

Benchmark study of Adaptive Sampling algorithms: Simulation of various viral mixtures

Benjamin Vacus, Edith Le Floch, Zuzana Gerber, Jean-François Deleuze,

Arnaud Gloaguen

► To cite this version:

Benjamin Vacus, Edith Le Floch, Zuzana Gerber, Jean-François Deleuze, Arnaud Gloaguen. Benchmark study of Adaptive Sampling algorithms: Simulation of various viral mixtures. Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM), Jun 2024, Toulouse, France. cea-04734006

HAL Id: cea-04734006 https://cea.hal.science/cea-04734006v1

Submitted on 7 Nov 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Benchmark study of Adaptive Sampling algorithms: Simulation of various viral mixtures



Benjamin VACUS^{1,2,@}, Edith LE FLOCH², Zuzana GERBER², Jean-François DELEUZE^{2,3}, Arnaud GLOAGUEN²

¹ DGA CBRN Defense Center, Vert-le-Petit, France.

² Centre National de Recherche en Génomique Humaine (CNRGH), IBFJ, CEA, Université Paris-Saclay, Evry, France.

³ Centre de Ressources Biologiques, CEPH Biobank, Fondation Jean Dausset - CEPH, Paris, France.

Nanopore sequencing & Adaptive Sampling



By reducing the time spent sequencing off-targets molecules, it can be used to

- Longer off-targets reads lead to higher enrichment, and Adaptive Sampling becomes inefficient when the average molecule length is lower than 2 kb.
- At the start of the experiment, it seems that Readfish cannot handle the signal that flows from Icarust and performs poorly – thus, hardware configuration matters.
- The target enrichment depends on the SNR, and more simulations are needed to find the best target proportion for Adaptive Sampling – however, the more target the better!
- BOSS-RUNS shows a lower target enrichment overall but focuses on regions of interest: a **more thorough** investigation is required to verify whether it allows for better strain identification.
- **BOSS-RUNS encounters several issues** that need to be tackled: it forbids reference sequence shorter than 10^5 bp by default, and we had to use an extra live basecalling solution for it to update its target strategy.
- The link between enrichment and other parameters has yet to be explored, like for the distance between the strain in the sample and the strain used as reference.

the monitoring of the epidemic [5]. However, the low-concentration of target viruses and the complex nature of wastewater samples poses a challenge for sensitive detection and precise identification. Sequencing viruses in wastewater gives the opportunity to quantify the relative abundances of the different circulating strains [5]. We believe that long-read sequencing is an asset for precise strain deconvolution and pathogen identification. To preserve the information contained in long molecules, we are currently developing a wet lab protocol using hybridization-capture rather than amplicon-PCR:

Since this protocol aims at preserving long reads, Adaptive Sampling should perform well and help increasing the amount of target sequences. When it comes to strain deconvolution, BOSS-RUNS could actually be the boss, because it focuses on positions carrying the most uncertainty: that is the case for SNVs, because there viruses might differ from the reference, and SNVs are key to identifying the lineage. However, we will also have to verify whether Adaptive Sampling biases the different strains relative abundances.

Maybe another Adaptive Sampling paradigm would perform better, e.g. neural networks-based methods [7], and we will possibly try to design our own method, building up on the conclusions drawn from these sequencing experiments.

[1] A. Payne et al., "Readfish enables targeted nanopore sequencing of gigabase-sized genomes," Nature biotechnology, vol. 39, no. 4, pp. 442–450, 2021. [2] H. Li, "Minimap2: Pairwise alignment for nucleotide sequences," Bioinformatics, vol. 34, no. 18, pp. 3094–3100, 2018. [3] L. Weilguny et al., "Dynamic, adaptive sampling during nanopore sequencing using bayesian experimental design," Nature Biotechnology, vol. 41, no. 7, pp. 1018–1025, 2023. [4] R. Munro et al., "Icarust, a real-time simulator for oxford nanopore adaptive sampling," Bioinformatics, vol. 40, no. 4, btae141, 2024. [5] N. Cluzel et al., "A nationwide indicator to smooth and normalize heterogeneous sars-cov-2 rna data in wastewater," Environment International, vol. 158, p. 106 998, 2022. [6] G. Rios et al., "Monitoring sars-cov-2 variants alterations in nice neighborhoods by wastewater nanopore sequencing," The Lancet Regional Health–Europe, vol. 10, 2021. [7] Y. Lin et al., "Nanodeep: A deep learning framework for nanopore adaptive sampling on microbial sequencing," Briefings in Bioinformatics, vol. 25, no. 1, bbad499, 2024.

^e benjamin.vacus@cnrgh.fr • Journées Ouvertes en Biologie, Informatique, et Mathématiques • June 25 - 28 2024 • Université Paul Sabatier, Toulouse (FR) •