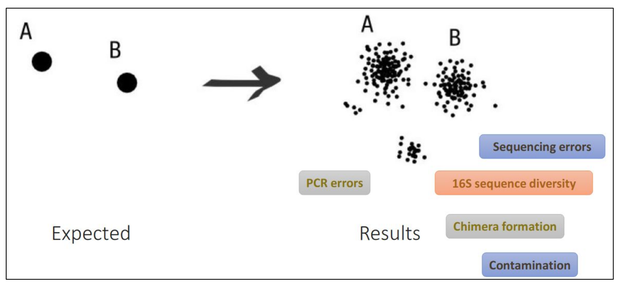
CuReSim-LoRM: a tool to simulate metabarcoding long reads

Y. Mesloub, D. Beury, F. Vandermeeren and S. Caboche

Context

- Metabarcoding = a rapid means of biodiversity assessment by identifying multiple taxa simultaneously in a sample using DNA sequencing
- Based on molecular markers conserved and shared across various taxonomic groups (*e.g.* 16S rRNA for bacteria)
- A routine application with the development of high-throughput short-read sequencing but:
  - a maximum of 500 bp-regions can be sequenced allowing one to reach, at best, the genus level and, for some bacterial taxa, not allowing one to discriminate between them

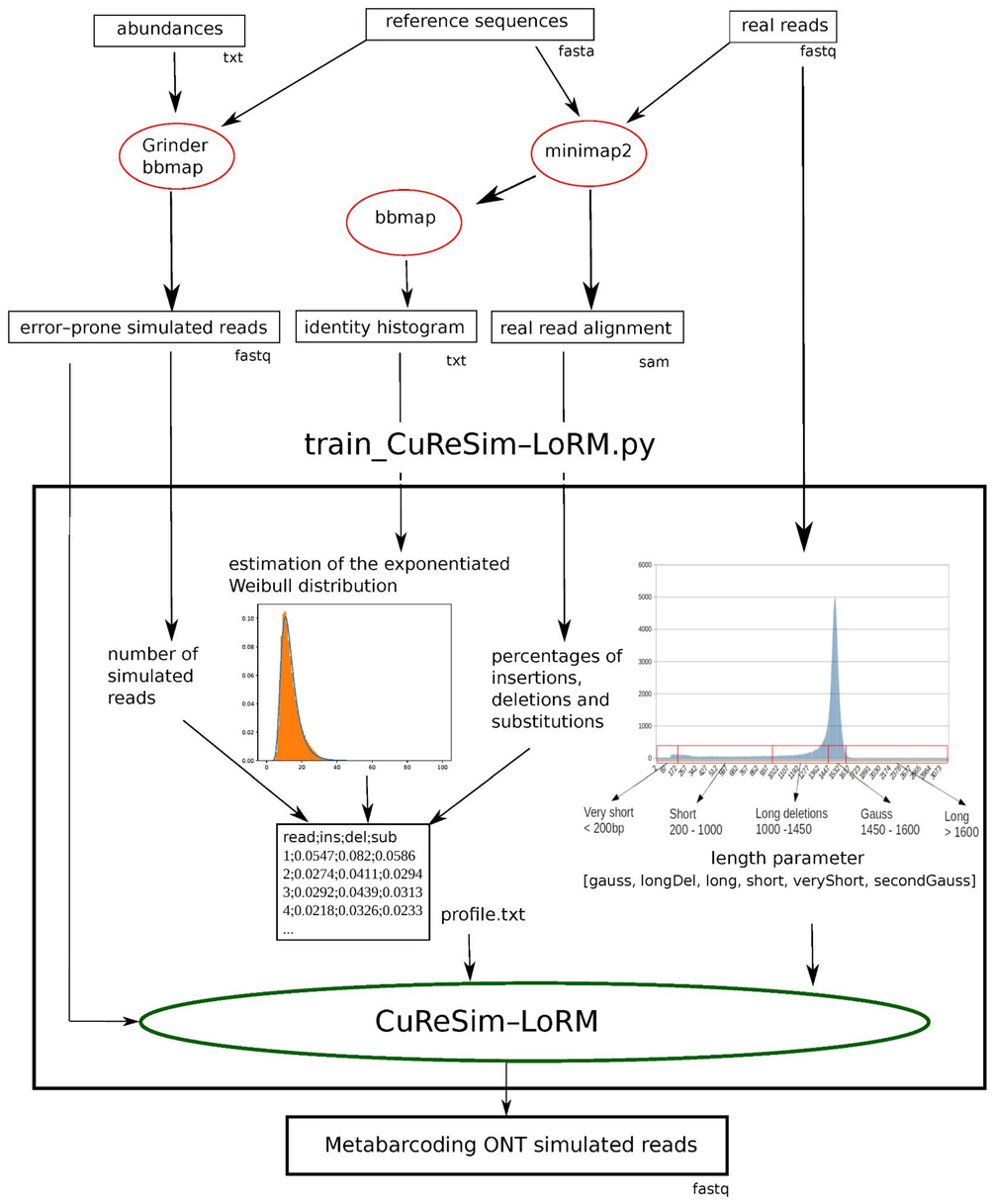
- Third-generation sequencing, such as Oxford Nanopore Technologies (ONT), produces long reads (kb-Mb)
- Obtaining the full-length ribosomal RNA gene would permit one to reach a better taxonomic resolution at the species or the strain level but:
  - ONT sequencing produces reads with high error rates (8-15%) which will introduce biases during the analysis process



