

Peer Review

Review of: "METAGENE-1: Metagenomic Foundation Model for Pandemic Monitoring"

Yang Yang¹

1. Shanghai Jiao Tong University, Shanghai, China

The study introduces a novel metagenomic dataset comprising 1.5 trillion base pairs derived from human wastewater samples, capturing highly diverse microbial and viral sequences. This dataset addresses gaps in existing genomic models that lack viral or unclassified sequences. The authors further assessed their models on benchmark datasets. METAGENE-1 outperforms DNABERT-2, Nucleotide Transformer (NT), and others on pathogen detection and human-virus classification tasks. Despite its focus on short sequences, the model achieves competitive results on the GUE benchmark. Through continual pretraining, it demonstrates cross-task generalization potential.

Overall, this study represents a pioneering effort in metagenomic foundation models, offering a robust framework for pandemic monitoring and biosurveillance. Its novel dataset, task-specific benchmarks, and open-source ecosystem provide valuable tools for public health research.

Major concerns :

1. METAGENE-1 relies on the Llama-2-7B architecture without domain-specific modifications for metagenomic data (e.g., handling sequence heterogeneity or short-read dependencies). Compared to recent genomic models like HyenaDNA (long-context processing) or Evo (hybrid architectures), its dependence on existing frameworks may limit breakthroughs in complex tasks.
2. Key comparisons (e.g., DNABERT-2, NT) are trained on human or curated species genomes lacking viral sequences. METAGENE-1's inherent viral data (2% of the pretraining corpus) may inflate its advantage in pathogen detection. Future work should include models trained on viral data (e.g., Evo).
3. The study focuses on task performance but lacks insights into learned biological features (e.g., viral k-mer patterns or host-pathogen interactions).

Declarations

Potential competing interests: No potential competing interests to declare.