NanoAmpli-Seq: A workflow for amplicon sequencing from mixed microbial communities on the nanopore sequencing platform.

Szymon T Calus[1], Umer Z Ijaz[1*], Ameet Pinto[2*]

Affiliations:

[1]School of Engineering, University of Glasgow, Glasgow, G12 8QQ, United Kingdom

[2]Department of Civil & Environmental Engineering, Northeastern University, 269 SN, 360 Huntington Avenue, Boston, MA, USA

Keywords: Amplicon, 16S rRNA, Consensus, Sequencing, Next Generation, Nanopore

* Correspondence to:

Ameet Pinto, Department of Civil & Environmental Engineering, Northeastern University. a.pinto@northeastern.edu

Umer Ijaz, School of Engineering, University of Glasgow, UK. umer.ijaz@glasgow.ac.uk

Abstract

Small subunit (SSU) rRNA gene sequencing on second generation platforms leverages their deep sequencing and multiplexing capacity, but is limited in genetic resolution due to short read lengths. While third generation platforms overcome this limitation, their application has been limited due to high error rates. In this study, we introduce an amplicon sequencing workflow, i.e., NanoAmpli-Seq, that builds on Intramolecular-ligated Nanopore Consensus Sequencing (INC-Seq) approach, for full-length SSU rRNA gene sequencing. NanoAmpli-Seq includes key improvements to INC-Seq that reduces sample processing time while significantly improving sequence accuracy. NanoAmpli-Seq adds chopSeq for correction of INC-Seq consensus reads and nanoClust for read partitioning-based *de novo* clustering and within cluster consensus calling to obtain full-length 16S rRNA gene sequences. NanoAmpli-Seq accurately estimates diversity of tested mock communities with average sequence accuracy of 99.5% for 2D and 1D2 sequencing on the nanopore sequencing platform. Residual errors in NanoAmpli-Seq sequences originate from deletions in homopolymers, indicating that homopolymer aware basecalling or error correction may allow for sequence accuracy nearing 100%.

Introduction

Amplicon sequencing, particularly sequencing of the small subunit rRNA (SSU rRNA) gene and internal transcribed spacer (ITS) regions, is widely used for profiling of microbial community structure and membership[1-4]. The wide scale application of amplicon sequencing has been largely driven by the ability to multiplex 100's of samples on a single sequencing run and obtain millions of sequences of target communities on second generation sequencing platforms (i.e., Illumina, Ion Torrent, etc)[1,5]. The primary limitation of second generation sequencing platforms is their short read lengths which ranges from 150-400 bp[6]. While excellent at bulk profiling of microbial communities through multiplexed deep sequencing, short read lengths are limited in the taxonomic resolution of sequenced reads and more so, are not amenable to robust phylogenetic analyses to assess the relationship between sequences originating from unknown microbes with those in publicly available databases. An important effect of the proliferation in short read sequencing applications has been a decrease in the rate at which long higher quality sequences, particularly for SSU rRNA genes, are being deposited in public databases. This effect is to some extent being mitigated assembly and curation of near full length SSU rRNA genes from metagenomic datasets[7,8] and will continue to be mitigated with novel approaches for SSU rRNA sequencing with use of synthetic long read approaches[9].

The introduction of third generation sequencing platforms, such as PacBio's single molecule real-time sequencing (SMRT) and single molecule sensing technologies on the Oxford Nanopore Technologies (ONT) MinION™ platform, have opened the possibility of obtaining ultra-long reads[6,10]. While sequencing throughput and raw data quality of long read third generation sequencing platforms is yet to rival that of short read second generation platforms, the ability to

obtain ultra-long reads can overcome several limitations of the latter[11]. For instance, long-read sequencing combined with various error correction approaches[12,13] have been used to obtain high quality single contig microbial genomes[13,14] or increase assembly quality of previously sequenced but fragmented eukaryotic genomes assemblies[15,16], which was not feasible using a short read only approach. Long read sequencing capabilities have also been recently leveraged to sequence near full-length SSU rRNA genes (e.g., 16S rRNA)[17-20] or even the entire *rrn* operon[21]. Majority of the studies utilizing either the SMRT or nanopore sequencing platforms have limited data analyses to sequence classification, demonstrating that widely used sequence classifiers can tolerate high sequencing error rates. Only a few studies have provided insights into factors influencing the amplicon sequencing error rates and provided sample preparation and/or post processing strategies to overcome these high sequencing error rates. Notable among these efforts has been the use of (1) SMRT sequencing of near full length 16S rRNA gene accompanied by quality filtering and clustering approaches to reduce error rates below 0.1%[17] and (2) the development of the Intramolecular-ligated Nanopore Consensus Sequencing (INC-Seq) protocol for consensus based error correction of nanopore sequencing reads with a median accuracy of 97-98%[18].

Schloss et al[17] utilized circular consensus sequencing on the PacBio platform coupled with a range of quality filtering (i.e., mismatches to primer, quality scores) and sequence clustering (i.e., pre-cluster) to significantly enhance data quality for full length 16S rRNA gene sequencing. Furthermore, the number of operational taxonomic units (OTUs) of the processed data was highly similar to the theoretical number of OTUs in tested mock communities, indicating that the application of this protocol for naturally derived mixed microbial communities is likely to result in robust diversity estimates. The INC-Seq workflow involves amplicon concatermization to

generate multiple identical copies of the same amplicon on a single DNA molecule, followed by sequencing of the concatemer pool using 2D sequencing chemistry on the nanopore sequencing platform, and consensus based error correction after aligning the physically linked concatemers on each sequenced DNA strand. By using this approach, Li et al[18] were able to increase the median sequence accuracy to 97-98%. While this significant improvement allows for taxonomic classification of sequences to the species level, it does not allow for sequence clustering for diversity estimation due to residual median error rates of approximately 2-3%.

In this study, we leverage and expand on the INC-Seq protocol developed by Li et al[18] to provide a complete workflow for amplicon sequencing and *de novo* data processing, NanoAmpli-Seq (v1.0), applied to near full length 16S rRNA gene of mock communities that results in high quality sequences with a mean sequence accuracy of 99.5±0.08%. The current version of NanoAmpli-Seq includes modifications to the library preparation protocol for INC-Seq and fixes a key issue with INC-Seq consensus sequences while adding a novel read partitioning-based sequence clustering approach which results in accurate estimation of diversity of mixed microbial communities and allows higher sequence accuracy by allowing within OTU pairwise sequence alignment and consensus calling. Further, we demonstrate that NanoAmpli-Seq works equally well on the (now deprecated) 2D sequencing chemistry and the recently released 1D2 sequencing chemistry on the MinION™ device. While important limitations such as suboptimal re-construction of community structure and error rate of 0.5% (5-10 fold higher than has been reported by Schloss et al[17] using SMRT sequencing) remain, the proposed approach may be used for sequencing of long amplicons from complex microbial communities to assess community

membership with cautious utilization of sequences from low abundance OTUs due to likely lower sequence accuracy ranging from 99-99.5% accuracy.

Results

Experimental design and workflow. The NanoAmpli-Seq protocol was developed and validated using amplicon pools consisting of near full-length 16S rRNA gene of a single organism (*Listeria monocytogens*) or an equimolar amplicon pool of near full-length 16S rRNA genes from 10 organisms (Supplementary Table S1). The amplicon pools were generated by PCR amplifying near full length 16S rRNA genes from genomic DNA of target organism(s) using primers and PCR reaction conditions as described in the materials and methods section. The respective amplicon pools were subsequently prepared for sequencing using the INC-Seq workflow as outlined in Figure 1, with a few important modifications. Briefly, the amplicon pools were self-ligated to form plasmid-like structures which was followed by digestion with plasmid-safe DNAse to remove the remaining non-ligated linear amplicons. The DNA pool consisting of plasmid-like structures was subject to rolling circle amplification (RCA) using random hexamer-free protocol using a combination of primase/polymerase (PrimPol) and hi-fidelity Phi29 DNA polymerase. The RCA product was subject to two rounds of T7 endonuclease debranching and g-TUBE fragmentation followed by gap filling and DNA damage repair. A detailed protocol description include reagent volumes and incubation conditions is provided in the materials and methods section. The prepared amplicon pools for both single organism and 10 organism mock community samples were then subject to library preparation using the standard 2D (SQK-LSK208) (Runs 1 and 2) and 1D2 (SQK-LSK308) (Runs 3 and 4) kits using ONT specifications and sequenced on the MinION™ MK1b device followed by basecalling using Albacore 1.2.4.

Each resulting read consisted of multiple concatamerized 16S rRNA gene amplicons. The reads were subject to INC-Seq based concatemer alignment and consensus error correction using three different alignment options (i.e., blastn, Graphmap, and partial order alignment (POA)) and followed by iteratively running PBDAGCON on the consensus for error correction (INC-Seq flag "iterative"). Manual inspection of INC-Seq corrected reads revealed a vast majority of sequences with incorrect orientation of primers (Figure 2A). Specifically, the forward and reverse primers did not occur at the ends of the INC-Seq consensus reads, but rather were co-located at varying positions along the length of the read. Efforts to manually split INC-Seq reads and re-orient the forward and reverse splits based on primer orientation revealed the presence of tandem repeats (ranging from 2-10 repeats) of nearly identical sequences, which affected efforts to merge the forward and reverse read splits (Figure 2B, 2C). To this end, we developed *chopSeq* algorithm as part of the NanoAmpli-Seq workflow which (1) identifies user provided primer sequences in the read, (2) splits the read at the end of the primers, (3) re-orients the forward and reverse splits, (4) identifies and removes tandem repeats by iteratively measuring the sequence similarity between co-occurring segments using window size ranging from 10 bp to 350 bp with diminishing sequence similarity threshold with increasing window size, (5) merges the forward and reverse splits into a single read, and (6) discards merged reads that do not meet user specified size threshold (i.e., 1300-1450 bp for this study). Finally, the chopSeq corrected reads are processed using *nanoClust*, a read-partitioning based *de novo* clustering approach which uses VSEARCH[22] to delineate operational taxonomic units (OTUs) at a user specified sequence similarity (i.e., 97% in this study) followed by within OTU pairwise reads alignment and consensus calling for each OTU. The nanoClust algorithm (1) splits each read into three partitions based on user specified partition

length, (2) implements VSEARCH for dereplication, chimera detection,[23] and removal, and clustering for each partition to identify the partition category with optimal (i.e., maximum) number of OTUs (not counting singleton OTUs) and (3) discards singletons. Following this, nanoClust (4) extracts read IDs for each OTU bin from the best performing partition, (5) the extracted read IDs for each OTU bin are then used to obtain full-length chopSeq corrected reads, a subset of reads that fall within 10% of the average full length read distribution within each OTU bin are aligned using Multiple Alignment using Fast Fourier Transform (MAFFT)[24,25] with G-INS-i option, followed by (6) consensus calling to obtain full length representative sequence for each OTU. The entire data processing workflow is shown in Figure 3.

Modifications to the original INC-Seq protocol significantly reduces time required for amplicon concatemer pool preparation. While the proposed DNA preparation protocol is based on the previously developed INC-Seq[18] approach, it contains multiple improvements that allow for faster and more efficient library preparation. These modifications include reduced incubation times for self-ligation step and plasmid-safe DNAse digestion process. More importantly, the current protocol utilizes Tth PrimPol and Phi29 DNA polymerase enzymes for RCA which minimizes the formation of unspecific products that may occur when using random hexamers. Similarly, the NanoAmpli-Seq protocol utilizes T7 endonuclease I enzyme for enzymatic debranching of RCA product combined with mechanical fragmentation step involving use of the g-TUBE. Thus, while our protocol increases the number of intermediate steps for sample preparation, by optimizing each step it reduces the overall time required for sample DNA preparation to 6 hours (approximately 70% reduction compared to the original INC-Seq protocol). These improvements not only result in analyses of near full-length 16S rRNA gene (i.e., twice the

amplicon size of the previously developed INC-Seq approach), but the combination of the improved protocol with appropriate data processing modifications resulted in significant increase in high-quality data post-processing.

NanoAmpli-Seq data yield for 2D and 1D2 experiments. Runs 1, 2, 3, and 4 resulted in 29420, 59490, 142233, and 301432 raw records with post basecalling read lengths ranging from 5bp to 43kbp and 5bp to 234kbp for 2D and 1D2 sequencing protocols (Supplementary Figure S1). The pass reads to total raw reads ratio ranged from 28% for 2D and 7-9% for the 1D2 experiments (Table 1). It is unclear if the low yield of pass reads, particularly for the 1D2 experiments, were due to the concatemerization process or due to other potential issues associated with basecalling. All pass reads were subjected to INC-Seq processing to allow for consensus based error correction using reads with a minimum of three concatemers per read (i.e., reads with less than three concatemers were excluded for any subsequent analyses) as compared to six used by Li et al[18]. The number of concatemers per read passing INC-Seq threshold ranged from 3 to 21 and 3 to 42 for 2D and 1D2 data (Supplementary Figure S1). The total number of reads passing the three concatemer threshold ranged from 36-75% of the base called reads depending on the experiment, sequencing protocol, and alignment approach during INC-Seq processing (Table 1). This was significantly higher than those reported by Li et al[18], primarily due to the use of three compared to six concatemer threshold recommended previously.

