100 cycles for Read 2. Only Read 1 sequences were used in downstream analysis.

*cpn*60 UT amplicon sequence data from the vaginal (n = 25) and infant stool (n = 47) microbiome generated in unrelated studies using the method described above were selected for analysis. The original collection of the human microbiome samples was approved by the University of British Columbia Children's & Providence Health Care Research Ethics Board (certificate no. H14-01819) and UBC Children's & Women's Research Ethics Board (certificate no. H17-02253). All sequence data used in comparison of methods is deposited available in the

National Center for Biotechnology Information (NCBI) Sequence Read Archive under BioProject accession PRJNA562310).

Cutadapt version 1.15 (Martin 2011) and trimmomatic version 0.32 (Bolger et al. 2014) were used for amplification primer sequence removal and quality/length trimming.

Variant calling

Filtered sequence reads were imported to QIIME2 (Bolyen et al. 2019) and DADA2 (Callahan et al. 2016) with truncation lengths of 25, 50, 100, 150, 200, 250 or 300 was executed. OTU (variant) sequences were exported from QIIME2 and aligned using watered-BLAST (Schellenberg et al. 2009) to cpnDB_nr, a non-redundant version of the chaperonin database cpnDB (Hill et al. 2004; Vancuren and Hill 2019) that consists of *cpn*60 UT sequences with a single record for each species, and a preference for the type strain if available. OTU sequences that were over 55% identical to anything in cpnDB_nr were considered to be *cpn*60 sequences and retained for analysis (Johnson et al. 2015). The feature table was also exported, and sequence read counts were summed for OTU with the same nearest neighbour (NN) species based on results from the watered-BLAST analysis.

De novo assembly

Trinity version 2.4.0 (Grabherr et al. 2011) with kmer size of 31 was used to assemble quality filtered reads into OTU. OTU sequence frequencies were determined by mapping quality filtered reads on to the assembled OTU sequences using Bowtie2 version 2.3.3.1 (Langmead and Salzberg 2012 p. 2). The resulting feature table was created using components of mPUMA (Links et al. 2013b). OTU sequences were aligned to cpnDB_nr to determine nearest neighbours (NN, best sequence match). For some analyses, read counts for individual OTU with the same nearest neighbour species were combined.

Reference mapping

Filtered reads were mapped to the reference dataset with Bowtie2 version 2.3.3.1 (Langmead and Salzberg 2012). For the vaginal samples and mock community, the VOGUE Reference Assembly was used as the reference set. The VOGUE Reference Assembly is a curated reference dataset of OTU previously generated from a de novo assembly of 546 vaginal microbiomes (Albert et al. 2015) and labeled according to nearest neighbour in cpnDB_nr. The mock community (MVP) was also mapped on to the 20 *cpn*60 UT sequences that it was composed of. A manually curated reference dataset was not available for infant stool, so cpnDB_nr was used. Following Bowtie2 mapping, a feature table containing OTU names and read abundances was created as previously described (Freitas et al. 2018).

Microbial community composition analysis

Alpha diversity metrics Shannon diversity and Chao1 estimated species richness were calculated for each sample as mean values over 100 subsamples with a sequence depth of 1000 for vaginal samples, and 20000 for infant stool samples, or all reads when fewer were available, using QIIME2 (Bolyen et al. 2019). Alpha rarefaction plots were visualized to ensure adequate sampling depth was achieved. Results are reported as mean values over all samples.

To determine consistency of the microbiome profiles resulting from each OTU generation method in terms of species presence and abundance, clustering was conducted with a Jensen-Shannon distance matrix in R using a custom distance function that calculates the square root of the Jensen-Shannon divergence (Endres and Schindelin 2003). Hierarchical clustering was conducted using the hclust function in R with the Ward 2 linkage method.

RESULTS AND DISCUSSION

Optimal length for variant calling

Variant calling using the DADA2 denoising algorithm in QIIME2 employs a truncation length parameter at which all input sequence reads are trimmed to uniformly. If a read is under the length then it is not included in the process (Callahan et al. 2016). It was therefore desirable to determine a uniform length to use with the *cpn*60 UT sequences, given input reads of up to 400 bp and the trade-off between minimum length requirement, and proportion of reads retained in variant calling. The vaginal libraries (n = 25) were used to investigate optimal variant length. DADA2 was run using seven different lengths from 25 to 300 bp, and the resulting profiles and statistics were used to determine at which length unambiguous species level identification of OTU sequences could be achieved. To evaluate the quality of the OTU generated from variant calling at each length, the variant sequences formed from each sample were pooled from each length and were aligned to cpnDB_nr with watered-BLAST returning all significant hits (Schellenberg et al. 2009). An OTU sequence was considered unambiguously identified if there was only one resulting match to the database over the 55% identity threshold that did not tie with any other results. By 100 bp 88% of OTU returned a single best match when aligned to the database, and only improved slightly to 93.5% at 300 bp (Table 1). Given the protocol with which the samples were sequenced, all reads start from the 5´ end of the *cpn*60 UT. Any OTUs aligned in the reverse direction are therefore considered nonsense and cannot be interpreted for species identification. By 150 bp, 99% of the OTU were aligned in the correct direction (Table 1). Although the proportion of OTU alignments in the correct direction increased between 50 and 300 bp, there was a corresponding decrease in the number of nearest neighbours identified. This was likely due to decreasing numbers of reads available for variant calling as the length requirement increased. Interestingly, there was a spike of alignments in the correct direction

