SeqScreen-Nano: functional and taxonomic characterization of long read metagenomic data

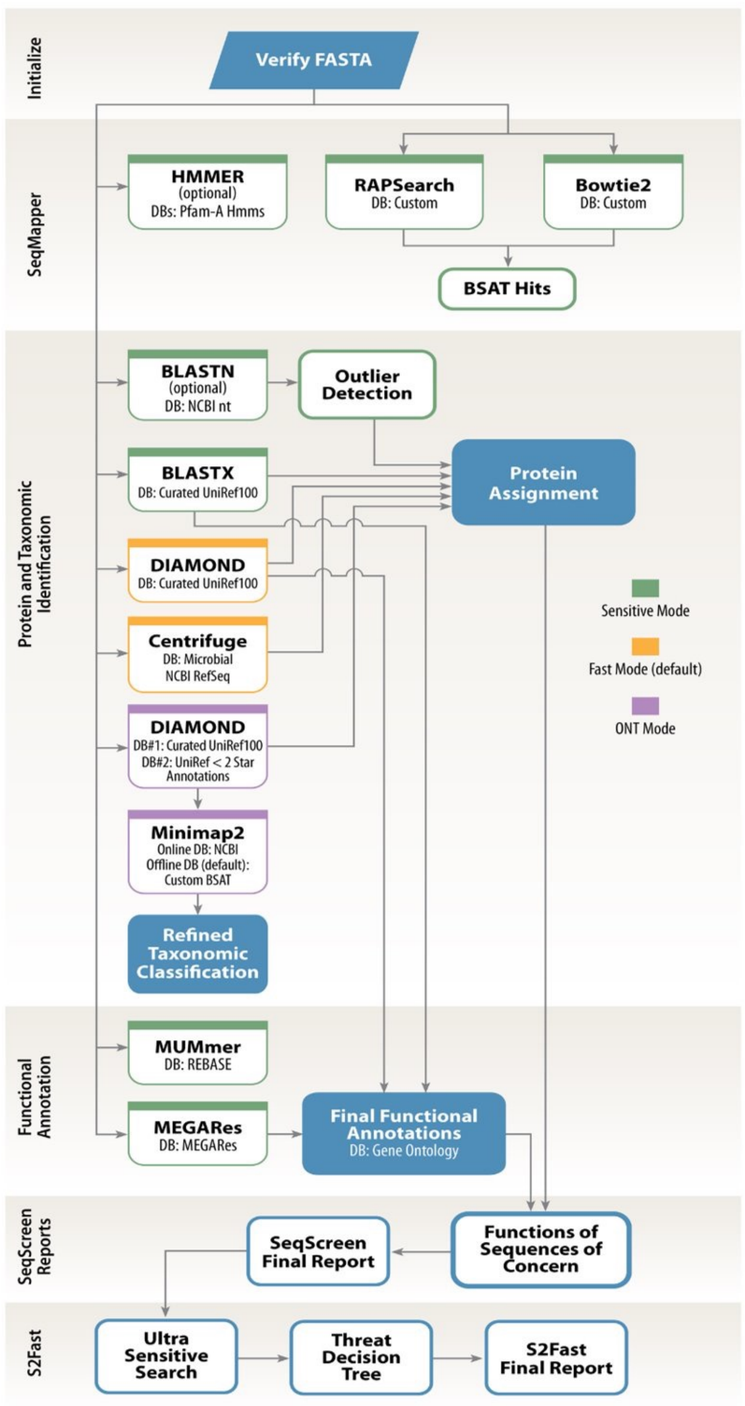
Advait Balaji[1], Yunxi Liu[1], **Michael G. Nute**[1], Bingbing Hu[1], Anthony D. Kappell[2], Gene D. Godbold[3], Krista L. Ternus[2] and Todd J. Treangen[1]

[1]Department of Computer Science, Rice University, Houston, TX, [2]Signature Science, LLC, 8329 North Mopac Expressway, Austin TX,
[3]Signature Science, LLC, 1670 Discovery Drive, Charlottesville, VA

ABSTRACT

Affordable long read sequencing has enabled a wide variety of metagenomic analysis tasks, from obtaining high quality genome assemblies to identifying structural variants. Though long reads offer better resolution, accurate assignment of functional and taxonomic labels to ONT sequences remains an open challenge. Here we present a solution to this challenge, building upon SeqScreen and adapting it to identify Functions of Sequences of Concern (FunSoCs) on ONT data. The taxonomic assignment over the entire read is carried out using a combination of a majority voting heuristic and greedy weighted min-set cover approach and refined using a reference-based approach that uses breadth of coverage information to separate closely related species in the sample. We show that on simulated and synthetic metagenomic data, SeqScreen-Nano can identify Open Reading Frames (ORFs) across the length of raw ONT reads and use it to accurately assign functional and taxonomic labels.

