porefile: automatic profiling of microbial communities using full-length 16S rRNA gene sequencing data

Cecilia Salazar[1,2], Ignacio Ferrés[1,2] & Gregorio Iraola[1,2,3,4]

[1]Laboratorio de Genómica Microbiana. Institut Pasteur de Montevideo, Uruguay; [2]Centro de Innovación en Vigilancia Epidemiológica. Institut Pasteur de Montevideo, Uruguay; [3]Wellcome Sanger Institute, Hinxton. United Kingdom; [4]Centro de Biología Integrativa. Universidad Mayor, Chile

csalazar@pasteur.edu.uy

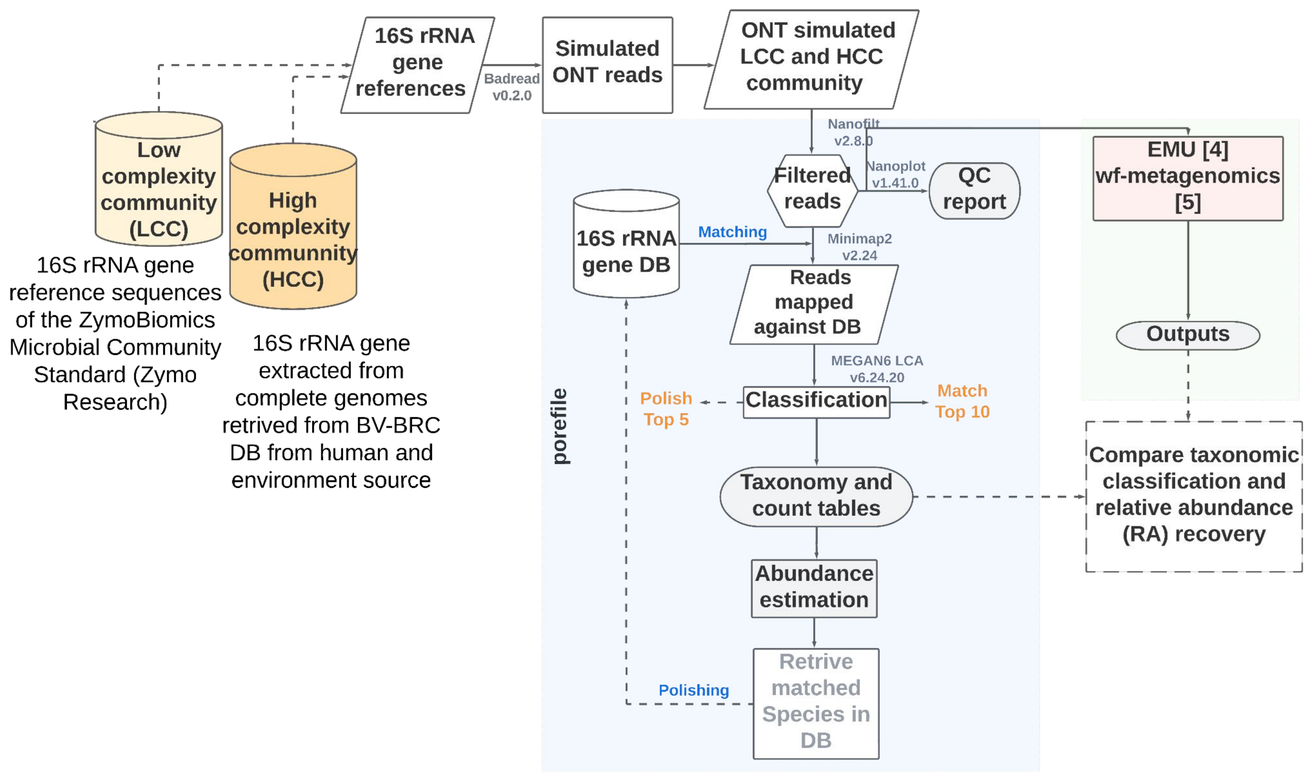
BACKGROUND

The 16S rRNA gene is a widely used taxonomic marker that has been compiled in quality-controlled databases such as the SILVA database [1]. High-throughput sequencing of the full or near-full length 16S rRNA gene using third generation sequencing approaches have proved to increase species level resolution from complex microbial communities [2]. Here we present *porefile*, a workflow that gathers different tools for read pre-processing and taxonomic profiling based on the 16S rRNA gene sequencing data generated with third generation sequencing platforms, such as Oxford Nanopore Technologies (ONT). *Porefile* sub-workflows are managed using the Nextflow system and uses a mapping strategy against the latest SILVA database and the lower common ancestor (LCA) algorithm implemented in MEGAN6 [3] to classify reads at the major taxonomic ranks. After the species-level polishing step, porefile recovers the composition of synthetic microbial communities generated with simulated ONT 16S rRNA gene sequencing data.