when variant calling was run at a length of 25 bp and the 89 OTU produced at this length corresponded to only 79 nearest neighbours. Since so few OTU formed at this length it can be concluded that this region of the *cpn*60 UT must be very similar among species. The alignments were likely in the correct direction at 25 bp as a reflection of a quality of the beginning 5´ region of the *cpn*60 UT.

Overall, the quality of the OTU formed increased with increasing length. Lengths 25 and 50 bp were shown to be insufficient as the OTU formed were poorly identified. Using a truncation length of 300 bp, however, would not be advisable since there is a corresponding loss of data. After quality trimming of reads, 173321 reads were available for input to DADA2, but only 53031 reads were retained when the length requirement was increased to 300 bp (Table 1). The effect of this loss is that the overall sequencing depth of the samples is reduced, and thus the chance of detecting rare organisms, which likely outweighs the benefit of the small increase in resolution that would be achieved using greater lengths. Shorter read lengths could also decrease the cost of sequencing.

To further investigate the optimal truncation length for cpn60 variant calling with DADA2 and remove the confounding effect of diminishing read numbers retained as the truncation length requirement increases, we generated a set of input reads from the infant stool data set that were all at least 300 in length after quality filtering. This was possible since there were 2594209 raw reads available. The Trimmomatic quality filtering step was performed so that only reads above the 300 bp threshold with a quality of 30 over a 4 bp sliding window. As expected, we found that without loss of reads, there were more variants (OTU) produced with greater sequence truncation lengths (Table 2). This was expected given the way in which DADA2 determines variants; longer lengths mean the algorithm has more chances to detect

differences. We also found that for all the truncation lengths used, all of the OTU sequences aligned to a cpnDB record in the forward direction. Similar to what was found with the vaginal data (Table 1), the proportion of OTU unambiguously identified reached its maximum of ~90% at a truncation length of 150 bp (Table 2), providing further evidence that 150 bp from the 5´ end of the cpn60 barcode sequence is sufficient for microbiome profiling with species-level resolution. This conclusion is consistent with the known features of the cpn60 barcode sequence (Links et al. 2012a). We also examined the pairwise percent identities calculated from multiple sequence alignments of bacterial sequences from cpnDB_nr, trimmed to lengths ranging from 50 bp to >500 (full length) from the 5´ end (Figure S1). The mean and median percent identities of the pairwise comparisons for alignments of 150 and 200 bp were significantly lower than for longer (or shorter) regions of the cpn60 barcode. This finding is consistent with that of a previous investigation demonstrating relatively higher inter-species sequence differences in the 5´ end of the cpn60 barcode sequence relative to the 3' end (Links et al. 2012a).

Considering all the trade-offs of data retention, cost of sequencing, and species resolution, a DADA2 truncation length of 150 bp was used for the subsequent comparison of methods of OTU generation.

Accuracy of OTU generation methods from synthetic community sequence data

The MVP synthetic community was used to determine accuracy the three methods of OTU generation. Following amplification primer removal and quality trimming, 54702 and 40128 reads were available for analysis from the duplicate sequencing runs. The number of OTU produced from de novo assembly, variant calling (truncation length 150), and reference mapping to either the VOGUE reference assembly or MVP sequences was determined (Figure 1).

Reference mapping directly to the MVP gave the expected result of 20 OTU in both library duplicates, confirming that all constituents of the mock community were amplified and sequenced. When the MVP OTU were aligned to cpnDB_nr using watered-BLAST, they aligned to only 19 different nearest neighbours. This difference was not unexpected since two of the cloned cpn60 barcode sequences in the MVP (corresponding to *Lactobacillus gasseri* and *L. johnsonii*) are 97% identical. Reference mapping to the VOGUE reference assembly was included in the comparison to determine the effect that more diverse mapping database could have on the results and we found that far more OTU and NN were generated from the MVP sequence libraries using this approach (Figure 1). Further analysis of the profiles showed that many of the 'extra' OTU that were produced were the result of reads mapping to highly similar sequences in the reference database. This result illustrates a limitation of the reference mapping approach, which is dependent on the Bowtie2 algorithm to determine the best match for each read. Bowtie2 alignments close in score may be chosen over one anther unpredictably, or if the program's search limit has been reached it may not determine a better alignment if one exists, which may lead to changes in resulting profiles (Langmead and Salzberg 2012). Variant calling and de novo assembly of MVP derived sequence reads produced more OTU than the 20 expected; up to 29 in the second replicate analysed by variant calling. Additional OTU were most likely formed as a result of minor sequence difference due to PCR errors in the barcode sequence amplification, which was not done with a proof-reading polymerase. Support for this explanation is also found in the observation that the 'extra' OTU corresponded to the expected number of nearest neighbours in the second replicate.