INC-Seq processed reads demonstrated incorrect read orientation and presence of tandem repeats. While the median read lengths for post INC-Seq were generally in the expected range (i.e., 1350-1450 bp) (Table 1) (Supplementary Figure 2), manual inspection of the reads revealed

several instances of incorrect read orientation (Figure 2). The amplicon pool preparation protocol relies on RCA of plasmid-like structures constructed through self-ligation of linear amplicons followed by combination of enzymatic debranching and mechanical fragmentation to generate linear molecules with multiple concatemers. Considering the fragmentation and debranching steps are not driven by sequence specificity, it would be reasonable to assume that the resulting linear amplicon is unlikely to have the correct orientation, i.e. 16S rRNA gene specific forward and reverse primers do not flank the entire amplified region. Indeed, we found a vast majority of the 2D and 1D2 INC-Seq consensus reads were incorrectly oriented for the single organism sequencing runs, with forward and reverse primers not located at the ends of the reads. As a result, the reads were chopped at the primer sites and re-oriented to allow for the forward and reverse primers to be correctly oriented. However, during the process of read re-orientation, we also discovered the presence of inserts in the form of tandem repeats. Additional inspection of these inserts revealed that they were composed of multiple repetitive sequences, with the length of these inserts ranging from 10 bp to in excess of 1500 bp (for rare cases) with median tandem repeat size ranging from 12 to 62 bp. The proportion of INC-Seq processed reads with tandem repeats varied from 60-75%, but did not reveal any significant effect of type of aligner used during INC-Seq consensus calling or the sequencing chemistry itself. Interestingly however, the length distribution for the tandem repeats was strongly associated with the sequencing chemistry. Specifically, the 1D2 reads had longer tandem repeats as compared to the 2D reads and demonstrated a bimodal distribution of tandem repeat lengths as compared to the 2D data which showed a unimodal tandem repeat length distribution (Figure 4). While the template and complements in the 2D sequencing chemistry are physically linked by a hairpin adapter, they are not physically linked in the 1D2

sequencing chemistry; this could likely be the cause of differences in tandem repeat length distribution between 2D and 1D2 experiments.

Read re-orientation and tandem repeat removal significantly improves sequence quality. BLASTn analyses of INC-Seq reads against reference database composed of 16S rRNA gene sequences of one (Run1 and Run4) or ten organisms (Run 2 and 3) revealed that a combination of incorrect read orientation and presence of tandem repeats significantly affected overall sequence quality. While the average sequence similarity between INC-Seq consensus reads and the reference sequence was 97±0.37%, the portion of the INC-Seq consensus read demonstrating a contiguous alignment to the reference sequence varied significantly (Figure 5); the remaining section of the read typically resulted in shorter secondary alignments with similar sequence similarity to that of the primary alignment. However, post chopSeq the average proportion of the read aligning to the reference sequences increased to 96±2.3% with additional step of discarding reads less than 1300 bp and greater than 1450 bp increasing the average proportion of the read aligned to 98.4±0.7% while the sequence similarity between chopSeq (97.5±0.42%) and chopSeq followed by size selection (98±0.23%) remained similar to or slightly better than INC-Seq processed read. This demonstrates that read reorientation and tandem repeat removal resulted in reconstruction of reads with high level of similarity to the reference sequences (Figures 5).

Inspections of the read to reference alignment length ratio indicated that the major source of sequence error for both INC-Seq and chopSeq corrected reads originated from deletions; i.e. percent similarity of the read to the reference decreased in proportion to the read to reference alignment ratio for all experiments and INC-Seq aligners used. While deletions in reads were also strongly associated with percent identity for post chopSeq and size selected reads, a small

proportion of chopSeq corrected reads showed read to reference alignment ratios of greater than 1 (Figure 6). This may be due to the retention of small tandem repeats (typically less than 10 bp) during the chopSeq correction resulting in insertion related errors. This suggests that additional optimization of chopSeq may be necessary to remove all tandem repeats within sequences.

***De novo* clustering of chopSeq corrected sequences followed by within cluster consensus calling significantly enhances sequence accuracy.** The overall sequence accuracy increased to an average of 97.9±0.23% following chopSeq read correction and size selection with 98±0.23% of the read aligning to the reference (Figure 5). However, approximately 5 and 10% of reads for the 2D and 1D2 runs, respectively, exhibited sequence accuracy of less than 95% with some sequences aligning over less than 50% of the read length even after chopSeq correction. These poor-quality reads could not be selectively filtered out based on any commonly used quality filtering criteria (e.g., maximum homopolymers length, primer mismatches, etc.) and significantly affected clustering of reads into OTUs. For instance, VSEARCH based clustering of full length post chopSeq and size selected reads (INC-Seq aligner: blastn) at a 97% sequence similarity threshold resulted in 817 (with 777 singletons) and 1301 (with 1238 singletons) for 2D and 1D2 data for single organism experiment and 2122 (with 1742 singletons) and 2725 (with 2447 singletons) for 2D and 1D2 data for 10 organism experiments. We hypothesized that accrual of residual errors over the entire read length hampered the accuracy of the OTU clustering and that accurate clustering was more likely over shorter regions of the reads due to fewer absolute errors. To this end, we developed and implemented nanoClust algorithm (see above) by splitting each read into multiple partitions (we used three splits from 0-450 bp, 451-900 bp, and 901-1300 bp) followed by clustering using VSEARCH. This resulted in an additional loss of 15-26% of chopSeq

corrected reads, which were primarily due to singleton OTUs and/or chimeric sequences (Table 1). The nanoClust approach was far superior than direct clustering of full-length reads and resulted in accurate determination of the number of OTUs, with one-two spurious OTUs and no false negatives (Table 2) depending on the type of experiment and INC-Seq aligner used. MAFFT-G-INS-i alignment of 50 reads from each OTU resulted in consensus reads with the entire read aligned to the reference and average consensus read of 99.5% and accuracy values for individual OTUs ranging from 99.2 to 100%. Nearly all errors in the nanoClust consensus reads originate from single base pair deletions in a few homopolymers regions (homopolymers > 4 bp) and no detectable insertions, with one-two mismatches associated with the spurious OTUs (Figure 7). Phylogenetic analyses of the consensus sequences demonstrated close placement of the OTU consensus sequences with their corresponding references, with excellent pairwise alignment between the two (Supplementary Figures S3 and S4). While accurate OTU estimation allowed for single OTUs to be detected in the one organism experiment, the overall community structure deviated from the theoretical community structure for the ten organism experiments (Figure S5) and thus additional protocol optimization is essential to ensure levels of deviation from theoretical community structure does not exceed what may be seen from PCR biases[26]. The nanoClust implementation in this study included a specified threshold of a maximum of 50 reads per OTU to generate OTU consensus sequences. This was feasible because our study focusses on single organism and even mock community of ten organisms. Thus, the process of consensus construction was not limited by number of reads that could be recruited. However, it would be critical to determine the potential for poor quality consensus sequence due to fewer reads with an OTU in naturally derived mixed microbial communities. To this end, we varied the number of reads used for consensus sequence construction from 5 to 100 for 2D and 1D2 data from the one organism

experiment (Figure 8). While, the consensus sequence accuracy surpasses 99% with the use of more than five reads for consensus sequence construction and plateaus at approximately 10-15 reads, the variability in accuracy with repeated random sampling of data was much more pronounced when fewer than 50 reads were used for both 2D and 1D2 data. This suggests that consensus sequence accuracy is reliably high only for OTUs where a minimum of 50 reads are available for use in constructing the consensus sequence. This would have an impact on sequence quality of low abundance OTUs.

Discussion

The current study uses mock communities to develop and validate the NanoAmpli-Seq workflow for long amplicon (>1 kbp) sequencing on the nanopore sequencing platform. While this study focusses on the near full length 16S rRNA gene, in principle the approach outlined by the NanoAmpli-Seq workflow should be amenable to amplicons generated from PCR amplification of any target gene irrespective of target gene length. While leveraging the previously described INC-Seq protocol, NanoAmpli-Seq adds several novel components which significantly enhances the amplicon sequencing workflow for the nanopore platform. The improvements over the previously described INC-Seq protocol involve modifications to both library preparation (i.e. PrimPol based *in silico* primer synthesis for RCA, debranching and fragmentation, shorter protocol length) and to the data analyses. Specifically, we identify and fix the issues associated with incorrect read-orientation and presence of tandem repeats in INC-Seq consensus reads, thus allowing for nearly the entire length of the chopSeq corrected reads to be aligned to the reference with accuracies (97-98%) similar to those described by Li et al[18]. While the original INC-Seq protocol prescribed a concatemer threshold of six, we halved the concatemer threshold to three,

thus more than doubling the number of INC-Seq consensus reads available as a proportion of the base called reads. We could use a lower INC-Seq concatemer threshold due to the ability to perform another round of alignment and consensus calling during the *nanoClust* step.

While the correction of INC-Seq consensus reads using chopSeq did not allow for sequences with high enough quality for direct OTU clustering using VSEARCH, the read-partitioning based sequence clustering allowed for accurate determination of the number of OTUs in the mock community. Further, *de novo* sequence clustering using nanoClust provided the opportunity to significantly increase the number of sequences used for consensus calling. In this study, we used 50 reads for consensus calling (i.e. 150x coverage considering three concatemer threshold set for INC-Seq) which resulted in average sequence accuracy of 99.5%. The use of more than 50 reads for consensus calling in nanoClust did not improve sequence accuracy, while reducing the number of reads resulted in reduced precision. This threshold of 50 reads for both the 2D and 1D2 sequencing data suggests that the OTUs with fewer than 50 reads are likely to have sequence quality lower than those OTUs with greater than 50 reads. It should however be noted that using a 10 read threshold (i.e., 30x coverage when including three concatemer threshold for INC-Seq) consistently allowed for sequence accuracy consistently greater than 99% and thus sequence classification to species level using any of the current sequence classification approaches[27,28] would be reliable even for lower abundance OTUs.

While the NanoAmpli-Seq workflow represents a significant improvement in amplicon sequencing on the nanopore platform, some key limitations remain. For instance, the NanoAmpli-Seq sequence accuracy is still lower than those reported for short amplicons sequenced on the Illumina Platform [3] and the full-length 16S rRNA sequencing on the PacBio platform using the

approach described previously [17]. Our analysis shows that the sequence accuracy does not improve with more than 50 sequences used in the nanoClust based consensus calling process. Nearly all the errors in the OTU consensus sequences originate from single deletions at homopolymers regions, specifically for homopolymers greater than 4 bp. This homopolymer error issue on the nanopore platform is well known[13,29] and is likely to best resolved during the base calling process rather than subsequent data processing or by processing signal data (rather than base called data) all the way through clustering, followed by base calling as the final step. The second key limitation of our approach is the low data yield at the base calling step, i.e. base called reads represents only a small portion (i.e., 7-9% for 1D2 data) of the raw records. This data loss is significant and could potentially deter the widespread use of the nanopore platform for amplicon sequencing. While the precise cause of the low yield of pass reads post base calling is unclear, it is possible that this could be due to DNA strands that represent poor diversity, i.e., same amplicon sequence repeated multiple times. If this is determined to be an issue, one remedy would be to fragment and size select the RCA product to minimize sequencing of DNA strands with more concatemers than is essential, i.e. limit the number of concatemers between 3-10. In our study, DNA strands with as many as 42 concatemers were sequenced but we saw marginal to no improvements in sequence accuracy for reads with more than 10 concatemers. The NanoAmpli-Seq workflow includes a *de novo* clustering step and as long as the sequence accuracy post chopSeq is ~97% (3-10 concatemers required), the binning process should provide for sufficient coverage for consensus based sequence correction to accuracies in excess of 99%.

Materials and Methods

Mock community description and preparation

Two different mock communities were constructed for the experiments outlined in this study. First, single organism mock community was constructed by amplifying the near full length of the 16S rRNA gene from genomic DNA of *Listeria monocytogens* using primers sets 8F (5'-AGRGTTTGATCMTGGCTCAG-3') and 1387R (5'-GGGCGGWGTGTACAAG-3'), both with 5' phosphorylated primers (Eurofins Genomics) [26]. Phosphorylated ends are essential for the subsequent self-ligation step. PCR reaction mix was prepared in 25µl volumes with use of 12.5µl of Q5® High-Fidelity 2X Master Mix (New England BioLabs Inc., M0492L), 0.8µl of 10pmol of each primer, 9.9µl of nuclease-free water, (Roche Ltd.) and 1ng of bacterial DNA in total followed by PCR amplification as described previously[26]. PCR amplicons from replicate PCR reactions were combined and purified with use of HighPrep™ PCR magnetic beads (MagBio, AC-60050) at 0.45x ratio. The ten organism mock community was constructed from purified near full-length 16S rRNA amplicons of 10 organisms. Briefly, genomic DNA from 10 bacteria was obtained from DSMZ, Germany (Supplementary Table S1) and the aforementioned primers, PCR reaction mix and thermocycling conditions were used to independently PCR amplify the near full length 16S rRNA gene, followed by purification using HighPrep™ PCR magnetic beads as detailed above. The purified amplicons from each organism were quantified on the Qubit using dsDNA HS kit, normalized to 4ng/µl, and combined to generate an amplicon pool consisting of equimolar proportion of the 16S rRNA gene amplicons of the 10 organisms.