PIPELINE



SeqScreen-Nano: Reference Inference

1. Collect all candidate taxids and filter for abundance at species level (0.2%)

1st read mapping

2. Download reference genomes for set of taxids and map reads to each reference individually and calculate coverage score (Calculated/Expected)

2nd read mapping

3. Concatenate references with greater than 0.7 coverage score in step 2 and re-map.

4. Call genomes present based on ANI calculations and difference of coverage scores from both stages of mapping.

Challenges in functional and taxonomic characterization of long read data

Challenge #1

Error in long reads causes frameshifts and no tool dd for ORF identification

Solution:

Carefully select DIAMOND parameters and interval trees to identify and extract ORFs efficiently

Challenge #2

Accurate protein-based read level taxonomic prediction is challenging dure to presence of multiple ORFs in a read.

Solution:

Combination of greedy majority vote and min-set cover enables more species level annotations at read level

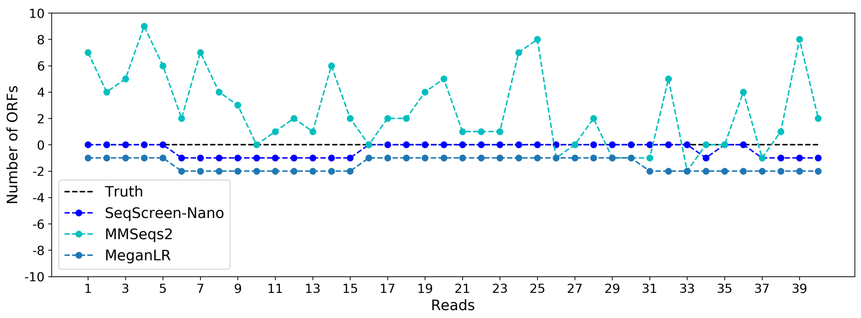
Challenge #3

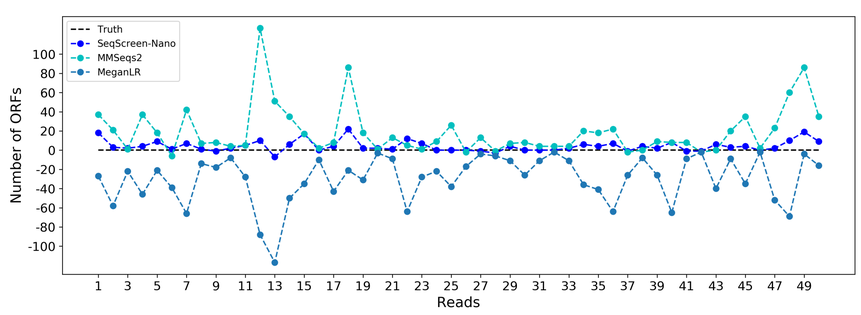
Accurate prediction of low-abundance microbes in the sample

Solution:

Reference-based inference pipeline that uses breadth of coverage information and ANI to ascertain presence/absence of a microbe.

Predicting number of ORFs in a nanopore read





Deviation from truth:

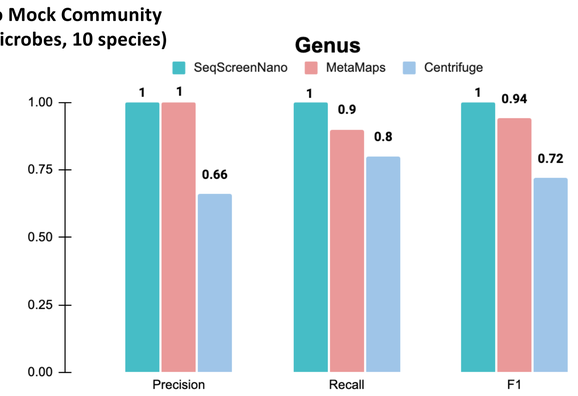
Tool	Median	Mean	Stdev
SeqScreen-Nano	0	0.37	0.49
MMSeqs2	2	2.97	0.6
MeganLR	1.5	1.5	0.5