SOFTWARE AND DATA AVAILABILITY

microgenlab/porefile
microgenlab/simulations_16S

METHODS



RESULTS

Simulated LCC

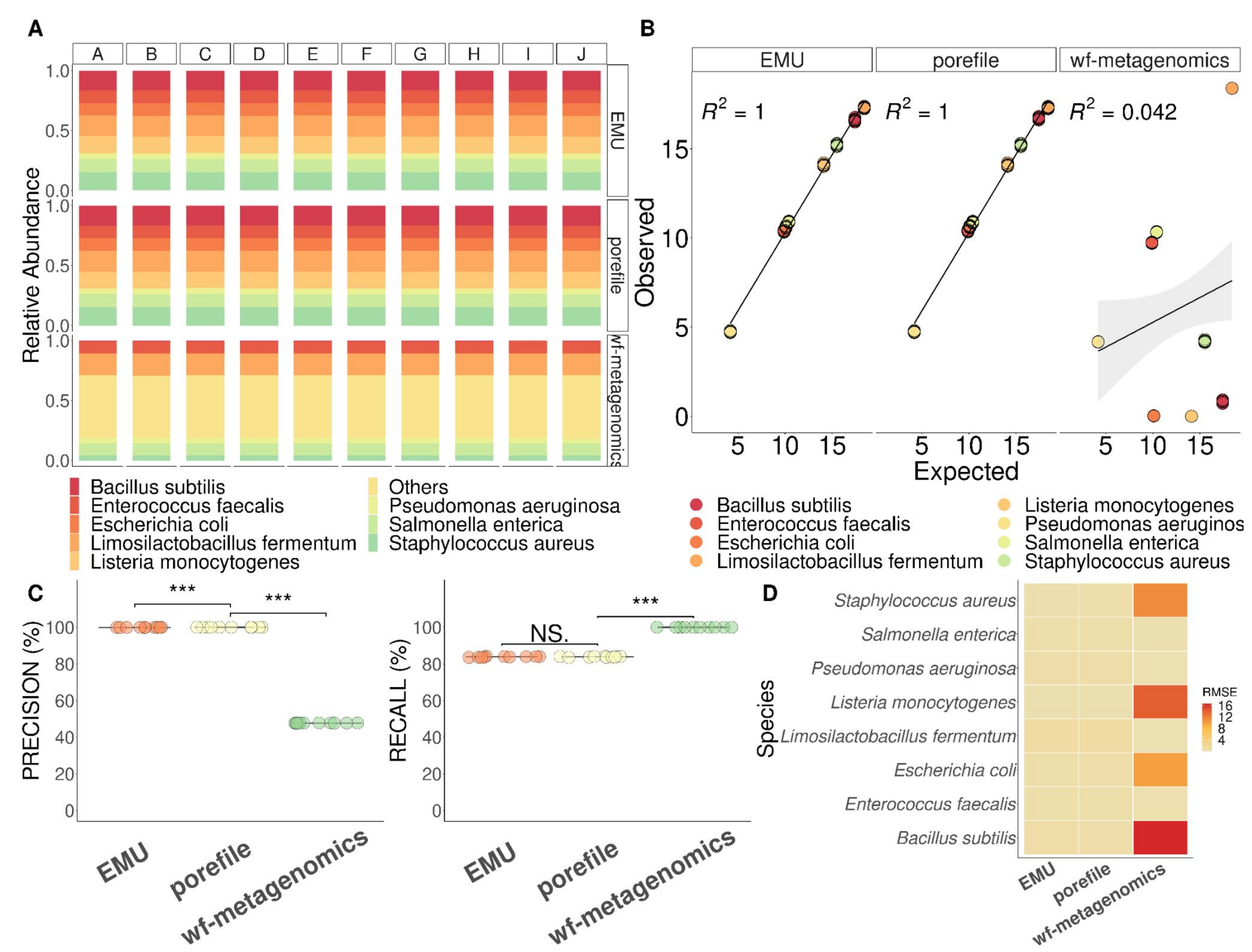


Fig. 1: Taxonomic classification of a simulated LCC dataset. **A)** RA and taxonomic classification of simulated LCC recovered from different replicates (A-J) with *porefile*, EMU and wf-metagenomics (minimap2) workflows. **B)** Linear fit between expected and observed relative abundances **C)** Precision and recall. **D)** Root-mean-square error (RMSE) of Species-level classification of the simulated LCC for each workflow.