DNA sequencing library preparation.

To circularize the linear amplicons into plasmid-like structures, 5µl of Blunt/TA Ligase Master Mix (New England Biolabs, M0367L) was added to 55 µl of amplicon pool at a concentration of 1ng/µl and incubated for 10min at 15°C then 10min at room temperature to (total time = 20

minutes). Not all linear amplicons self-ligate into plasmid-like structures, but some are likely to cause long chimeric linear amplicons. These long chimeric structures were removed using magnetic bead based purification with the following modifications. HighPrep™ PCR magnetic beads were homogenized by vortexing followed by aliquoting 50μl into sterile 2ml tube and placed on magnetic rack for 3min. A total of 25μl of supernatant was carefully removed using a sterile pipette to concentrate the beads to 2x its original concentration. The tube was removed from the magnetic rack and vortexed vigorously to resuspend the beads. This concentrated bead solution was used at a ratio of 0.35x to remove any amplicons greater than 2000 bp in the post-ligation reaction mix. Briefly, the post-ligation product was mixed with concentrated bead solution at 0.35x ratio by vortexing followed by incubation for three minutes at room temperature. The tube was placed on the magnetic rack to separate the beads from solution, followed by transferring of clear liquid containing DNA structures less than 2000 bp into new sterile tubes. Sample containing short self-ligated molecules was subject to another round of concentration using standard magnetic beads at 0.5x ratios according to manufacturer instructions and eluted in 15μl of warm nuclease-free water. Concentrated and cleaned DNA pool consisting of plasmid-like structures and remaining linear amplicons was then processed with Plasmid-Safe™ ATP-Dependent DNase (Epicentre, E3101K) reagents to digest linear amplicons using the mini-prep protocol according to manufacturer instructions and was followed by another round of cleanup with magnetic beads at 0.45x ratio as described before and then eluted in 15μl of warm nuclease-free water.

The pool containing plasmid-like structures was subject to RCA with use of TruPrime™ RCA Kit (Sygnis, 390100) random hexamer-free protocol. Samples were prepared in triplicate and processed according to manufacturer protocol with all incubations performed in triplicate for 120-150 min depending of the assay efficiency. The progress of RCA was monitored by measuring the

concentration of DNA using Qubit® 2.0 Fluorometer at 90, 120 or 150 min time points. Negative controls were processed and analyzed concomitantly with the samples. The final concentration of the RCA product after 150 minutes of incubation was typically 70 ng/µl when using a starting DNA concentration of 0.35-0.4 ng/µl with no detectable unspecific product formation in the negative control.

Replicates RCA products were combined (~4.5µg of DNA in total) and subject to de-branching and fragmentation of post-RCA molecules to remove hyperbranching structures generated during RCA. The RCA product was first treated with T7 endonuclease I enzyme (New England BioLabs, M0302S) by adding 2µl of the reagent to the 65µl of RCA product followed by vortexing and incubation as recommended by manufacturer. Subsequently, the reaction mix was transferred into a g-TUBE (Covaris, 520079) and centrifuged at 1800 rpm for 4min or until the entire reaction mix passed through the fragmentation hole. The g-TUBE was reversed and centrifugation process was repeated. Post debranching and fragmentation, short fragments were removed using the modified bead based clean up step using concentrated bead solution (see above for concentration procedure). Concentrated beads were mixed with fragmented RCA product at 0.35x ratio, vortexed for 15sec, and incubated at room temperature for 3min then placed on magnetic rack until beads separated and the supernatant was removed. The beads were subsequently washed with 70% freshly prepared ethanol according to manufacturer protocols. Size selected amplicons bound to the beads were eluted in 41µl of warm nuclease-free water. Preliminary experiments indicated that one round of de-branching did not completely resolve the hyperbranching structure, which was inferred based on poor sequencing yield likely caused due to pore blocking by hyperbranched DNA. As a result, a second round of enzymatic de-branching using T7 endonuclease I was added and the de-branched product was cleaned a second time using

the bead based clean-up step. Figure S6 shows an example BioAnalyzer traces of the RCA product post-debranching/fragmentation and post-cleanup using magnetic bead based protocol.

Finally, the de-branched RCA product was treated with NEBNext® FFPE DNA Repair Mix (New England BioLabs, M6630S) for gap filling and repairing DNA damages caused during g-TUBE fragmentation and T7 endonuclease I enzyme. All reagent components were combined with de-branched RCA product according to manufacturer recommendations and incubated at 12°C for 10min then at 20°C for another 10min. Post incubation, the repaired RCA product was cleaned using standard magnetic beads at 0.5x ratio, washed with 70% ethanol, and eluted in 46µl of warm nuclease-free water. Concentration of the DNA product was measured using Qubit and was approximately 20-25 ng/µl with a total yield of ~1000 ng of DNA with product size typically ranging from 1500bp to 20,000 bp. A total of 45µl DNA pool of concatamerized amplicons was prepared for sequencing using the standard 2D and 1D2 library preparation protocol by Oxford Nanopore Technologies (SQK-LSK208, SQK-LSK308) according to manufacturer specifications to obtain pre-sequencing mix. And the final concentration for prepared libraries was determined using dsDNA HS kit on the Qubit instrument. Detailed step-by-step protocol is provided in the supplementary text.

DNA sequencing.

The MinION MkIB was connected to Windows personal computer compatible with Oxford Nanopore Technologies requirements. R9.4 (FLO-MIN106) and R9.5 (FLO-MIN107) flow cells were placed onto the MinION Mk1B (Oxford Nanopore Technologies). Platform quality control was performed using MinKNOW software (v1.4.2 for 2D and v1.6.11 for 1D2 libraries). Only flow cells containing above 1100 active pores were used in this study. Each flow cell was primed

twice according to ONT specifications using priming buffer consisting of equal parts of running buffer (RBF1) and nuclease-free water with 10 min breaks between subsequent primes. The loading mix was prepared with 12µL pre-sequencing mix, 75µL of RBF1, and 63µL of nuclease-free water. Loading mix was sequenced with MinKNOW settings appropriate for 2D or 1D2 options and standard 48h processing time for every run. Albacore 1.2.4 was used to convert raw signals into HDF5 file format using switch options FLO-MIN106 and SQK-LSK208 for 2D data and FLO-MIN107 and SQK-LSK308 for 1D2 data.

Data processing

HDF5 raw signals from each sequencing run were converted to FASTQ format using Fast5-to-Fastq (https://github.com/rrwick/Fast5-to-Fastq) and then from FASTQ to FASTA with seqtk (https://github.com/lh3/seqtk). The resultant data was subject to INC-Seq processing using blastn, Graphmap, and POA aligners with concatemer threshold of 3 with the iterative flag for consensus error correction using PBDAGCON. INC-Seq consensus reads were subject to chopSeq by specifying forward and reverse primer sequences and upper (1450 bp) and lower (1300 bp) read size thresholds for size selection (chopSeq_SS). The chopSeq processed reads were subsequently processed using nanoClust with partition size limits (flag "-s") of 0,450,451,900,901,1300 which splits the reads into three partitions of 450, 450, and 400 bp respectively prior to further processing. VSEARCH was used for chimera removal (using uchime) followed by clustering of reads in each partition at a hardcoded sequence similarity threshold of 97%. For the optimal binning results, nanoClust then outputs consensus sequence for each OTU based on MAFFT-G-INS-i pairwise alignment as described previously and outputs an OTU table with reads corresponding to each OTU after discarding singletons.

Fasta files from all stages (i.e. raw, INC-Seq, chopSeq, chopSeq_SS, and OTU consensus sequence) were analyzed for read lengths in R[30]. The number of concatemers on each raw read was estimated by dividing the length of each raw read with the length of its corresponding INC-Seq consensus read. At each stage of processing, the reads were aligned to reference dataset using blastn and only the match with the highest bitscore was considered. The ratio of read to reference alignment at appropriate points (as discussed above) was estimated based on blastn results. The percent identity from the blastn results were used as measure consensus sequence accuracy for the nanoClust output. All figures for the manuscript were generated in R using packages "ggplot2"[31], "gridExtra", and "cowplot" (https://github.com/wilkelab/cowplot), as appropriate. Neighbor-Joining tree construction (Figure S3 and S4) was performed after muscle alignment[32] (default parameters) and using Jukes-Cantor model was constructed in Geneious (version 8) using 100 bootstraps.

Data availability: All data is available on European Nucleotide Archive (ENA) under primary accession number: PRJEB21005.

Code availability: All code used for this study as part of development of the NanoAmpli-Seq workflow is available on https://github.com/umerijaz/nanopore

Author contribution: STC, UZI, and AJP designed and developed the experiments. STC performed the experiments. UZI wrote the code for chopSeq and nanoClust. STC, UZI, and AJP analyzed the data and contributed equally to writing this manuscript.

Acknowledgements: This research was supported by EPSRC award no: EP/M016811/1. STC is supported by the EPSRC Doctoral Training Center at University of Glasgow. UZI is supported by a NERC Fellowship (NE/L011956/1).

References.

1 Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* **6**, 1621-1634 (2012).
2 Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Nat Acad of Sci* **108**, 4516-4522, (2011).
3 Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol* **79**, 5112-5120, (2013).
4 Schoch, C. L. *et al.* Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc Nat Acad Sci* **109**, 6241-6246, (2012).
5 Jünemann, S. *et al.* Bacterial community dhift in treated periodontitis patients revealed by Ion Torrent 16S rRNA gene amplicon sequencing. *PLOS one* **7**, e41606, (2012).
6 Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**, 333-351, (2016).
7 Yuan, C., Lei, J., Cole, J. & Sun, Y. Reconstructing 16S rRNA genes in metagenomic data. *Bioinformatics* **31**, i35-43, 2015.
8 Miller, C., Baker, B., Thomas, B., Singer, S. & Banfield, J. EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biol* **12**, R44 (2011).
9 Karst, S. M. *et al.* Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. *Nat Biotech*, online publication, (2018).
10 Mardis, E. R. DNA sequencing technologies: 2006–2016. *Nat Protocols* **12**, 213, (2017).
11 Wommack, K. E., Bhavsar, J. & Ravel, J. Metagenomics: Read length matters. *Appl Environ Microbiol* **74**, 1453-1463, (2008).
12 Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**, 722-736, (2017).
13 Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* **12**, 733-735, (2015).
14 Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228-232, (2016).
15 Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* **13**, 1050-1054, (2016).
16 Michael, T. P. *et al.* High contiguity A*rabidopsis thaliana* genome assembly with a single nanopore flow cell. *bioRxiv*, doi:10.1101/149997 (2017).

17 Schloss, P. D., Jenior, M. L., Koumpouras, C. C., Westcott, S. L. & Highlander, S. K. Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. *PeerJ* **4**, e1869, (2016).
18 Li, C. *et al.* INC-Seq: accurate single molecule reads using nanopore sequencing. *Gigascience* **5,** 34, (2016).
19 Benítez-Páez, A., Portune, K. J. & Sanz, Y. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. *GigaScience* **5**, 4, (2016).
20 Cusco, A. *et al.* Using MinION to characterize dog skin microbiota through full-length 16S rRNA gene sequencing approach. *bioRxiv*, doi:10.1101/167015 (2017).
21 Kerkhof, L. J., Dillon, K. P., Häggblom, M. M. & McGuinness, L. R. Profiling bacterial communities by MinION sequencing of ribosomal operons. *Microbiome* **5**, 116, (2017).
22 Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahe, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, **4**, e2584, (2016).
23 Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194-2200, (2011).
24 Katoh, K., Misawa, K., Kuma, K.-i. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucl Acid Res* **30**, 3059-3066 (2002).
25 Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics* **9**, 286-298, (2008).
26 Pinto, A. J. & Raskin, L. PCR Biases Distort Bacterial and Archaeal Community Structure in Pyrosequencing Datasets. *PLoS one* **7**, e43093, (2012).
27 Wang, Q., Garrity, G., Tiedje, J. & Cole, J. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**, 5261-5267, (2007).
28 Werner, J. J. *et al.* Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. *ISME J* **6**, 94-103, (2012).
29 Feng, W. *et al.* Improving alignment accuracy on homopolymer regions for semiconductor-based sequencing technologies. *BMC Genomics* **17**, 521, (2016).
30 R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria, 2014).
31 Wickham, H. ggplot2: elegant graphics for data analysis. (2009).
32 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797, (2004).