Taken together, the results of analysis of the MVP synthetic community sequence data demonstrates that while variant calling and de novo assembly fairly accurately produced the

expected community composition, results of reference mapping were heavily influenced by the qualities of the database used and could lead to a significant over-estimation of the number of species in the community.

Effects of OTU generation method on overall microbial community profiles

Microbial community profiles derived from barcode sequence amplification and sequencing include both an inventory of the distinct OTU present as well as their proportional abundance based on sequence read numbers. Various statistical approaches and clustering methods can then be applied to investigate beta-diversity within collections of samples with the goal of identifying similarities and differences among samples that may be associated with defined groups in a study, for example pre- and post-treatment groups, or subjects with different clinical outcomes or phenotypes. To determine the effects of the different OTU formation, identification and frequency calculation methods on overall microbiome profiles, we exploited the vaginal microbiome sequence data, and a previously generated data set derived from infant stool. Results of de novo assembly, reference mapping, and variant calling (using a 150 bp truncation length) were compared to determine if the methods produce different results in terms of outcome microbiome profiles.

The number of OTU formed and the number of nearest neighbour species from the three methods are indicated in Table 3 for all sample types. For vaginal samples de novo assembly, reference mapping to the VOGUE reference assembly, and variant calling producing 126, 183 and 191 OTU, respectively, reflecting the nature of the methods. The variant calling program is designed to separate variants or sequences by small differences that could indicate different strains (Callahan et al. 2017), whereas the de novo method is assembling reads into OTU by

similarity and can have trouble resolving very close sequences (Links et al. 2013b). The overestimation of community richness resulting from the reference mapping approach that we observed in the synthetic community sequencing experiment was also observed here. Reference mapping approach resulted in 183 OTU, which corresponded to 135 nearest neighbour species, notably more than identified for de novo assembly (103) and variant calling (109).

Stool samples were chosen for comparison because the fecal microbiome is known to be more diverse than the vaginal microbiome (Costello et al. 2009). When reads from 47 samples were combined, 213 OTU were generated from de novo assembly, 318 from reference mapping to cpnDB_nr, and 422 from variant calling (Table 3). Similar to what was observed for vaginal samples, variant calling produced the greatest number of *cpn*60 OTU sequences, a reflection of the resolution achievable with this technique. It was again observed that when the OTU were aggregated to nearest neighbour species, de novo assembly and variant calling produced similar numbers (130 and 160, respectively), while reference mapping produced approximately twice as many (318). Unlike the vaginal microbiome, a reference set of curated *cpn*60 sequences were not available for the infant stool microbiome, so cpnDB_nr was the target for reference mapping. cpnDB_nr is, by definition, a non-redundant database and so no collapsing to NN species occurred. This example further highlights the limitations of reference-mapping, especially when a custom database is not available for the environment under study.

The resulting microbiome profiles of each sample from each method were compared to determine if the OTU formation method affected the overall microbiome profile. Dendrograms were created for each sample type using Jensen-Shannon distance calculation and a clustering algorithm to determine relationships of the sample profiles resulting from each method of OTU generation (Figure 2). For the vaginal microbiome samples, the three profiles for each sample

clustered together regardless of OTU generation method (Figure 2A). Only samples 6 and 7 did not cluster immediately together, but upon further investigation these samples had profiles comprised almost entirely of *Lactobacillus crispatus*, so small differences in the other, low abundance sequences accounted for intermixing of profiles.

When the infant samples were initially clustered, many samples had different resulting profiles over the three methods. Upon further investigation this was found to be due to very similar OTU sequences aligning to either *Escherichia coli* or *Shigella sonnei*. Despite being in different genera, these organisms are difficult to differentiate and it has been argued that they are in fact the same species (Lan and Reeves 2002). The *cpn*60 barcode sequence from these species are 99.3% similar, so the frequencies from these were merged to *Escherichia coli* for further analysis. Cluster analysis was done with the resulting profiles from fifteen randomly selected samples (Figure 2B). In most cases, profiles clustered by sample, regardless of method of OTU generation. Minor differences in the microbiome profiles for samples 4, 25 and 36 were due to differences in classification of some OTU sequences with similarity to *Bifidobacterium* spp., which were not unexpected given the demonstrated differences in resolution of the OTU generation methods used (Table 3).