Simulated HCC

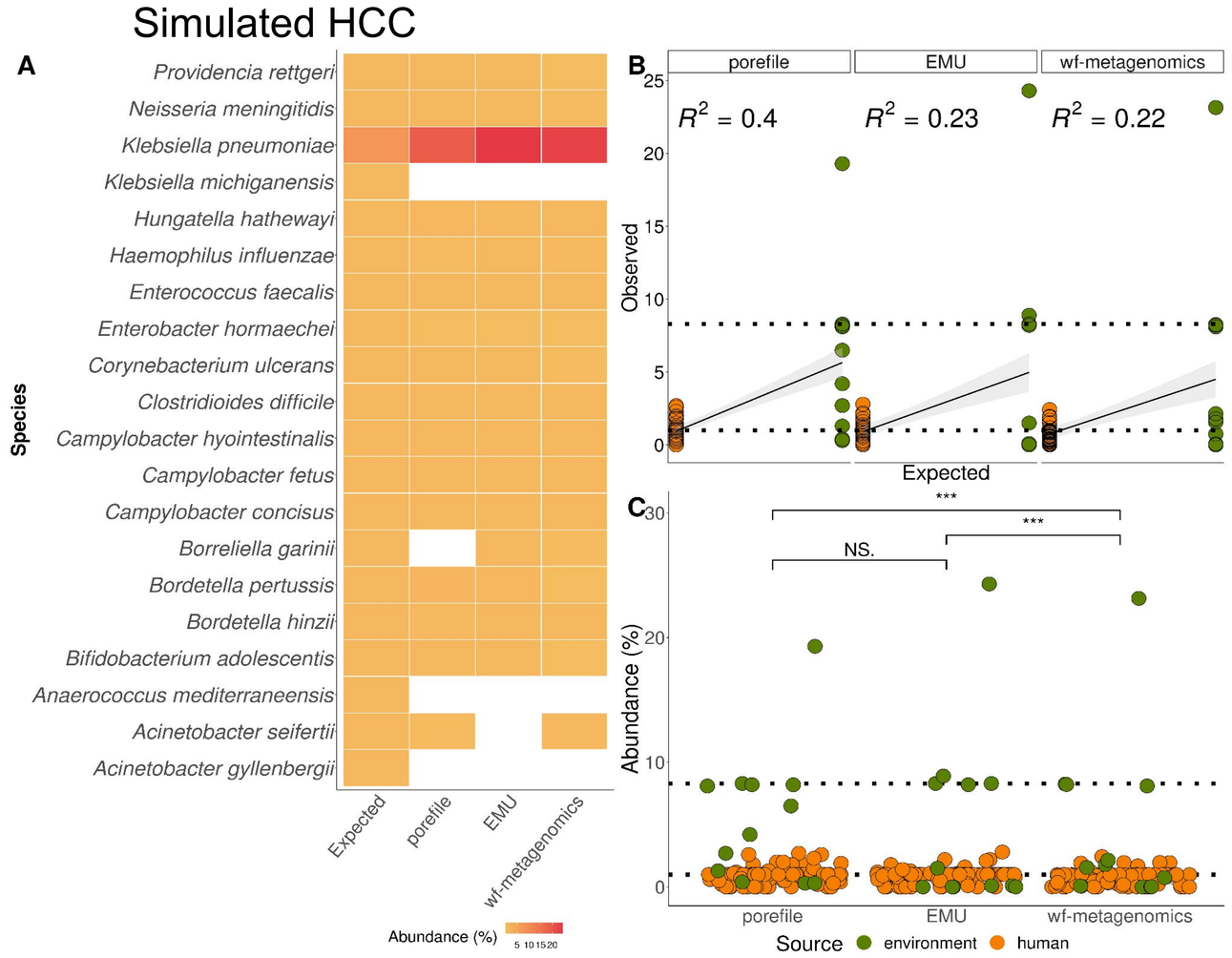


Fig. 2: Taxonomic classification of a simulated HCC dataset. **A)** Species-level detection of a sample of 20 random HCC components generated from human (n = 101 components) and environment (n = 12) sources with *porefile*, EMU and wf-metagenomics (minimap2). **B)** Linear fit between expected and observed RA. **C)** Mean RA obtained with *porefile*, EMU and wf-metagenomics. Dotted lines indicate the expected relative abundance for the human (~1%) and environment dataset (~8%).

CONCLUSION

Our results suggest that the *porefile* workflow with the polishing step module is suitable for the generation of taxonomic profiles of complex microbial communities at the species level compared to other available tools. *Porefile*, EMU and wf-metagenomics detect all LCC components, however EMU and porefile showed improved recovery of the expected RA (**Fig.1**). After generating HCCs, wf-metagenomics was able to detect more components of the synthetic microbial communities, however *porefile* showed improved recovery of the RA compared to EMU and wf-metagenomics. (**Fig. 2**).

FUTURE WORK

The next step is the generation of taxonomic profiles from 16S rRNA gene datasets using recent ONT sequencing chemistries to validate results obtained with simulated data from both low and high complexity microbial communities.

REFERENCES

[1] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2012 Nov 27;41(D1):D590–6. [2] Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun. 2019 Nov 6;10(1):5029. [3] Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007 Mar;17(3):377–86. [4] Curry, K.D., Wang, Q., Nute, M.G. et al. Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data. Nat Methods 19, 845–853 (2022). [5] https://github.com/epi2me-labs/wf-metagenomics