Table 1: Summary of the total number of reads and median read lengths at each step of the data processing workflow for all experiments.

	Number	Of	Reads		Median	Read	Length	
Protocol	2D	1D2	2D	1D2	2D	1D2	2D	1D2
Experiment	One organism	One organism	Ten organism	Ten organism	One organism	One organism	Ten organism	Ten organism
Run	1	4	2	3	1	4	2	3
Raw records	29420	301432	59490	142233	-	-	-	-
Pass reads	8108	20888	16403	12011	4868	7305	5401	6418
INC-Seq aligner	Blastn							
INC-Seq	3911	7618	7011	9081	1394	1433	1396	1413
chopSeq	3911	7618	7011	9081	1383	1387	1377	1374
chopSeq-size select	3748	7288	5186	7265	1384	1388	1377	1375
otuSeq	2997	6153	3677	5465	1396	1398	1384	1386
INC-Seq aligner	Graphmap							
INC-Seq	3902	7631	7004	9169	1400	1439	1401	1420
chopSeq	3902	7631	7004	9169	1384	1388	1378	1377
chopSeq-size select	3765	7399	5190	7496	1384	1389	1377	1377
otuSeq	2981	6179	4141	5490	1397	1396	1384	1386
INC-Seq aligner	POA							
INC-seq	3913	7643	7025	9191	1414	1457	1415	1443
chopSeq	3913	7643	7025	9191	1394	1396	1386	1387
chopSeq-size select	3779	7088	5076	6954	1394	1396	1385	1386
otuSeq	3184	5993	3916	5622	1398	1389	1384	1386

Table 2: Number of OTUs detected and consensus sequence accuracy for all experiments using the nanoClust for OTU clustering and consensus calling approach.

	Number	Of	OTUs		Average	Consensus	Accuracy	(%)
Protocol	2D	1D2	2D	1D2	2D	1D2	2D	1D2
Experiment	One organism	One organism	Ten organism	Ten organism	One organism	One organism	Ten organism	Ten organism
Run	1	4	2	3	1	4	2	3
Theoretical	1	1	10	10				
INC-Seq aligner	Blastn							
OTUs detected	1	1	11	11	99.36	99.5	99.47	99.61
Spurious OTUs	0	0	1	1	-	-	99.22	99.37
Non-Detect	0	0	0	0	-	-	-	-
INC-Seq aligner	Graphmap							
OTUs detected	1	1	11	11	99.43	99.43	99.44	99.61
Spurious OTUs	0	0	1	1	-	-	99.29	99.5
Non-Detect	0	0	0	0	-	-	-	-
INC-Seq aligner	POA							
OTUs detected	1	2	10	12	99.5	99.61	99.60	99.52
Spurious OTUs	0	1	0	2	-	99.57	-	98.67
Non-Detect	0	0	0	0	-	-	-	-

Figure 1: Overview of the sample preparation protocol for 16S rRNA gene pool preparation, plasmid-like structure preparation, enzymatic debranching and mechanical fragmentation, and 2D and 1D2 library preparation including intermediate clean-up steps

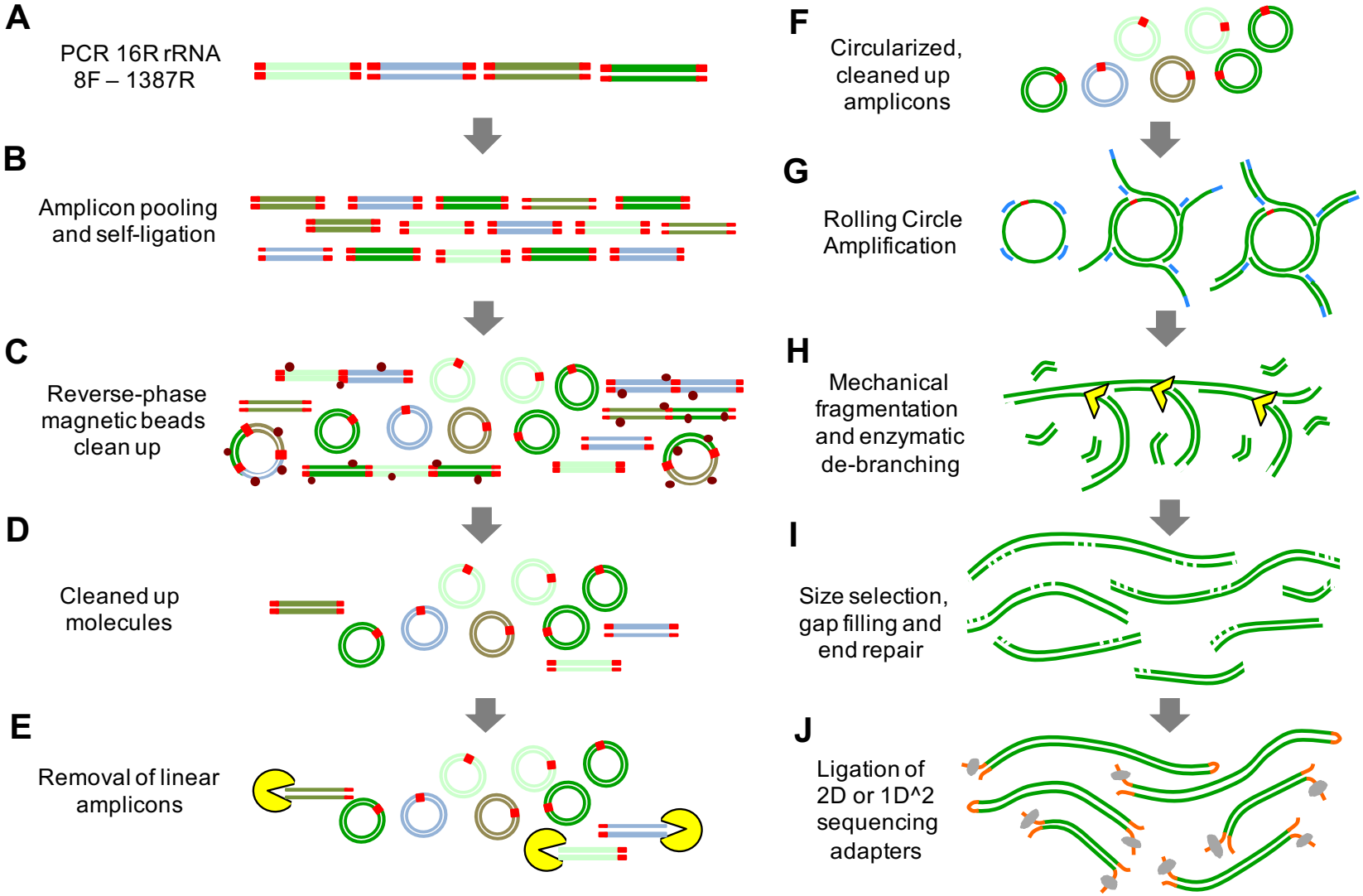


Figure 2: (A) Example of INC-Seq consensus reads showing the improper orientation with forward (maroon) and reverse (green) primers co-located and incorrectly oriented. **(B)** Manual splitting and re-orientation of the reads revealed the presence of tandem repeats in the forward and reverse splits which were identified and removed using chopSeq. **(C)** An expanded view of tandem repeat region in Figure 2B.

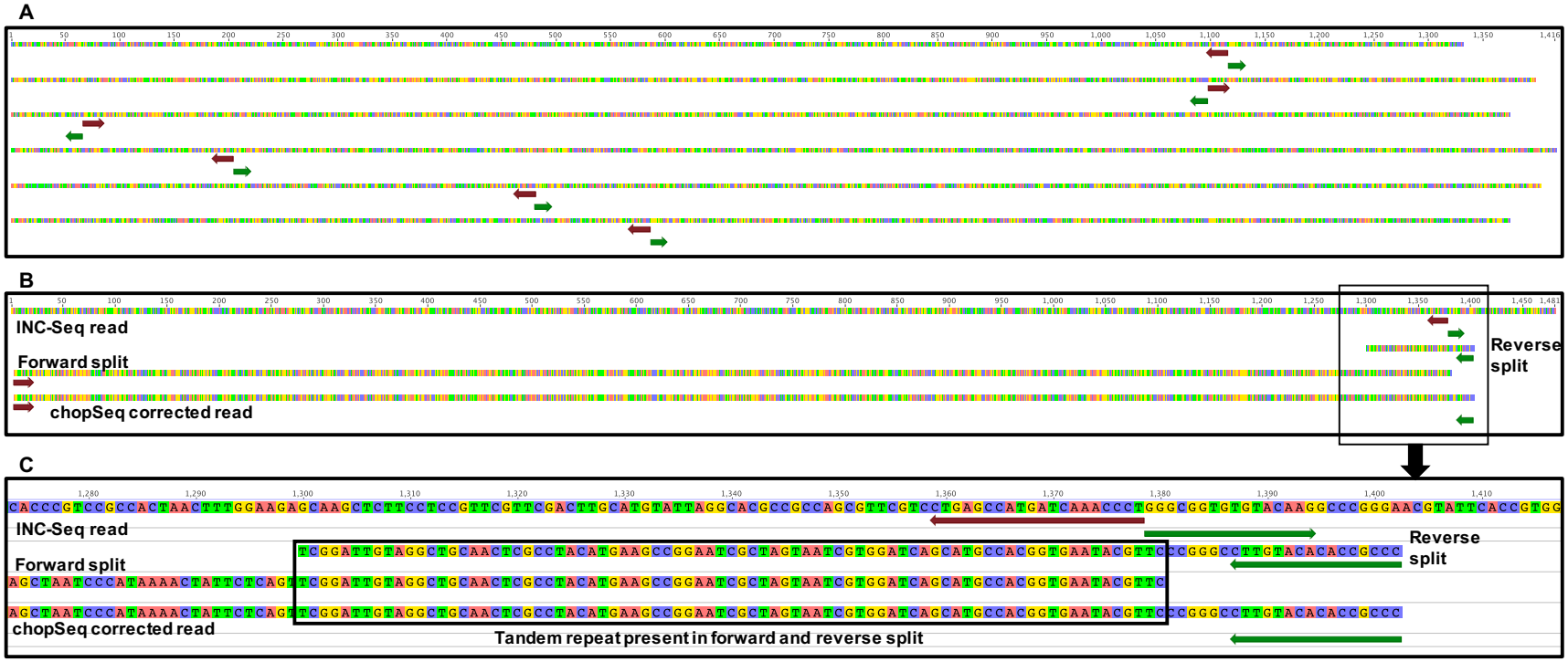


Figure 3: **(A)** Overview of the INC-Seq based anchor alignment and iterative consensus calling using PBDAGCON. **(B)** INC-Seq consensus reads were subject to chopSeq based read reorientation followed by tandem repeat removal and size selection to retain reads between 1300-1450bp. **(C)** chopSeq corrected reads are subject to partitioning followed by VSEARCH based binning to identify optimal binning results using partition that generates maximum number of OTUs (without singletons). MAFTT-G-INS-i was then used for pairwise alignment of a subset of full length reads from each OTU bin of for the best performing partition and the alignment was used to create the OTU consensus read.