Alpha diversity statistics (Chao1 and Shannon) were calculated for each method and sample type (Table 4). There was no difference in the average richness or diversity for the vaginal samples (Kruskal-Wallis, $p > 0.05$). No difference in Shannon diversity was detected between methods for the infant stool samples (Kruskal-Wallis, $p > 0.05$), but there was a difference in Chao1 richness (Kruskal-Wallis, $p < 0.001$) with the reference mapping to cpnDB_nr producing the richest microbiome profile, which was not surprising since Chao1 is

influenced by rare OTUs, and these were abundant in the stool microbiome data, likely due to the idiosyncrasies of Bowtie2 described above.

Practical considerations and recommendations

All three approaches use open source software so there is no cost factor to consider. The real time for all processes in each OTU formation method were summed for each sample type, conducted on a Dell PowerEdge R910 with 250 GB RAM running CentOS Linux 7 (Figure 3). In each case, variant calling was the fastest, but comparable to de novo assembly. Perhaps surprisingly, reference mapping was the most time-consuming method, and this difference was especially apparent when a larger and less specific reference dataset was used in the case of the infant stool analysis where the cpnDB_nr was used in place of a custom, environment-specific database.

The reference mapping approach can be efficient with a custom reference dataset, but its resolving power depends entirely on the quality of the reference dataset to which reads are mapped, and reference sequence datasets are not available for all bacterial communities. The de novo assembly method has been used extensively for *cpn*60-based microbiome profiling, but is slower, and requires more dependencies than variant calling. Variant calling was the fastest method we tested, and resulted in high resolution microbiome profiles compared to the other method with truncation lengths of at least 150 bp. The use of denoising methods for variant calling has also been suggested for detection of otherwise unobserved 16S rRNA sequence variants (Callahan et al. 2017). This approach is particularly attractive for users with limited experience with data analysis, since it can be easily installed and accessed with the QIIME2 package (Bolyen et al. 2019).

The choice of method for sequence classification from barcode sequence-based studies of microbial communities is influenced by a number of factors including pre-existing knowledge of the environment under study, numbers of samples, sequencing depth, desired level of taxonomic resolution, computational resources, and the need to pool results from independent studies. Exact sequence variant calling with DADA2 using 150 bp from the 5´ end of the cpn60 barcode sequencing is a recommended strategy for rapid, relatively simple to execute, high-resolution profiling of microbiomes that facilitates the comparison of microbiome profiles across studies.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Chelsea Elwood, Scott Dos Santos and the LEGACY team for access to prepublication sequence data.

REFERENCES

Albert, A.Y., Chaban, B., Wagner, E.C., Schellenberg, J.J., Links, M.G., van Schalkwyk, J., Reid, G., Hemmingsen, S.M., Hill, J.E., and Money, D. 2015. A study of the vaginal microbiome in healthy Canadian women utilizing cpn60-based molecular profiling reveals distinct *Gardnerella* subgroup community state types. PLoS One **10**(8): e0135620. doi:10.1371/journal.pone.0135620.

Bolger, A.M., Lohse, M., and Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**(15): 2114–20. doi:10.1093/bioinformatics/btu170.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodríguez, A.M., Chase, J., Cope, E.K., Silva, R.D., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall,

D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K.B., Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., Hooft, J.J.J. van der, Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., Hippel, M. von, Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R., and Caporaso, J.G. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol **37**(8): 852–857. doi:10.1038/s41587-019-0209-9.

Bondici, V.F., Lawrence, J.R., Khan, N.H., Hill, J.E., Yergeau, E., Wolfaardt, G.M., Warner, J., and Korber, D.R. 2013. Microbial communities in low permeability, high pH uranium mine tailings: characterization and potential effects. J Appl Microbiol **114**(6): 1671–1686. doi:10.1111/jam.12180.

Callahan, B.J., McMurdie, P.J., and Holmes, S.P. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal **11**(12): 2639–2643. doi:10.1038/ismej.2017.119.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods **13**(7): 581–583. doi:10.1038/nmeth.3869.

Chaban, B., Albert, A., Links, M.G., Gardy, J., Tang, P., and Hill, J.E. 2013. Characterization of the upper respiratory tract microbiomes of patients with pandemic H1N1 influenza. PLoS ONE **8**(7): e69559. doi:10.1371/journal.pone.0069559.

Chaban, B., Links, M.G., Jayaprakash, T., Wagner, E.C., Bourque, D.K., Lohn, Z., Albert, A.Y., van Schalkwyk, J., Reid, G., Hemmingsen, S.M., Hill, J.E., and Money, D.M. 2014. Characterization of the vaginal microbiota of healthy Canadian women through the menstrual cycle. Microbiome **2**(1): 23. doi:10.1186/2049-2618-2-23.

Costa, M.O., Chaban, B., Harding, J.C.S., and Hill, J.E. 2014. Characterization of the fecal microbiota of pigs before and after inoculation with "*Brachyspira hampsonii*." PLoS ONE **9**(8): e106399.

