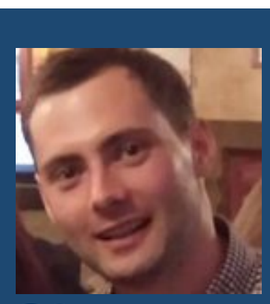
NanoAmpli-Seq: A workflow for amplicon sequencing from mixed microbial communities on the nanopore sequencing platform

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

[1] University of Glasgow, UK & [2] Northeastern University, USA



@Szymonome

Introduction: Short-read DNA sequencing platforms are restricted by the size of the DNA molecule that can be successfully[1]. This is particularly limiting in the context of 16S rRNA gene sequencing, where short reads can rarely be classified to the genus level and never provide a species-level resolution[2]. Long read sequencing platforms (i.e. Oxford Nanopore Technology's (ONT) MinION™) can allow for analysis of entire SSU rRNA genes. However, application of ONT platforms for amplicon sequencing of conserved marker genes has been limited due to higher raw error rates. In this study, we describe the development and validation of sample preparation and data analyses approach, (i.e., NanoAmpli-Seq)[3] that builds and expands on the previously described Intramolecular-ligated Nanopore Consensus Sequencing (INC-Seq)[4] for profiling of complex microbial communities using full-length 16S rRNA gene sequencing[3].

Materials and methods

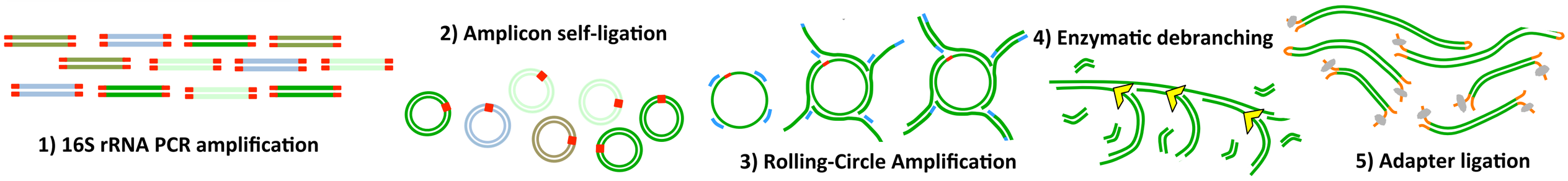


Figure 1. Overview of the NanoAmpli-Seq protocol for 16S rRNA gene analysis: 1) PCR amplification, 2) plasmid-like amplicon self-ligation, 3) Rolling Circle-amplification, 4) enzymatic branching of hyperbranched amplicons, 5) 2D and 1D^2 adapter ligation followed by nanopore sequencing.

Single organisms and ten organism mock communities were constructed by amplifying the near full length of the 16S rRNA gene from genomic DNA of included organisms using set of 8F and 1387R, both with 5' phosphorylated primers. Rolling circle amplification (RCA) of amplicon pool was performed using PrimPol, involving, in silico primer synthesis for RCA, and a combination of g-TUBE and T7 endonuclease I for mechanical and enzymatic debranching of RCA molecules. Sequenced reads were processed through the INC-Seq workflow to construct INC-Seq consensus reads. A vast majority of the INC-Seq consensus reads were incorrectly oriented and contained inserts between the forward and reverse primer. Read re-orientation and insert removal was removed by developing and applying chopSeq algorithm. Finally, to further correct for the residual errors we developed read-partitioning based de novo clustering and consensus calling approach to obtain high quality full-length 16S rRNA sequences.

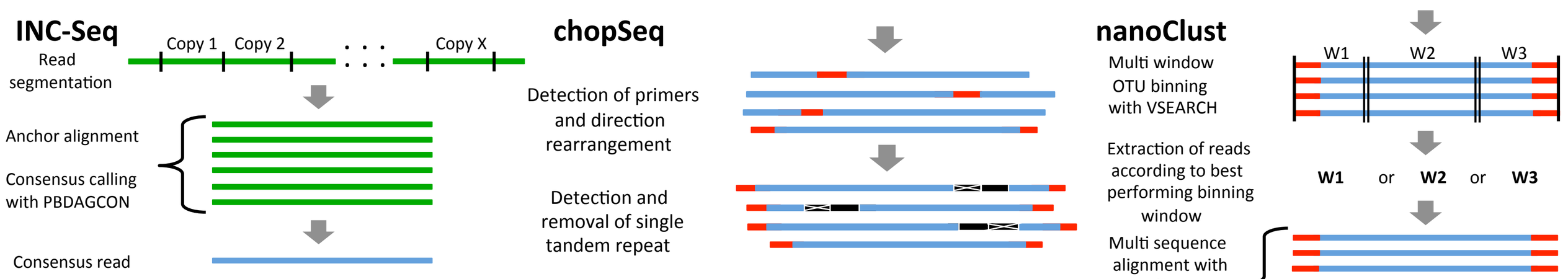


Figure 2. INC-Seq consensus reads were subject to chopSeq based read reorientation, insert removal, and size selection to retain reads between 1300-1450bp. chopSeq corrected reads were subject to partitioning followed by VSEARCH clustering to identify optimal partition size that generates maximum number of OTUs (without singletons). MAFTT-G-INS-i is then used pairwise alignment of a subset of full length reads from best performing partition OTU bin OTU consensus sequence is created from the alignment.