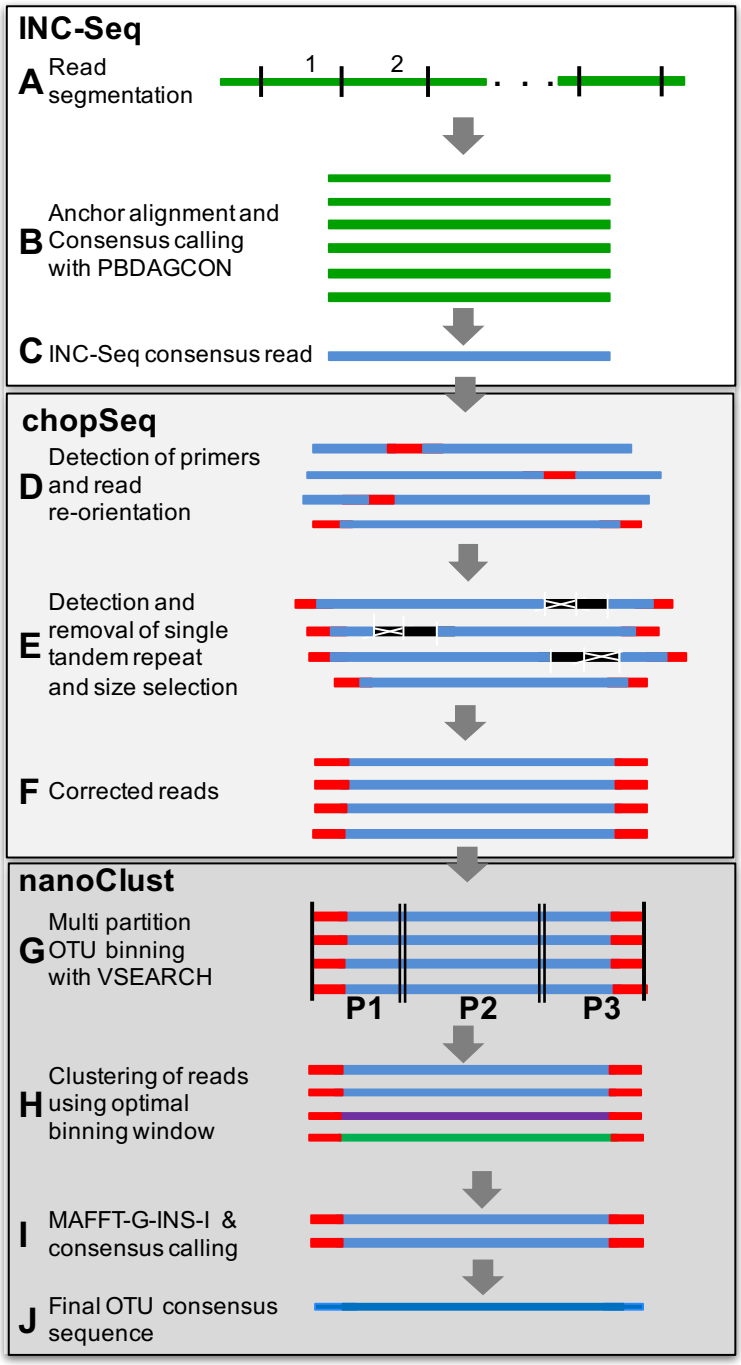


Figure 4: Histogram of tandem repeat length distribution of the INC-Seq processed reads did not show any effect of the aligner used during the INC-Seq process, but rather a marked effect of the sequencing chemistry. Results are only shown for INC-Seq reads generated using blastn aligner.

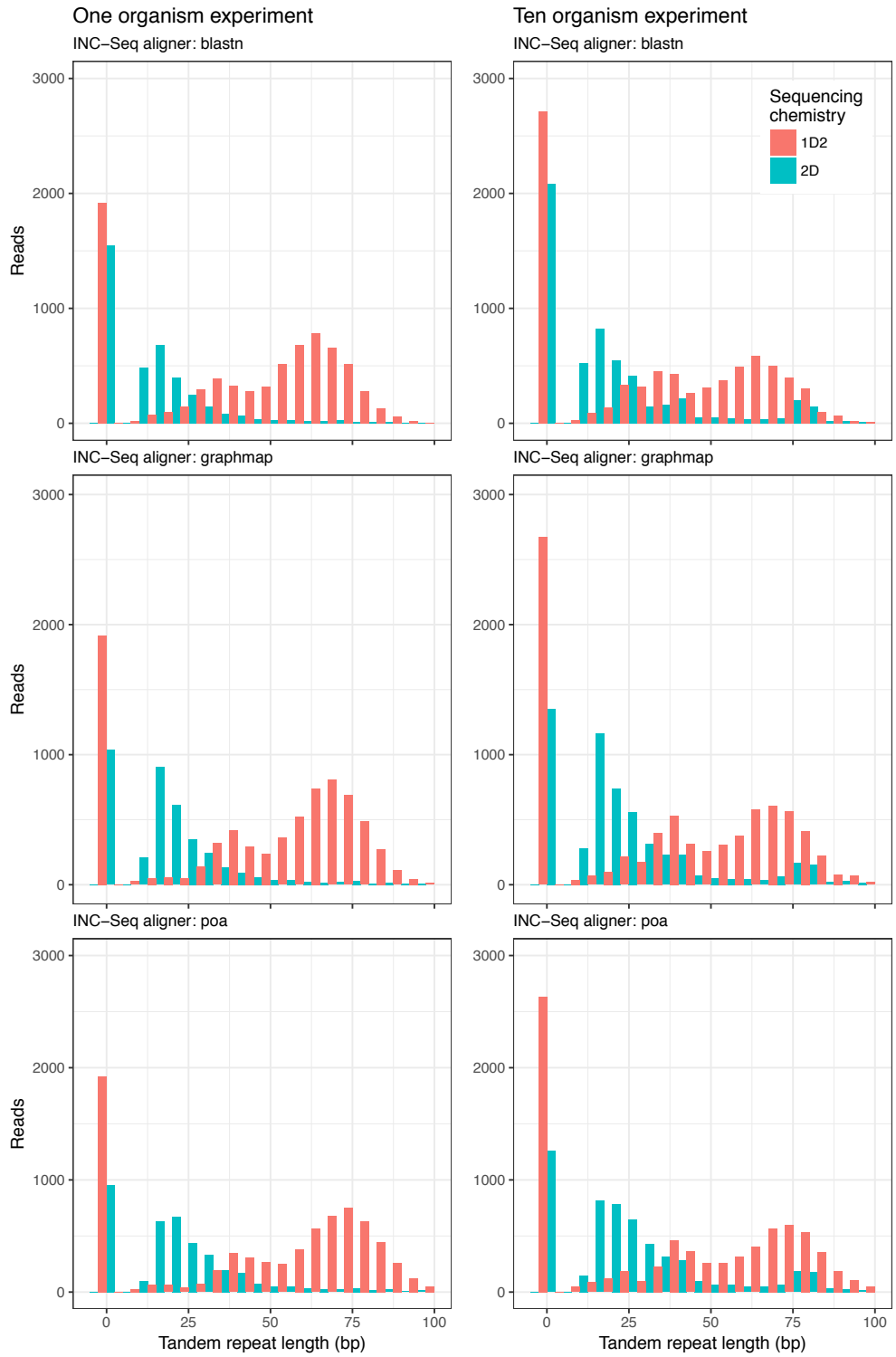


Figure 5: While the distribution of percent identities of INC-Seq and chopSeq processed reads (chopSeq_SS) was on average 97-98%, variable lengths of the INC-Seq processed reads aligned to the reference sequences. In contrast, nearly the entire length of the chopSeq processed reads aligned to the reference sequence without affecting overall sequence similarity. Results are only shown for INC-Seq reads generated using blastn aligner.

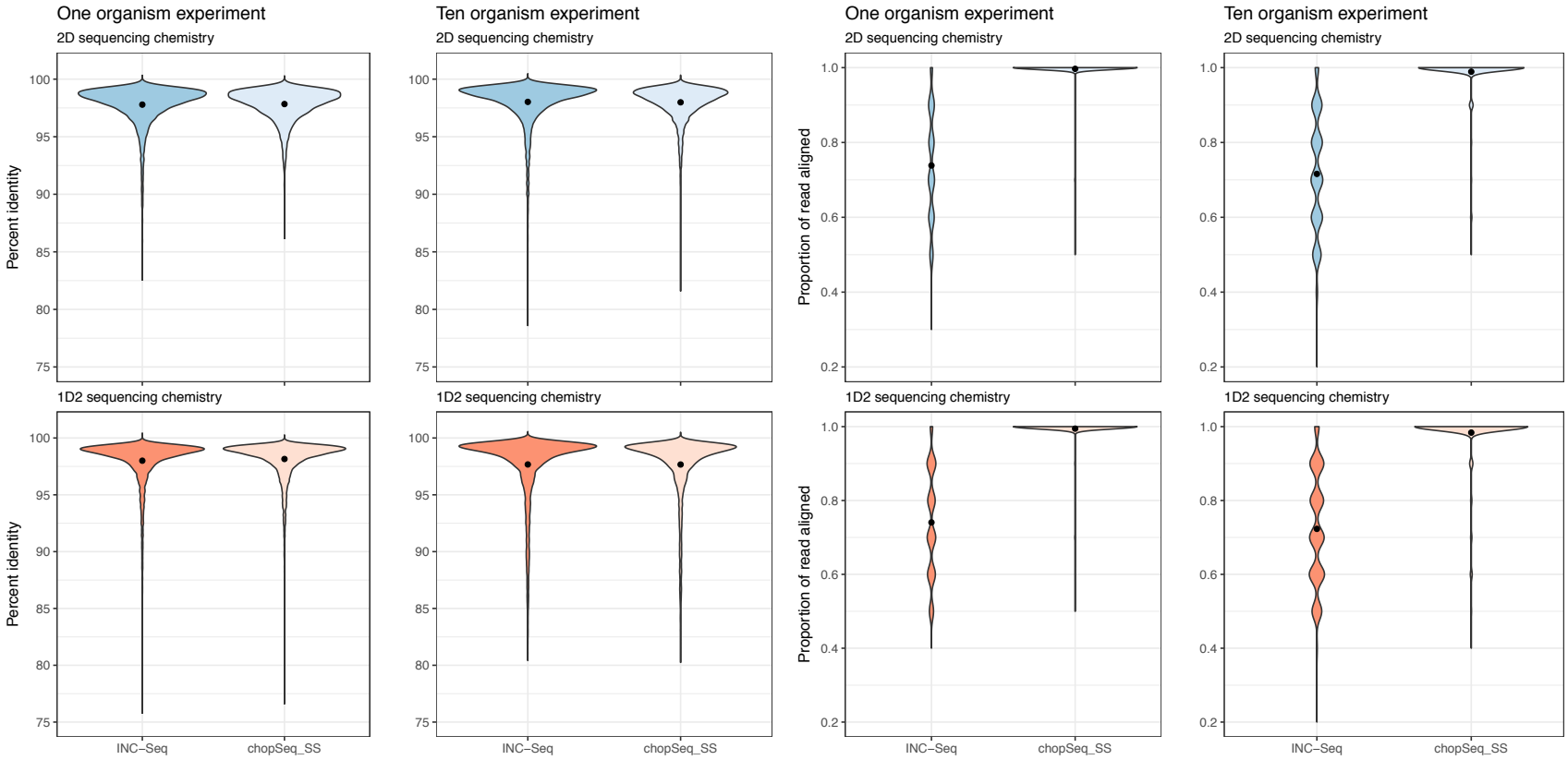


Figure 6: The ratio of the alignment length of INC-Seq and chopSeq corrected reads to that of the corresponding reference sequences was consistently lower than 1, suggesting that deletions in the INC-Seq and chopSeq corrected reads were the primary cause of dissimilarity with the reference sequences. Results are only shown for INC-Seq reads generated using blastn aligner.

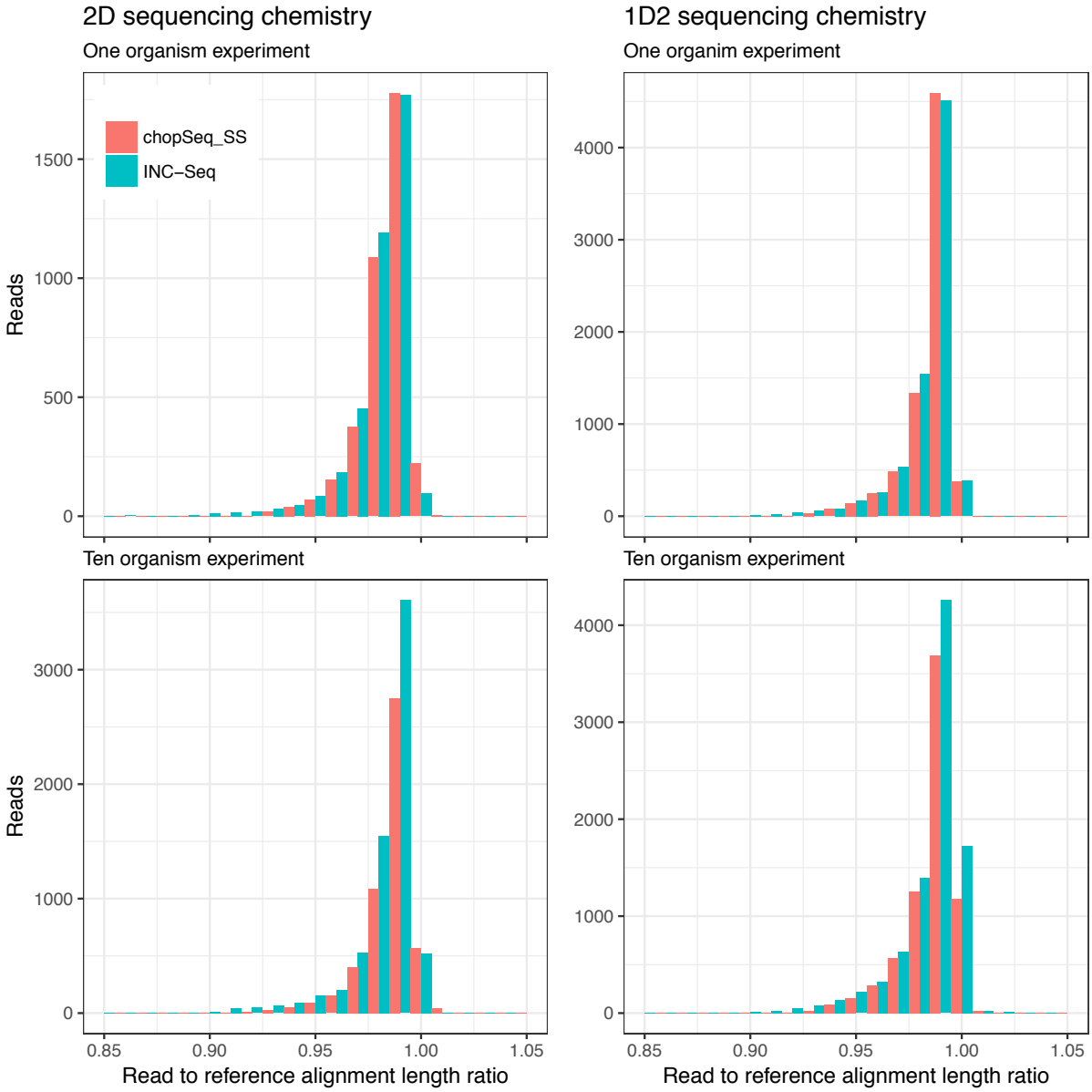


Figure 7: nanoClust clustering and consensus sequence generation resulted in few spurious OTUs and average similarity to the reference sequence of ~ 99.5%. Results are shown for one (A, B, C, G, H, I) and ten organism experiments (D, E, F, J, K, L) for 2D (blue data points) and 1D2 (red data points) experiments with the use of blastn (A, G, D, J), Graphmap (B, H, E, K), and poa (C, I, F, L) aligners during INC-Seq processing. Spurious OTUs are indicated for each panel.