Costello, E.K., Lauber, C.L., Hamady, M., Fierer, N., Gordon, J.I., and Knight, R. 2009. Bacterial community variation in human body habitats across space and time. Science **326**(5960): 1694–7. doi:1177486 [pii] 10.1126/science.1177486.

Desai, A.R., Musil, K.M., Carr, A.P., and Hill, J.E. 2009. Characterization and quantification of feline fecal microbiota using cpn60 sequence-based methods and investigation of animal-to-animal variation in microbial population structure. Vet Microbiol **137**: 120–128.

Dumonceaux, T.J., Hill, J.E., Hemmingsen, S.M., and Van Kessel, A.G. 2006a. Characterization of intestinal microbiota and response to dietary virginiamycin supplementation in the broiler chicken. Applied and Environmental Microbiology **72**(4): 2815–2823.

Dumonceaux, T.J., Hill, J.E., Pelletier, C., Paice, M.G., Van Kessel, A.G., and Hemmingsen, S.M. 2006b. Molecular characterization of microbial communities in Canadian pulp and paper activated sludge and quantification of a novel *Thiothrix eikelboomii*-like bulking filament. Canadian Journal of Microbiology **52**(5): 494–500.

Dumonceaux, T.J., Schellenberg, J., Goleski, V., Hill, J.E., Jaoko, W., Kimani, J., Money, D., Ball, T.B., Plummer, F.A., and Severini, A. 2009. Multiplex detection of bacteria associated with normal microbiota and with bacterial vaginosis in vaginal swabs using oligonucleotide-coupled fluorescent microspheres. Journal of Clinical Microbiology **47**(12): 4067–4077. doi:10.1128/jcm.00112-09.

Endres, D.M., and Schindelin, J.E. 2003. A new metric for probability distributions. IEEE Trans. Inf. Theory **49**(7): 1858–1860.

Freitas, A.C., Bocking, A., Hill, J.E., Money, D.M., Money, D., Bocking, A., Hemmingsen, S., Hill, J., Reid, G., Dumonceaux, T., Gloor, G., Links, M., O'Doherty, K., Tang, P., van Schalkwyk, J., Yudin, M., and the VOGUE Research Group. 2018. Increased richness and diversity of the vaginal microbiota and spontaneous preterm birth. Microbiome **6**(1): 117. doi:10.1186/s40168-018-0502-8.

Freitas, A.C., Chaban, B., Bocking, A., Rocco, M., Yang, S., Hill, J.E., and Money, D.M. 2017. The vaginal microbiome of healthy pregnant women is less rich and diverse with lower prevalence of Mollicutes compared to healthy non-pregnant women. Scientific Reports **7**: 9212. doi:10.1038/s41598-017-07790-9.

Gohl, D.M., Vangay, P., Garbe, J., MacLean, A., Hauge, A., Becker, A., Gould, T.J., Clayton, J.B., Johnson, T.J., Hunter, R., Knights, D., and Beckman, K.B. 2016. Systematic

improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nature Biotechnology **34**(9): 942–949. doi:10.1038/nbt.3601.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology **29**(7): 644–52. doi:10.1038/nbt.1883.

Hebert, P.D.N., Cywinska, A., Ball, S.L., and deWaard, J.R. 2003. Biological identifications through DNA barcodes. Proceedings of the Royal Society B: Biological Sciences **270**(1512): 313–321. doi:10.1098/rspb.2002.2218.

Hill, J.E., Penny, S.L., Crowell, K.G., Goh, S.H., and Hemmingsen, S.M. 2004. cpnDB: a chaperonin sequence database. Genome Research **14**(8): 1669–75.

Hill, J.E., Seipp, R.P., Betts, M., Hawkins, L., Van Kessel, A.G., Crosby, W.L., and Hemmingsen, S.M. 2002. Extensive profiling of a complex microbial community by high-throughput sequencing. Applied and Environmental Microbiology **68**(6): 3055–66.

Hill, J.E., Town, J.R., and Hemmingsen, S.M. 2006. Improved template representation in *cpn*60 PCR product libraries generated from complex templates by application of a specific mixture of PCR primers. Environmental Microbiology **8**(4): 741–746. doi:doi:10.1111.

Johnson, L.A., Chaban, B., Harding, J.C., and Hill, J.E. 2015. Optimizing a PCR protocol for cpn60-based microbiome profiling of samples variously contaminated with host genomic DNA. BMC Research Notes **8**: 253. doi:10.1186/s13104-015-1170-4.

Lan, R., and Reeves, P.R. 2002. *Escherichia coli* in disguise: molecular origins of *Shigella*. Microbes and Infection **4**(11): 1125–1132. doi:10.1016/S1286-4579(02)01637-4.