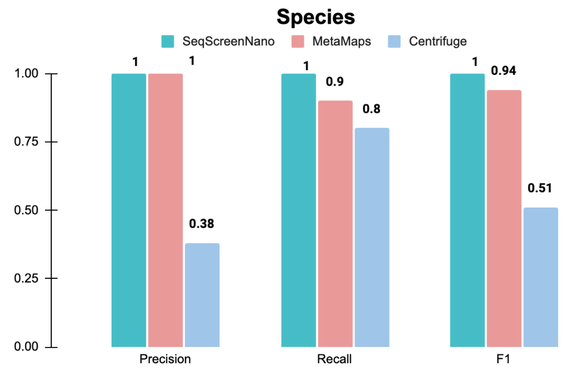
Devia tion from truth:

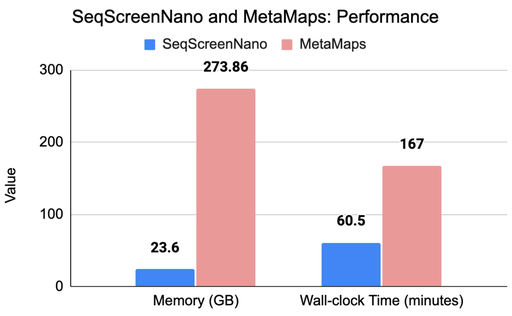
Tool	Median	Mean	Stdev
SeqScreen-Nano	3	4.8	5.27
MMSeqs2	8.5	19.56	25.33
MeganLR	26	29.96	24.63

Refence genome inference from SeqScreen-Nano (Zymo mock metagenome)

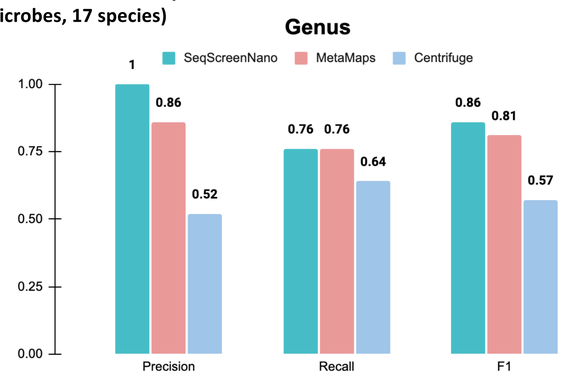
Zymo Mock Community (10 microbes, 10 species)

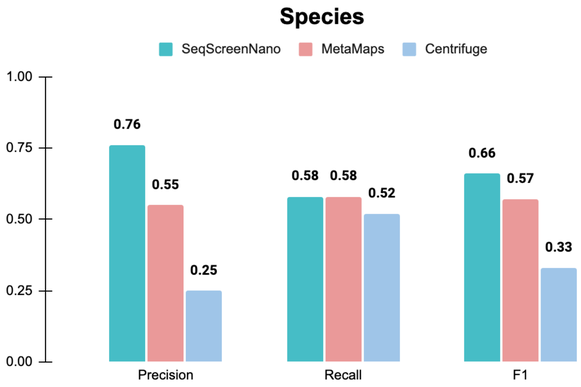






Zymo Gut Mock Community (21 microbes, 17 species)





REFERENCES

1. Balaji, A, et al. (2022) "SeqScreen: accurate and sensitive functional screening of pathogenic sequences via ensemble learning." *Genome biology* 23.1: 1-29.
2. Buchfink, B., Reuter, K., & Drost, H. G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature methods*, *18*(4), 366-368.
3.Dilthey, A. T., Jain, C., Koren, S., & Phillippy, A. M. (2019). Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. *Nature communications*, *10*(1), 1-12.
4. Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, *35*(11), 1026-1028.
5.Huson, D. H., Albrecht, B., Bağcı, C., Bessarab, I., Gorska, A., Jolic, D., & Williams, R. B. (2018). MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biology direct*, *13*(1), 1-17.

ACKNOWLEDGEMENTS

All of the co-authors were either fully or partially supported by the Fun GCAT program from the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the Army Research Office (ARO) under Federal Award No. W911NF-17-2-0089. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, ARO, or the US Government. The co-authors would like to thank ASM NGS for their generous registration and travel award.