Results and conclusions: The NanoAmpli-Seq protocol involves important modifications to INC-Seq's amplicon library preparation protocol by significantly reducing sample processing time (up to 70%). Further, NanoAmpli-Seq includes an algorithm, i.e. chopSeq, for correction of INC-Seq consensus reads and nanoClust for read partitioning-based *de novo* sequence clustering and within cluster consensus calling of chopSeq corrected reads. Through these improvements, NanoAmpli-Seq accurately estimates the diversity of tested mock communities and provides an average consensus sequence accuracy of 99.5±0.08% for both 2D and 1D^2 sequencing chemistry on the nanopore sequencing platform (Fig.3). Nearly all-residual errors (~0.5%) in NanoAmpli-Seq processed reads originate from deletions in homopolymer regions.

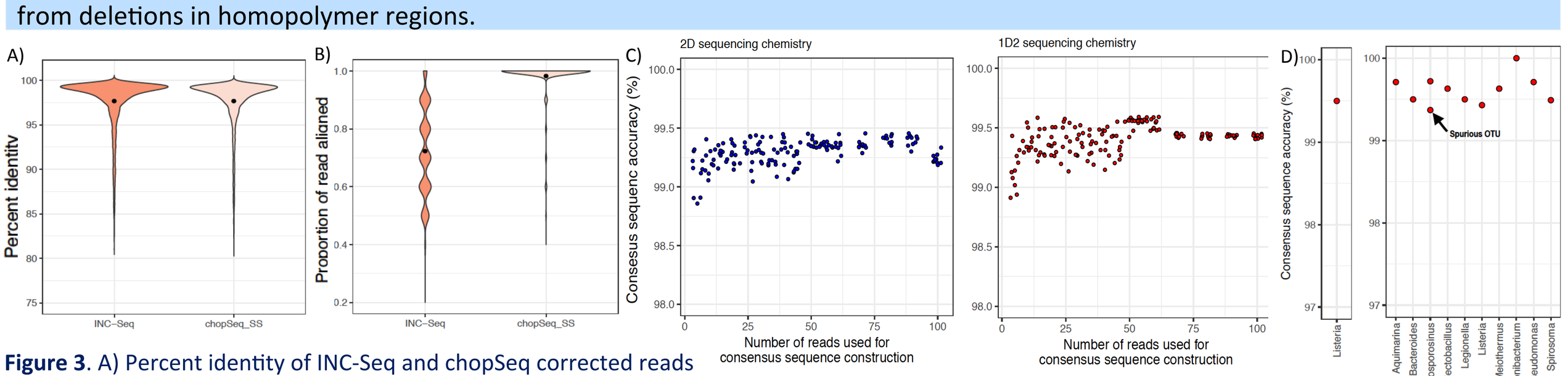


Figure 3. A) Percent identity of INC-Seq and chopSeq corrected reads was 97-98% with no significant difference between these two steps. B) However, proportion of read aligned for INC-Seq and chopSeq reads differ significantly indicating that reads direction rearrangement worked successfully. C) Optimization of nanoClust for number of reads used for consensus sequence construction did not improve sequence accuracy beyond 50 reads. D) The use 50 reads for consensus construction of OTU consensus sequences results in accurate estimate of mock community diversity and high sequence accuracy.

References:

1) Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J. 2012;6. 2) Wommack KE, Bhavsar J, Ravel J. Metagenomics: Read Length Matters. Applied and Environmental Microbiology. 2008;74(5):1453-63. 3) Calus, S.T., Ijaz, U., Pinto, A.J. NanoAmpli-Seq: A workflow for amplicon sequencing from mixed microbial communities on the nanopore sequencing platform. biorxiv. 2018. 4) Li C, Cheng KR, Boey EJ, Ng AH, Wilm A, Nagarajan N. INC-Seq: accurate single molecule reads using nanopore sequencing. (2047-217X).

Contact: **E-mail:** a.pinto@northeastern.edu
Twitter: @watermicrobe **Web:** www.pintolab.com

Northeastern University of Glasgow