Langmead, B., and Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods **9**(4): 357–9. doi:10.1038/nmeth.1923.

Links, M.G., Chaban, B., Hemmingsen, S.M., Muirhead, K., and Hill, J.E. 2013a. mPUMA: a computational approach to microbiota analysis by de novo assembly of operational taxonomic units based on protein-coding barcode sequences. Microbiome **1**(1): 23. doi:10.1186/2049-2618-1-23.

Links, M.G., Chaban, B., Hemmingsen, S.M., Muirhead, K., and Hill, J.E. 2013b. mPUMA: a computational approach to microbiota analysis by *de novo* assembly of OTUs based on protein-coding barcode sequences. Microbiome **1**: 23. doi:10.1186/2049-2618-1-23.

Links, M.G., Demeke, T., Gräfenhan, T., Hill, J.E., Hemmingsen, S.M., and Dumonceaux, T.J. 2014. Simultaneous profiling of seed-associated bacteria and fungi reveals antagonistic interactions between microorganisms within a shared epiphytic microbiome on *Triticum* and *Brassica* seeds. New Phytologist **202**(2): 542–553.

Links, M.G., Dumonceaux, T.J., Hemmingsen, S.M., and Hill, J.E. 2012a. The chaperonin-60 universal target is a barcode for bacteria that enables *de novo* assembly of metagenomic sequence data. PLoS ONE **7**(11): e49755. doi:10.1371/journal.pone.0049755.

Links, M.G., Dumonceaux, T.J., Hemmingsen, S.M., and Hill, J.E. 2012b. The Chaperonin-60 Universal Target Is a Barcode for Bacteria That Enables De Novo Assembly of Metagenomic Sequence Data. PLoS ONE **7**(11): e49755. doi:10.1371/journal.pone.0049755.

Mansfield, G.S., Desai, A.R., Nilson, S.A., Van Kessel, A.G., Drew, M.D., and Hill, J.E. 2010. Characterization of rainbow trout (*Oncorhynchus mykiss*) intestinal microbiota and

inflammatory marker gene expression in a recirculating aquaculture system. Aquaculture **307**(1–2): 95–104. doi:Doi 10.1016/J.Aquaculture.2010.07.014.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal **17**(1): 10–12. doi:10.14806/ej.17.1.200.

Schellenberg, J., Links, M.G., Hill, J.E., Dumonceaux, T.J., Peters, G.A., Tyler, S., Ball, B., Severini, A., and Plummer, F.A. 2009. Pyrosequencing of the chaperonin-60 universal target as a tool for determining microbial community composition. Applied and Environmental Microbiology **75**(9): 2889–2898.

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., and Chen, W. 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc Natl Acad Sci U S A **109**(16): 6241–6. doi:10.1073/pnas.1117018109.

Vancuren, S.J., and Hill, J.E. 2019. Update on cpnDB: a reference database of chaperonin sequences. Database (Oxford) **2019**. doi:10.1093/database/baz033.

Table 1: Variant calling results from vaginal microbiome sequence data (295868 raw reads)

Method	Number of filtered reads available for DADA2	Number of OTU	Number of NN	% OTU single hit to cpnDB_nr	% NN alignments in the correct direction
Variant calling 25 bp	249197	89	79	59.3	95.5
Variant calling 50 bp	218971	336	254	68.3	73.8
Variant calling 100 bp	201996	273	167	88	93.4
Variant calling 150 bp	173321	239	129	90.2	99.6
Variant calling 200 bp	141663	222	115	92.9	100
Variant calling 250 bp	100699	185	98	91.9	100
Variant calling 300 bp	53031	107	55	93.5	100

Table 2: Variant calling results with consistent input read numbers (n=186359 reads).

Method	Number of OTU	Number of NN	% OTU single hit to cpnDB
Variant Calling 50 bp	156	96	70.25
Variant Calling 100 bp	187	97	79.14
Variant Calling 150 bp	205	100	89.27
Variant Calling 200 bp	219	101	88.58
Variant Calling 250 bp	221	98	90.95
Variant Calling 300 bp	230	99	91.74

Table 3: Numbers of OTU and Nearest Neighbours resulting from de novo assembly, reference mapping and variant calling.