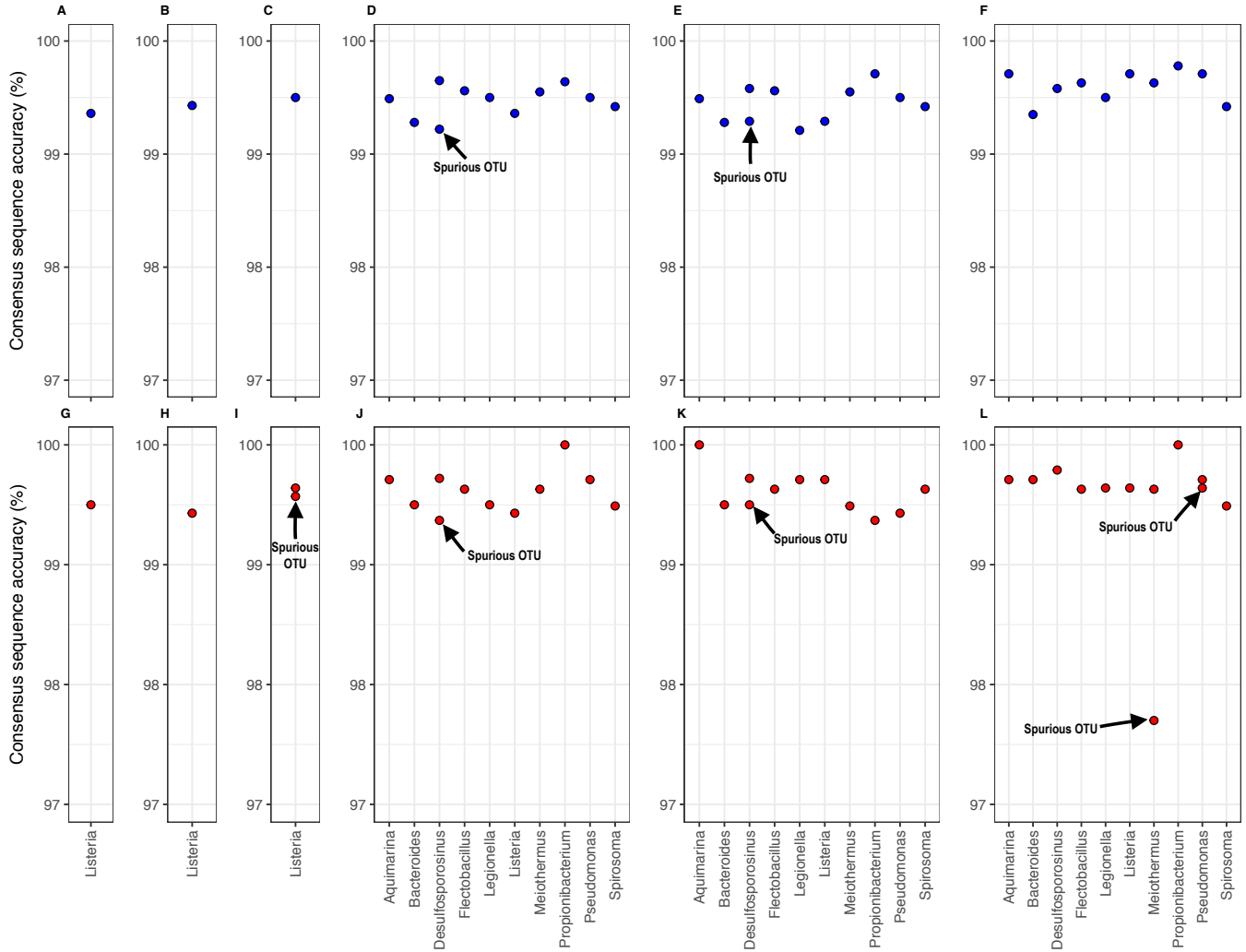
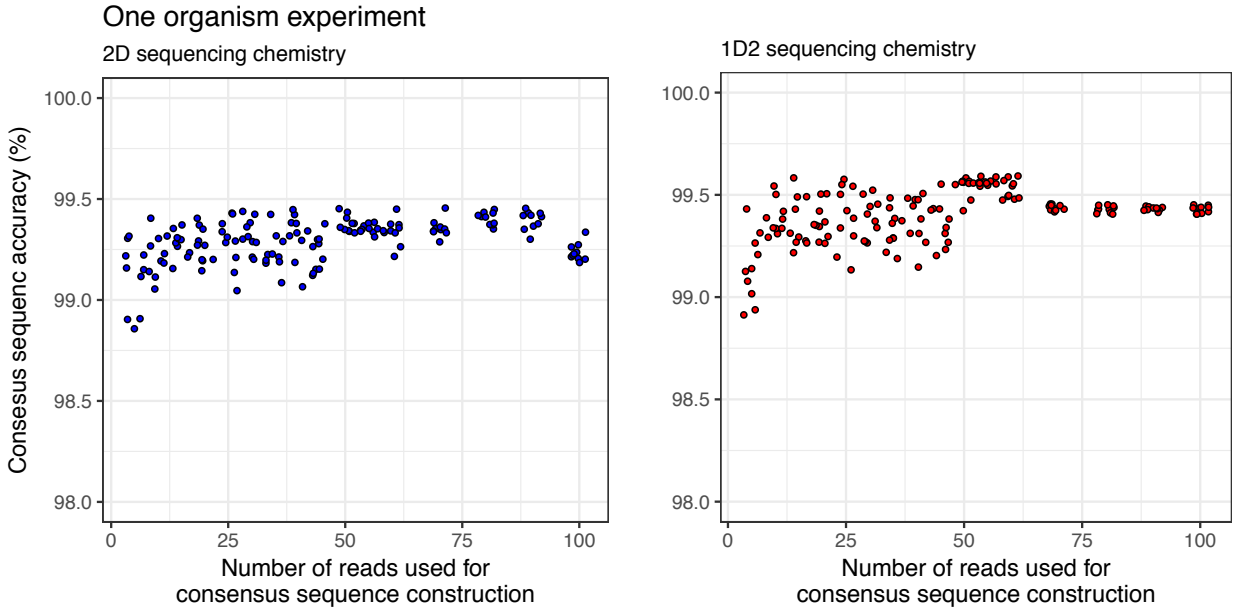


Figure 8: Consensus sequence accuracy plateaus with the use of 10-15 reads for MAFFT-G-INS-i alignment and consensus calling. However, with increasing number of reads used for consensus calling the variability in consensus sequence accuracy from repeated sampling of data diminishes significantly for both 2D and 1D2 sequencing chemistry. Data is shown for one organism experiment where the blastn aligner was used during INC-Seq.



Supplementary tables and figures

Table S1: Names and DSM catalog numbers of bacteria used to construct mock communities for the single organism and ten organism experiments and their corresponding accession numbers are shown below. The 16S rRNA genes were extracted from genome assemblies using RNAmmer[1] for bacteria for use in estimation of sequencing accuracy. Where genome assemblies were unavailable, the 16S rRNA gene sequence in Genbank was utilized.

Organism name	DSM catalog number	Genbank assembly accession number	16S rRNA gene accession number
One organism experiment			
Listeria monocytogenes	19094	HE999705.1	
Ten organism experiment			
Aquimarina intermedia	17527	-	AM113977
Bacteroides vulgatus	1447	CP000139	-
Desulfosporosinus orientis	765	CP003108	-
Flectobacillus major	103	ATXY00000000	-
Legionella pneumophila	7513	AE017354	-
Listeria monocytogenes	19094	HE999705.1	-
Meiothermus ruber	1279	CP001743	-
Propionibacterium acnes	16379	AE017283	-
Pseudomonas aeruginosa	1128	NC_009656	-
Spirosoma linguale	74	CP001769	-

1 Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research* **35**, 3100-3108, (2007).

Figure S1: Violin plots showing the read length distribution of raw data (i.e., post base calling with Albacore 1.2.4) and the number of concatemers on each base called read estimated using read lengths from INC-Seq processing using the "blastn" alignment approach for all four experiments involving both 2D and 1D2 chemistry.

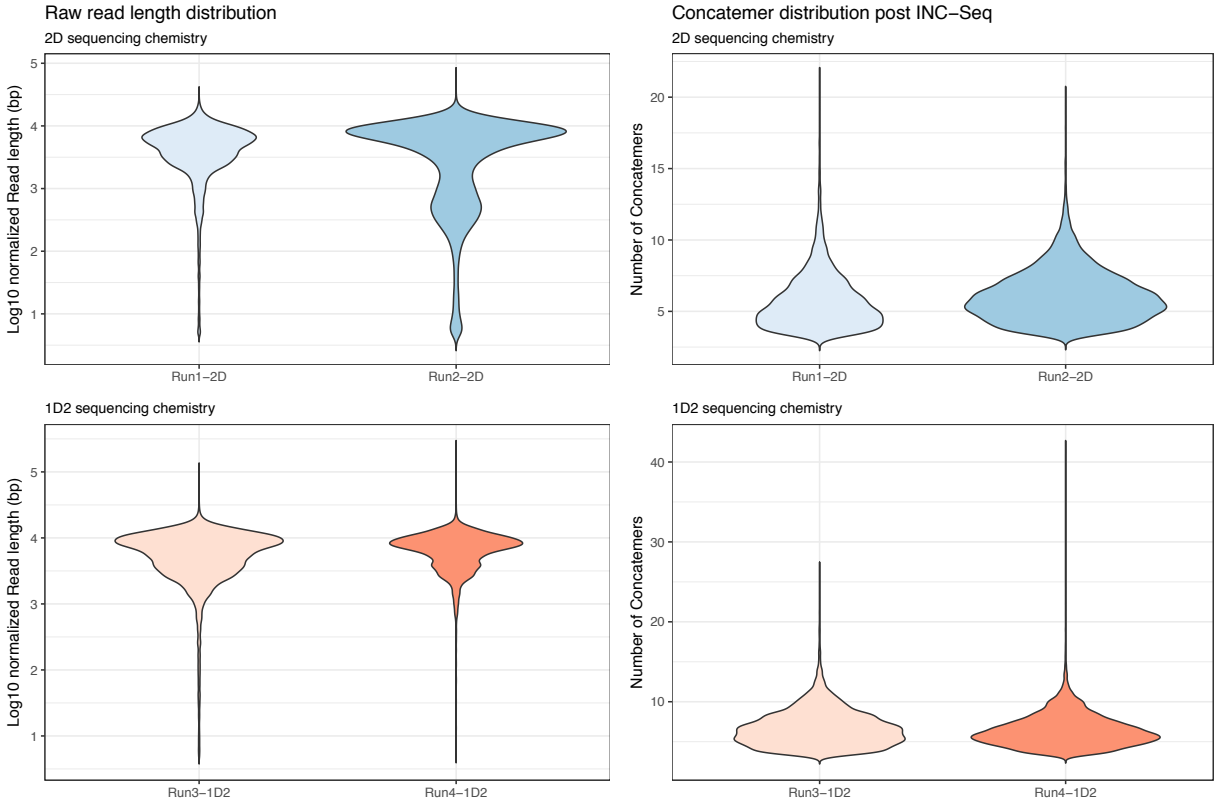


Figure S2: Violin plot of read length distributions for INC-Seq processed reads, chopSeq corrected reads after tandem repeat removal, and post size-selection of chopSeq corrected reads and size selected reads (chopSeq_SS) for one organism experiments using both 2D and 1D2 sequencing chemistry.