- Understanding the biases introduced during the analysis allows one to better interpret the biological results and take care of conclusions drawn from metabarcoding experiments
- To benchmark an analysis process, the ground truth, *i.e.* the real composition of the microbial community, needs to be known

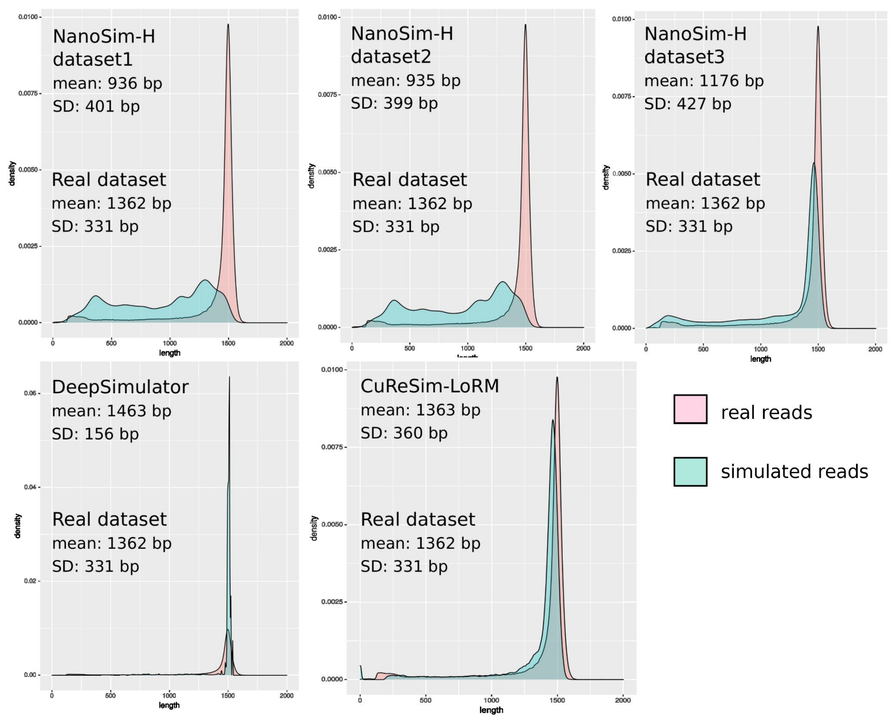
Need a tool to simulate metabarcoding long reads in order to evaluate biases in the databases and the bioinformatics analyses dealing with metabarcoding data