Sample Type	Number of Reads Post Quality Trimming	Average Number of Reads Per Sample	Method	Number of OTU	Number of NN
Vaginal (n=25)	91692	3056	De Novo	126	103
			Reference Mapping (to VOGUE assembly)	183	135
			Variant Calling	191	109
Infant Stool (n=47)	1326909	28232	De Novo	213	130
			Reference Mapping (to cpnDB_nr)	318	318
			Variant Calling	422	160

Table 4: Alpha diversity statistics for microbiome profiles resulting from de novo assembly, reference mapping or variant calling

Sample Type	Method	Shannon diversity	Standard deviation	Chao1 index	Standard deviation
Vaginal (n=25)	De novo	1.21	1.16	14.22	12.50
	Reference mapping	1.45	0.91	17.14	13.85
	Variant calling	1.21	1.16	10.94	9.98
Infant Stool (n=47)	De novo	1.34	0.79	29.45	9.77
	Reference mapping	1.57	0.82	41.86	15.31
	Variant calling	1.29	0.76	17.12	6.05

Figure Captions

Figure 1: Number of operational taxonomic units (OTU) and nearest neighbour species (NN) from each method applied to the MVP mock community of 20 vaginal bacteria. The synthetic community was sequenced in duplicate in separate runs. For each pair of duplicates, the lighter bar represents the first run and the darker bar the replicate. Methods: 1 = de novo assembly, 2 = reference mapping to MVP, 3 = reference mapping to VOGUE reference assembly, 4 = variant calling.

Figure 2: Comparison of microbiome profiles resulting from three methods of OTU generation. Clustering of profiles from vaginal samples (A) and 15 randomly selected infant stool samples (B) based on nearest neighbour species abundance with Jensen-Shannon distance calculation and hclust in R. Cases where the sample profiles from all methods have grouped together are indicated with a box. RM = reference mapping, DN = de novo assembly, VC = variant calling.

Figure 3: Timing for each method of operational taxonomic unit generation for each sample type. Real time execution using the Linux 'time' function was calculated and summed for every procedure required in processing microbiome profiles for each method for all samples in every sample type.

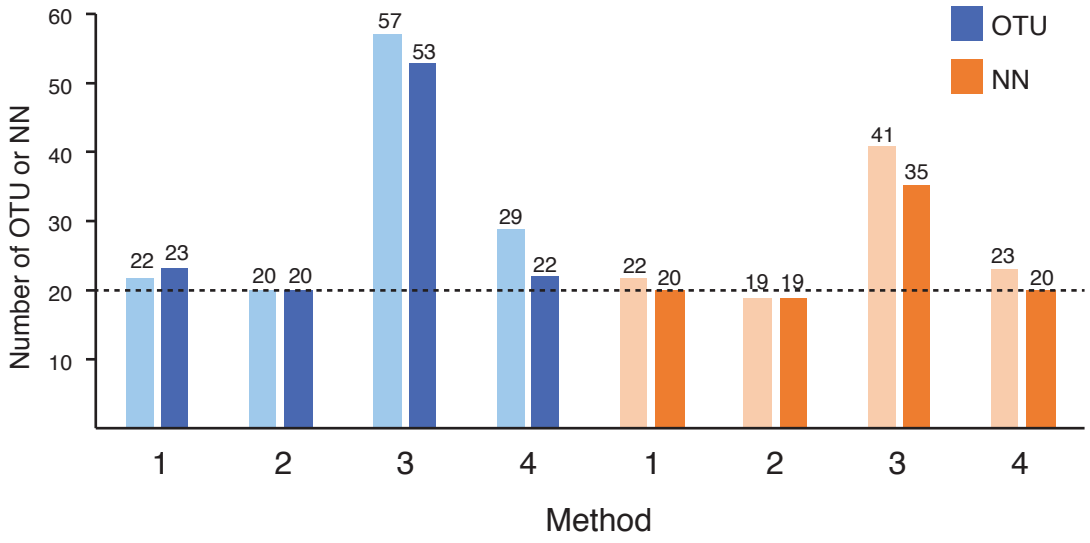

Figure 1: Number of operational taxonomic units (OTU) and nearest neighbour species (NN) from each method applied to the MVP mock community of 20 vaginal bacteria. The synthetic community was sequenced in duplicate in separate runs. For each pair of duplicates, the lighter bar represents the first run and the darker bar the replicate. Methods: 1 = de novo assembly, 2 = reference mapping to MVP, 3 = reference mapping to VOGUE reference assembly, 4 = variant calling.