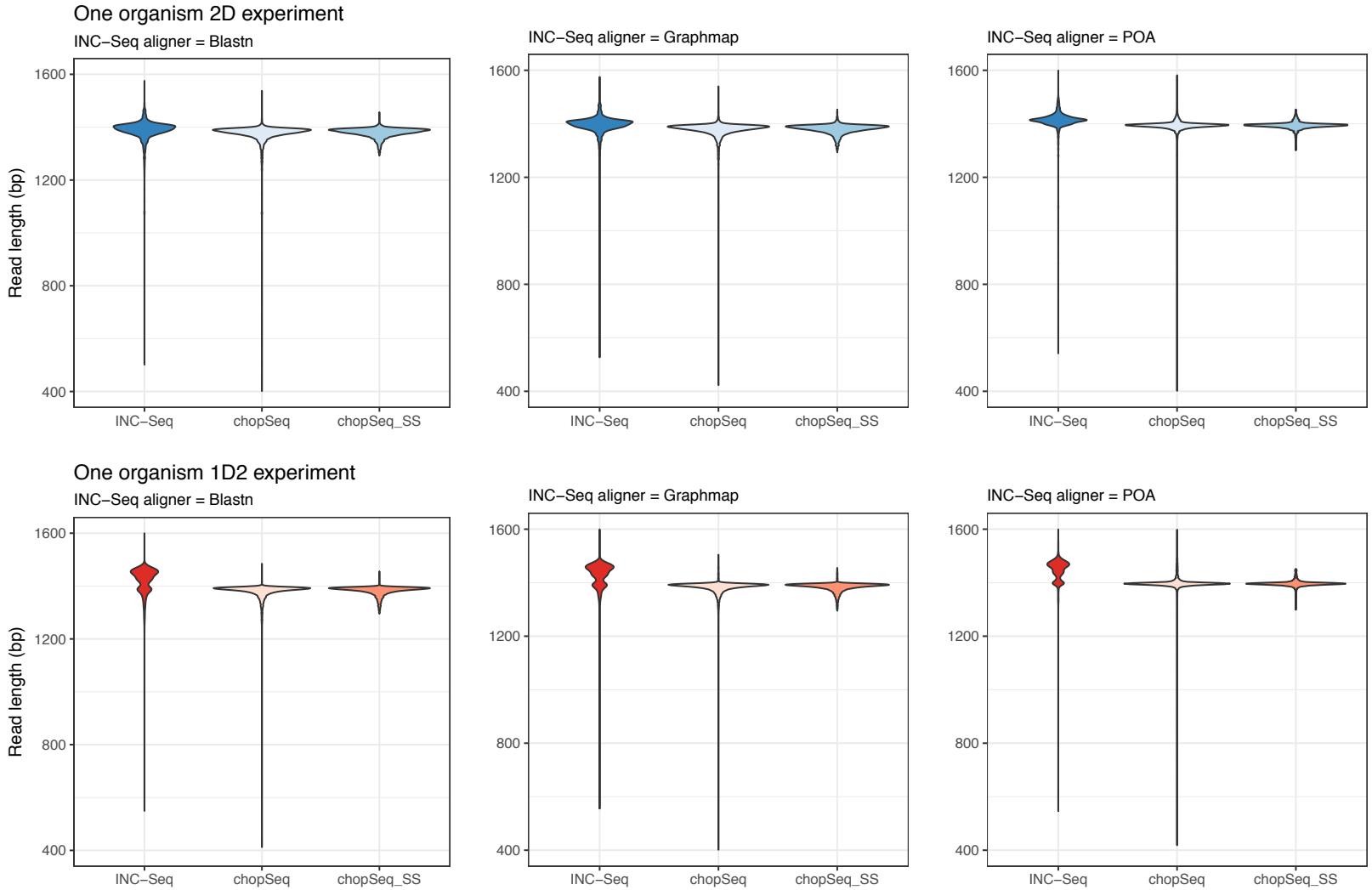


Figure S3: Consensus sequences from OTUs from Run 3 (1D2 sequencing chemistry, INC-Seq aligner: blastn) were combined with reference sequences and aligned using muscle (default parameters) and Neighbor-Joining tree using Jukes-Cantor model was constructed in Geneious (version 8) using 100 bootstraps. Reference sequences are labelled in green and OTUs are labelled in red.

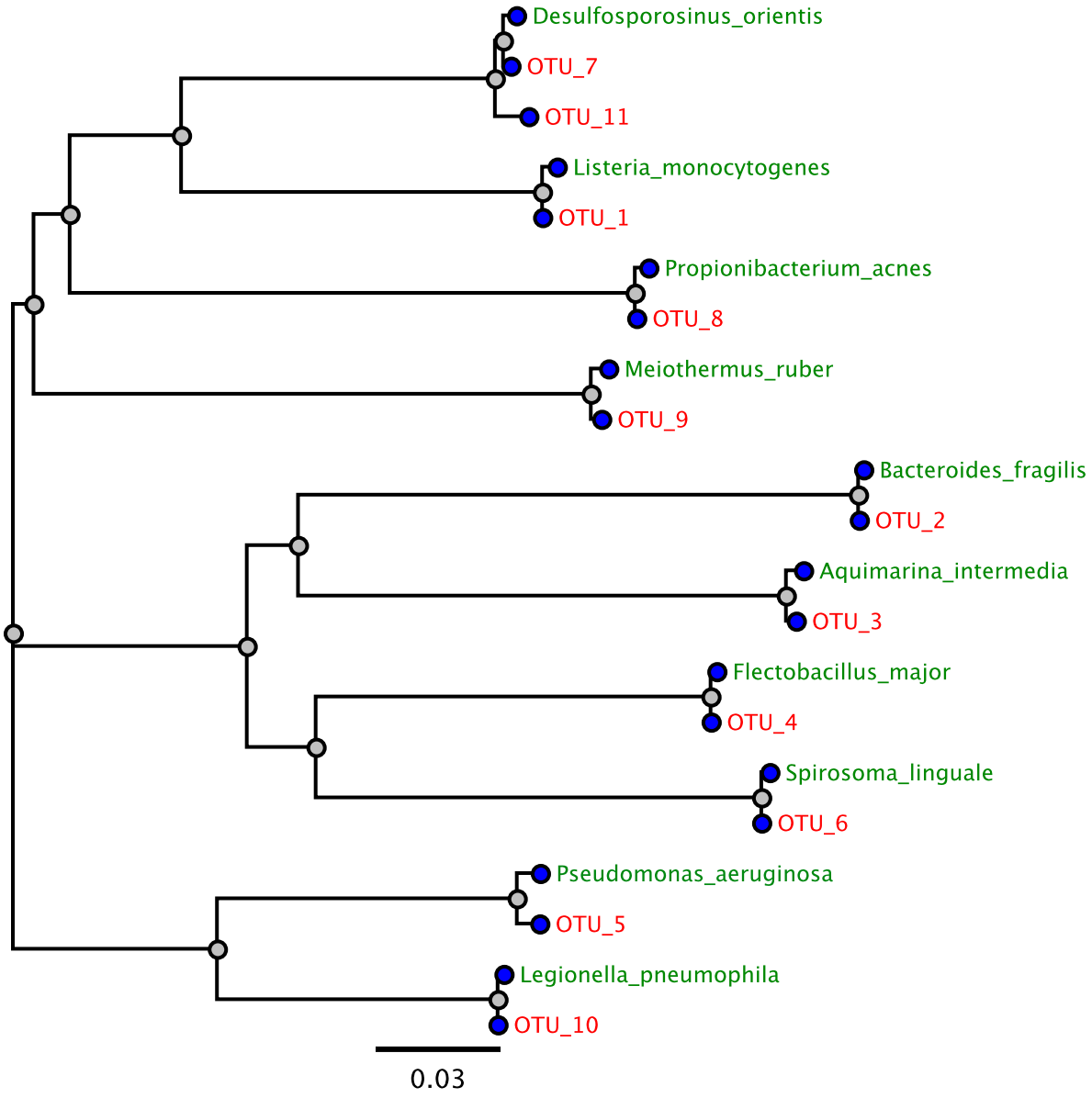


Figure S4: An overview of the muscle alignment (wrapped view) used to construct the phylogenetic tree in Figure S3.

Figure S5: Relative abundance of OTUs for one (A, B, C, G, H, I) and ten organism experiments (D, E, F, J, K, L) for 2D (blue data points) and 1D2 (red data points) experiments with the use of blastn (A, G, D, J), Graphmap (B, H, E, K), and poa (C, I, F, L) aligners during INC-Seq processing. For the one organism experiment the theoretical relative abundance is 100%, while for the ten organism experiment the theoretical relative abundance of each OTU is 10%. Spurious OTUs are indicated for each panel.

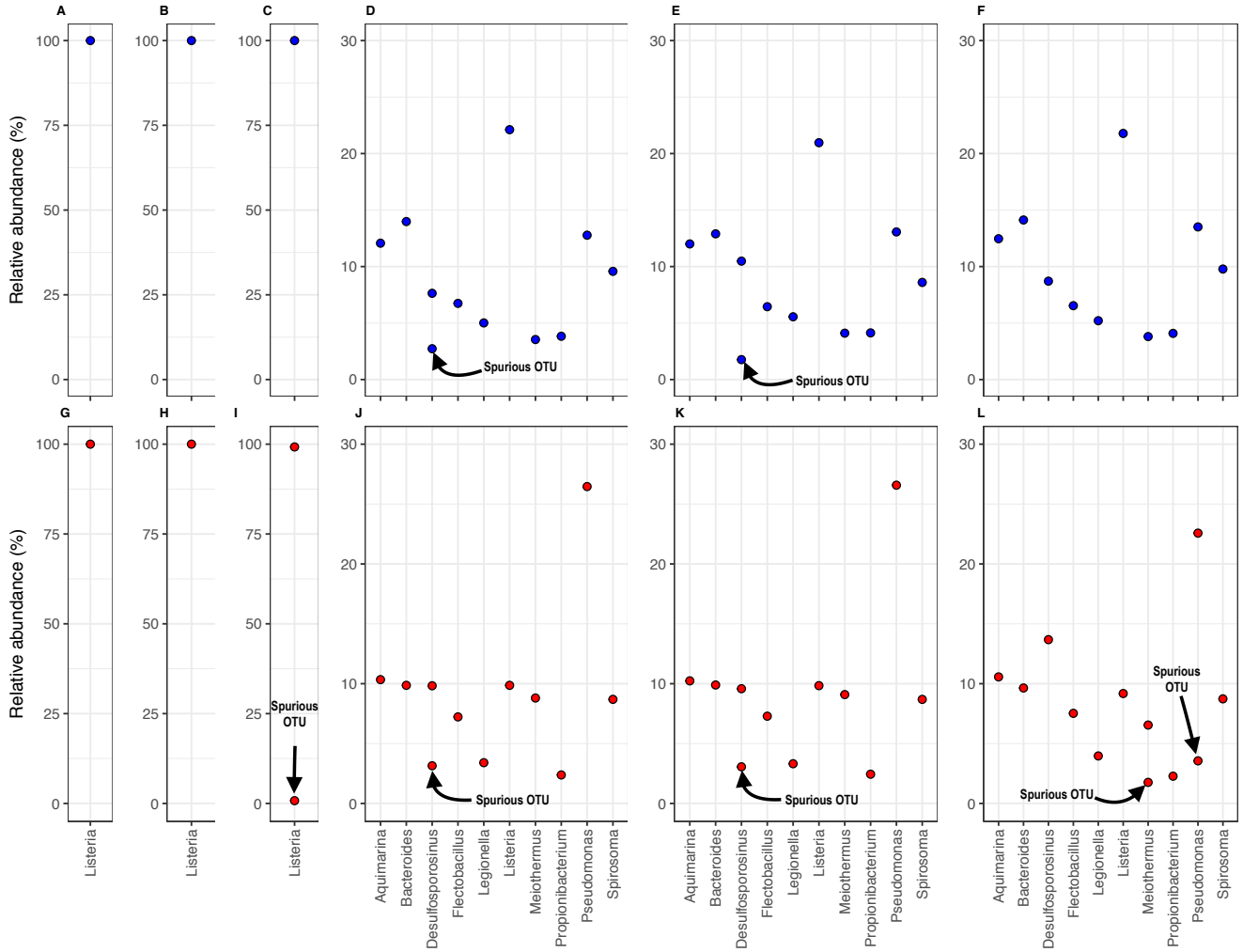
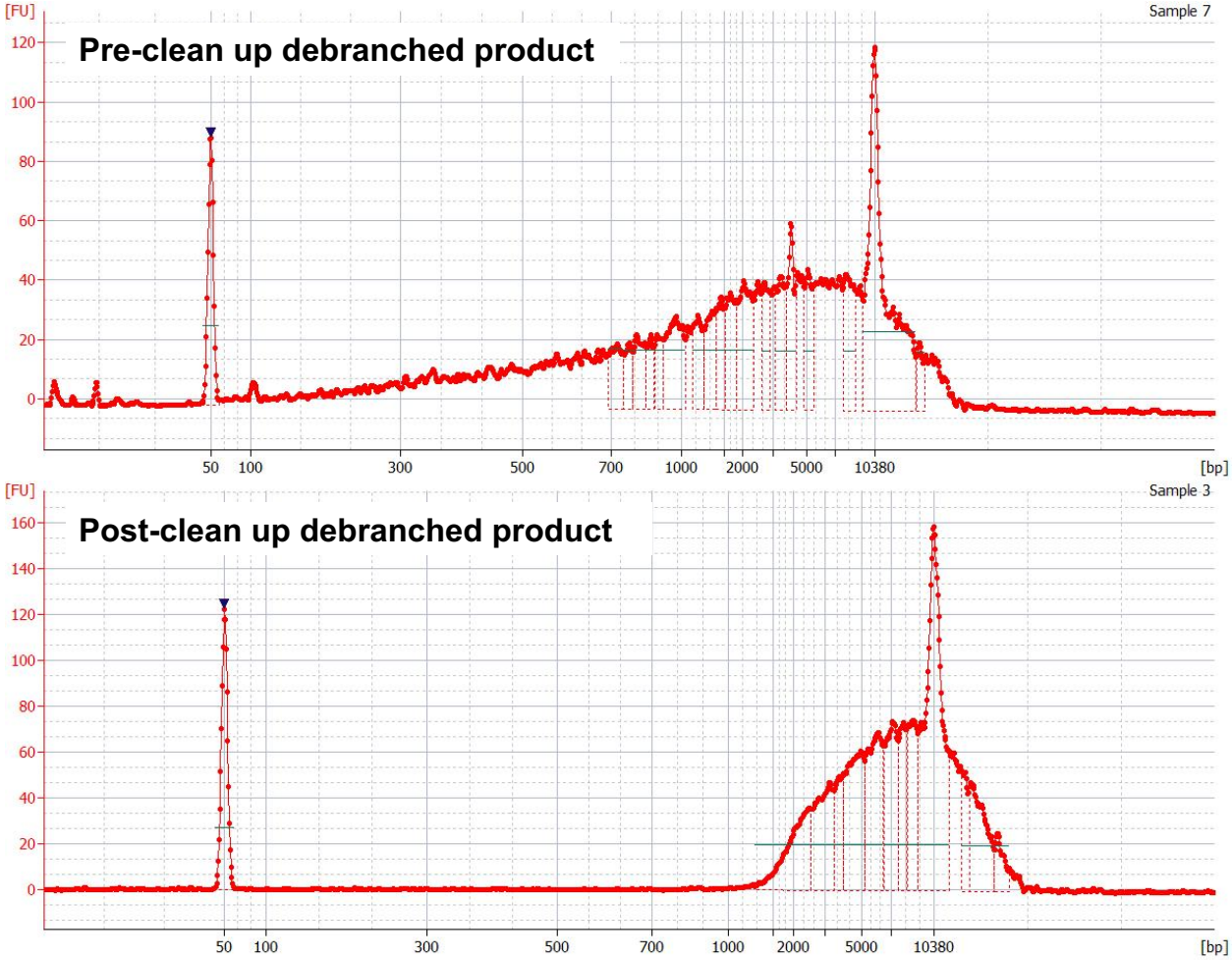