Methods



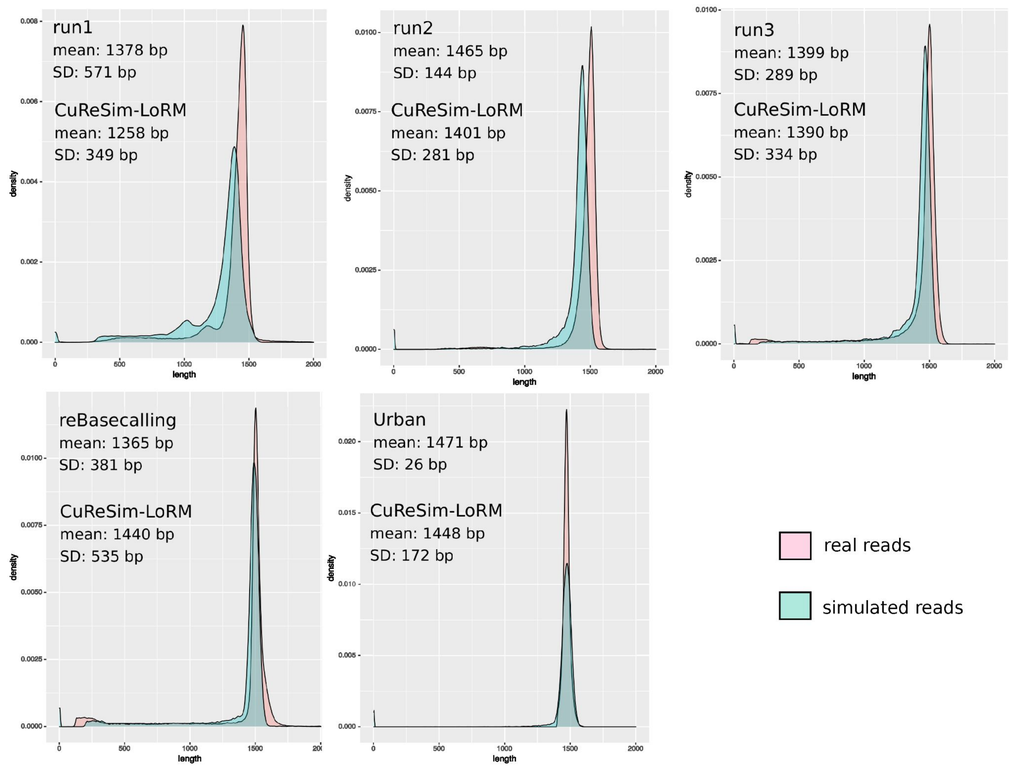
Results

Comparison of NanoSim-H, DeepSimulator, and CuReSim-LoRM simulated Data with real data



Metrics	NewLot	CuReSim-LoRM	NanoSim-H Dataset1	NanoSim-H Dataset2	NanoSim-H Dataset3	Deep-Simulator
error rate	14.45	**14.5**	12.94	13.28	18.77	10.84
%unmapped	3.4	**5.5**	6.5	6.6	9.6	0.01
%identity	86.6	**86.6**	87.8	87.2	82.7	89.4
SD	5.2	**5**	3.9	3.2	3.4	1.5
precision_Z	1	**1**	0.99	**1**	**1**	**1**
recall_Z	0.96	**0.94**	0.92	0.93	0.9	1
precision_R	0.94	**0.91**	0.87	0.87	0.86	0.84
recall_R	0.91	**0.88**	0.82	0.83	0.8	0.84
precision_S	0.74	0.76	0.67	0.64	0.62	**0.73**
recall_S	0.72	**0.72**	0.6	0.59	0.55	0.73

Evaluation of CuReSim-LoRM with challenging datasets



Metrics	run1	sim.	run2	sim.	run3	sim.	reBasecalling	sim.	Urban	sim.
Error rate	17.41	17.4	16.39	16.45	14.24	14.35	11.64	11.2	10.59	10.6
%unmapped	16.8	15.76	0.7	4.9	2.6	3.9	4.5	3.7	0	1.6
%identity	84.2	84.5	84.8	85	86.8	86.8	88.8	89.3	89.8	89.9
SD	4.7	4.6	4.3	4.6	5.2	5	4.9	4.5	3.6	3.8
precision_Z	1	1	1	1	1	1	1	1	1	1
recall_Z	0.83	0.84	0.99	0.95	0.97	0.96	0.95	0.96	1	0.98
precision_R	0.94	0.92	0.96	0.93	0.95	0.94	0.93	0.91	0.94	0.92
recall_R	0.8	0.81	0.95	0.91	0.94	0.92	0.9	0.89	0.94	0.9
precision_S	0.71	0.73	0.78	0.77	0.81	0.81	0.76	0.79	0.69	0.75
recall_S	0.6	0.61	0.77	0.73	0.78	0.78	0.72	0.76	0.69	0.74

Conclusion

CuReSim-LoRM is able (i) to produce simulated reads showing an error profile very close to the real data, (ii) to produce a read length distribution mimicking the real one, and (iii) to produce a wide range of data with varying error rates and length distributions.
CuReSim-LoRM is the first tool able to simulate ONT metabarcoding reads.

International Journal of Molecular Sciences

Mesloub Y, Beury D, Vandermeeren F, Caboche S. CuReSim-LoRM: A Tool to Simulate Metabarcoding Long Reads. Int J Mol Sci. 2023 Sep 12;24(18):14005. doi: 10.3390/ijms241814005. PMID: 37762307; PMCID: PMC10531135.