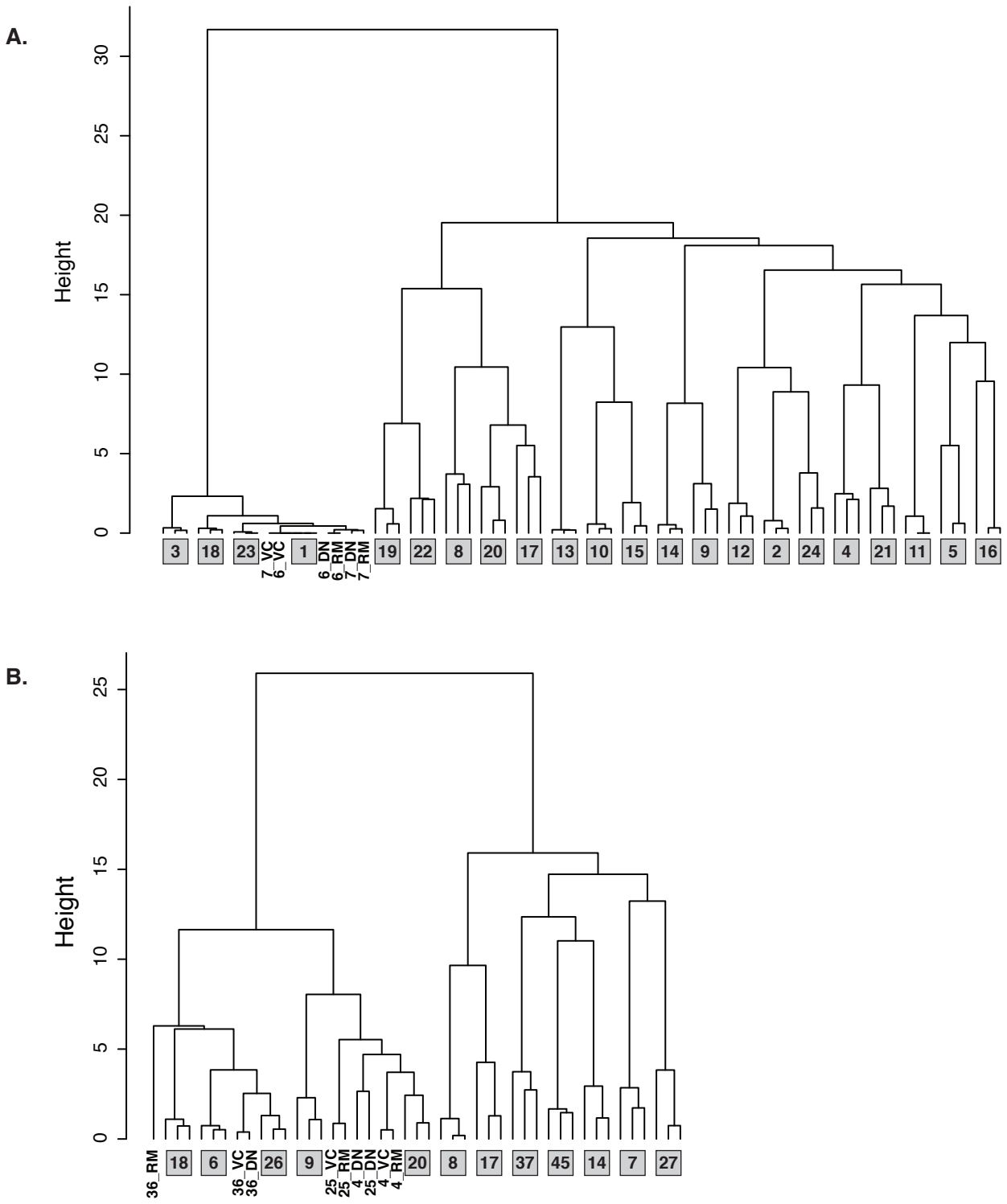

Figure 2: Comparison of microbiome profiles resulting from three methods of OTU generation. Clustering of profiles from vaginal samples (A) and 15 randomly selected infant stool samples (B) based on nearest neighbour species abundance with Jensen-Shannon distance calculation and hclust in R. Cases where the sample profiles from all methods have grouped together are indicated with a box. RM = reference mapping, DN = de novo assembly, VC = variant calling.

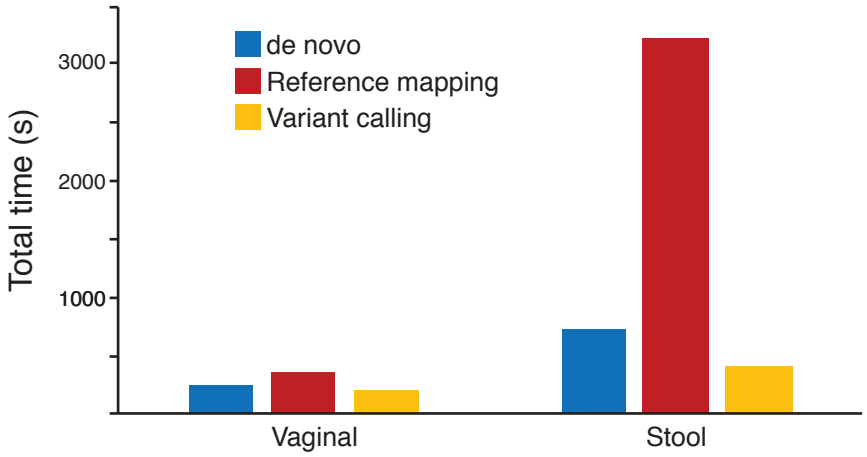

Figure 3: Timing for each method of operational taxonomic unit generation for each sample type. Real time execution using the Linux 'time' function was calculated and summed for every procedure required in processing microbiome profiles for each method for all samples in every sample type.