Figure S6: Example Bioanalyzer traces of de-branched product pre- and post-clean up. The post-cleanup product was then treated for gap filling and DNA damage repair prior to being used for library preparation of sequencing according to ONT protocol for 2D and 1D2 chemistries.



Supplementary text:

Protocol for sample preparation for NanoAmpli-Seq workflow.

Consumables required:

1. Sterile Filtered pipette tips from any vendor
2. 0.2 ml PCR grade tubes from any vendor
3. Wide Bore Filtered pipette tips from any vendor
4. Qubit dsDNA HS Assay kit. Vendor: ThermoFisher Scientific. Catalog number: Q32851.
5. Primers for PCR amplification of 16S rRNA gene with phosphorylated 5' ends can be ordered from any provider:
    a. Forward primer: 8F: [PHO] AGRGTTTGATCMTGGCTCAG
    b. Reverse primer: 1387R: [PHO] GGGCGGWGTGTACAAGRC
6. Master mix for PCR amplification: Q5® High-Fidelity 2X Master Mix. New England Biolabs, Inc. Catalog number: M0492S
7. Nuclease-free Water. Vendor: New England Biolabs, Inc. Catalog number: B1500S
8. HighPrep™ PCR paramagnetic bead solution. Vendor: MAGBIO. Catalog number: AC-60050
9. 70% ethanol (prepared from denatured ethanol).
10. Nuclease-free Water. Vendor: New England Biolabs, Inc. Catalog number: B1500S
11. Magnetic stand for 1.5 ml tubes. MagStrip Magnet Stand 10. Vendor: MAGBIO. Catalog number: MBMS-10
12. Blunt/TA ligase Master Mix. Vendor: New England Biolabs, Inc. Catalog number: M0367S
13. HighPrep™ PCR paramagnetic bead solution. Vendor: MAGBIO. Catalog number: AC-60050
14. DNA LoBind Tubes, 1.5 ml. Vendor: Eppendorf. Catalog number: 003018078
15. Plasmid-Safe™ ATP-Dependent DNAse. Vendor: Epicentre. Catalog number: E3101K.
16. TruePrime™ RCA kit. Vendor: Expedeon. Catalog number: 390100
17. T7 endonuclease I. Vendor: New England Biolabs, Inc. Catalog number: M0302S.
18. g-TUBE. Vendor: Covaris. Catalog number: 010145
19. NEBNext® FFPE DNA Repair Mix. Vendor: New England Biolabs Inc. Catalog number: M6630S
20. Ligation Sequencing Kit 1D. Catalog number: SQK-LSK108 and/or Ligation Sequencing Kit 1D2. Catalog number: SQK-LSK308. Vendor: Oxford Nanopore Technologies
21. Flow Cell (R9.4) and/or Flow Cell (R9.5). Vendor: Oxford Nanopore Technologies

Equipment required:

1. PCR thermocycler from any vendor
2. PCR hood
3. Thermal mixer with appropriate blocks from any vendor
4. Pipettes from varying volumes range from any vendor
5. Centrifuge for 2 ml and 0.2 ml tubes from any vendor
6. MinION™ Mk1b device and compatible personal computer

Step 1: PCR amplification of 16S rRNA gene.

1. Combine the following components using volumes below or as appropriate for your experiment in a PCR tube.
    a. Q5® High-Fidelity 2X Master Mix: 12.5 µl
    b. Forward primer (10 pmol): 0.8 µl
    c. Reverse primer: (10 pmol): 0.8 µl
    d. Template DNA: 1 µl
    e. Nuclease-free water: 9.9 µl.

2. Amplify PCR reaction mix at the following PCR conditions:

Segment	Temp (°C)	Time (sec)	Cycles
Initial denaturation	98	00:30	1
Denaturation	98	00:05	20
Annealing	59	00:10	
Extension	72	00:35	
Final extension	72	02:00	1
Hold	8	00:00	1

Step 2: PCR product clean up.

1. Follow vendor instructions (MAGBIO) to clean PCR product at 0.45x bead ratio as described at the following url: http://www.magbiogenomics.com/image/data/Literature/Protocols/HighPrep%20PCR%20Protocol.pdf

Step 3: PCR product concentration estimation.

1. Follow vendor instructions (ThermoFisher) to determine concentration of cleaned PCR product using Qubit™ dsDNA HS Assay kit as described at the following url: https://assets.thermofisher.com/TFS-Assets/LSG/manuals/Qubit_dsDNA_HS_Assay_UG.pdf

Step 4: Self-ligation for formation of plasmid-like structure.

1. Dilute PCR product from step 2 to 1 ng/µl using nuclease-free water
2. Mix 55 µl of diluted PCR product with 5 µl of Blunt/TA ligase Master Mix in 0.2 ml PCR grade tubes
3. Gently mix by flicking the tube a few times
4. Centrifuge tube for 10 seconds
5. Incubate for 10 min at 15°C in a PCR thermocycler
6. Gently mix by flicking the tube a few times
7. Centrifuge tube for 10 seconds
8. Incubate for another 10 min at room temperature.

Step 5: Reverse phase cleanup.

1. Vortex HighPrep™ PCR paramagnetic bead solution
2. Transfer 50 µl of bead solution into clean DNA LoBind 1.5 ml tube
3. Place tube on magnetic rack until beads separate from liquid

4. While the tube is on the magnetic rack, remove 25 µl of liquid from the tube using sterile pipette tip, taking care not to disturb the beads
5. Remove tube from magnetic rack and gently vortex to resuspend the beads.
6. Add the self-ligation mix from step 4 to concentrated beads at a ratio of 0.35x bead ratio.
7. Incubate the mixture for 3 minutes at room temperature.
8. Place tube on magnetic rack
9. Allow the beads to separate from the liquid. The beads contain long linear amplicons (potentially chimeric amplicons) while the liquid contains short linear amplicon and plasmid like structures. Carefully remove the clear liquid from the tube and move it to step 6.

Step 6: Plasmid and short amplicon clean-up.

1. Follow vendor instructions (MAGBIO) to clean plasmid and short amplicon mix at 0.45x bead ratio as described at the following url: http://www.magbiogenomics.com/image/data/Literature/Protocols/HighPrep%20PCR%20Protocol.pdf

Step 7: Removal of linear molecules from plasmid mix.

1. Combine 42 µl of clean product from step 6 (eluted in nuclease-free water) with 2 µl of 25 mM ATP solution, 5 µl of 10X reaction buffer and 1 µl of Plasmid-Safe DNASe in 0.2 ml PCR grade tube. The last three ingredients are provided with the Plasmid-Safe™ ATP-Dependent DNAse kit.
2. Incubate at 37°C for 15 minutes in a PCR thermocycler
3. Clean product as described in Step 2
4. Determine the concentration of DNA in cleaned product as described in Step 3.

Step 8: Rolling Circle Amplification (RCA).

Perform RCA in triplicate for each sample and include negative controls using nuclease-free water instead of cleaned product from step 7. Below are reaction conditions and volumes for a single RCA reaction:

1. Combine 2.5 µl of cleaned product from step 7 with 2.5 µl of Buffer D (provided with TruePrime™ RCA kit) in 0.2 ml PCR grade tube and incubate at room temperature for 10 minutes
2. While sample is being incubated, prepare the amplification mix consisting of 9.3 µl of nuclease-free water, 2.5 µl of reaction buffer, 2.5 µl of dNTPs, 2.5 µl of Enzyme 1 and 2.5 µl of Enzyme 2. All ingredients are included in the TruePrime™ RCA kit
3. After 10 minutes, add 2.5 µl of Buffer D (provided with TruePrime™ RCA kit) and prepared amplification mix to the tube
4. Mix by pipetting and incubate tube at 30°C for 150 minutes.
5. Follow Step 3 to determine the concentration of DNA in samples and negative controls.

Step 9: Enzymatic de-branching.

1. Combine triplicate reactions from step 8 into a single 0.2 ml PCR grade tube
2. Mix thoroughly to prepare single RCA product per sample
3. Combine 65 µl of RCA product with 2 µl of T7 endonuclease I and incubate at room temperature for 5 minutes.

Step 10: Mechanical fragmentation.

1. Transfer enzymatically de-branched RCA product into g-TUBE using wide bore pipette tips
2. Centrifuge at 1800 rpm for 4 minutes or until entire reaction mix passes through the fragmentation hole
3. Reverse the g-TUBE and centrifuge again at 1800 rpm for 4 minutes or until entire reaction mix passes through the fragmentation hole
4. Clean product with as described in Step 5, with the exception that the clear liquid after bead separation is discarded and bead bound DNA is eluted according to vendor outlined protocol
5. Repeat step 9 (i.e. Enzymatic de-branching) on g-TUBE fragmented product again.
6. Clean product as described in Step 5, with the exception that the clear liquid after bead separation is discarded and bead bound DNA is eluted according to vendor outlined protocol.

Step 11: DNA Damage repair.

1. Combine 53.5 µl of product from Step 10 with 6.5 µl of FFPE DNA Repair Buffer and 2 µl of NEBNext FFPE Repair mix in a 0.2 ml tube
2. Incubate at 20°C for 15 minutes
3. Clean product as described in Step 2
4. Quantify DNA concentration as described in Step 3.

Step 12: **DNA library preparation and nanopore sequencing.**

1. Prepare 2D or 1D2 libraries for nanopore sequencing according to the protocols described for the SQK-LSK108 or SQK-LSK308 kits by Oxford Nanopore Technologies.
2. Primer appropriate flow cell using protocols outlined by Oxford Nanopore Technologies.
3. Finally, sequence the prepare libraries on appropriate flow cells as outlined by Oxford Nanopore Technologies.