RNA-seq Analysis as a Cloud Service: Toward AI-Driven Computational Resource Efficiency

Elisson Silva[1], Rosana Blawid[2], Stefan Blawid[1]

[1]Centro de Informática, Universidade Federal de Pernambuco, Recife, 50740-560, PE, Brazil

[2]Department of Agronomy, Universidade Federal Rural de Pernambuco, Recife, 52171-900, PE , Brazil

{elgs,sblawid}@cin.ufpe.br, rosana.blawid@ufrpe.br

Abstract. *Analysis of RNA-seq data generated by high-throughput sequencing poses substantial computational challenges due to its scale and complexity, often exceeding the resources available in typical laboratory environments. In this work, we present a cloud-centric perspective on RNA-seq analysis by deploying a phytosanitary pipeline as a cloud service based on a conventional bioinformatics workflow and analyzing its computational characteristics. Building on this baseline, we investigate the potential of artificial intelligence to improve computational resource efficiency by introducing an attention-based neural network for early-stage read classification. Our results indicate that AI-based filtering can distinguish relevant reads and reduce the volume of data processed by downstream, resource-intensive stages. This suggests a means to reduce compute time and memory usage through selective data reduction, although full integration into the pipeline is left for future work. We discuss how the combination of cloud-native execution and AI-driven preprocessing can enable more resource-efficient and accessible RNA-seq analysis services.*

1. Introduction

High-throughput sequencing (HTS) of RNA samples has become a key technology for pathogen detection [Maree et al. 2018, Villamor et al. 2019, Vazquez-Iglesias et al. 2022], but its practical adoption remains constrained by the significant computational resources required for data processing. Conventional bioinformatics pipelines for viral identification rely on multi-stage workflows, including quality control, host read removal, de novo assembly, and alignment against large and growing reference databases [Li et al. 2016, Hu et al. 2023]. These steps are computationally intensive and often require high-performance computing (HPC) infrastructure, thereby limiting accessibility for laboratories without dedicated resources [Deshpande et al. 2023]. To address these limitations, cloud computing offers a promising paradigm by enabling on-demand access to scalable infrastructure and shifting the burden of resource provisioning away from end users. However, migrating existing pipelines to the cloud does not eliminate their computational complexity and may increase operational costs if resource usage is not carefully managed [Cinaglia et al. 2023, Kica et al. 2025]. This motivates approaches that leverage cloud scalability while reducing computational demand.

Meanwhile, alignment-free methods and machine learning techniques have emerged as alternatives to traditional alignment-based approaches. In particular,

artificial neural networks (ANNs) operating on k-mer representations enable efficient classification of sequencing reads without requiring full alignment or assembly [Sukhorukov et al. 2022, Mateos et al. 2021]. Building on these developments, this work adopts a cloud-centric perspective on RNA-seq analysis by deploying a conventional pipeline as a cloud service and investigating attention-based NNs [Vaswani et al. 2017] for early-stage read classification to reduce downstream computational load. While not yet integrated into the deployed pipeline, this approach provides a proof of concept for future resource-efficient workflows.

2. Read Classification Bottlenecks in Cloud-Native RNA-seq Workflows

A central computational bottleneck in phytosanitary RNA-seq pipelines lies in the classification of HTS reads and subsequent annotation steps. These stages dominate resource consumption due to their reliance on large reference datasets and memory-intensive data structures. When deployed in cloud environments, such workloads translate directly into increased demand for high-memory instances, extended runtime, and elevated operational costs.

State-of-the-art tools such as Kraken2 and Kaiju exemplify current approaches to read classification. Kraken2 [Wood and Salzberg 2014, Wood et al. 2019] operates on nucleotide-level k-mer matching against large taxonomic databases, while Kaiju [Menzel et al. 2016] performs protein-level comparisons via translated sequence alignment. Despite their algorithmic efficiency, both tools require substantial computational resources when used with comprehensive reference collections. Memory requirements can reach hundreds of gigabytes, accompanied by significant CPU and storage demands, making them challenging to scale efficiently in cloud settings. These tools are integral components of established workflows, including pipelines such as PhytoPipe [Hu et al. 2023], where read classification plays a critical role in filtering background sequences and enabling downstream analysis. However, processing large volumes of reads against expansive reference databases creates a fundamental scalability limitation. This results in a trade-off between classification accuracy and computational cost, particularly for samples with low pathogen abundance [Lambert et al. 2018, Wright et al. 2023, Silva et al. 2025].

Recent advances in machine learning offer an alternative paradigm for read classification [Bohnsack et al. 2023]. Deep learning models operating on k-mer representations can learn discriminative sequence features directly from data, avoiding explicit dependence on large reference databases. Attention-based neural networks further enhance this approach by capturing contextual relationships within sequences [Mock et al. 2022, Wichmann et al. 2023]. Building on these developments, we focus here on early-stage read classification.

3. Attention-Based Read Classification as a Pathway to Resource Efficiency

Beyond cloud deployment, we investigate artificial intelligence to reduce the computational burden of read classification in RNA-seq pipelines. Specifically, we explore an attention-based NN operating directly on short sequencing reads (300nt/100aa) as an alternative to reference-dependent methods. The approach replaces large-scale database matching with a compact model that learns intrinsic sequence patterns, addressing limitations of convolutional NNs for short-sequence classification [Ren et al. 2020]. The proposed approach [Silva et al. 2025] represents sequencing reads in a protein-level feature

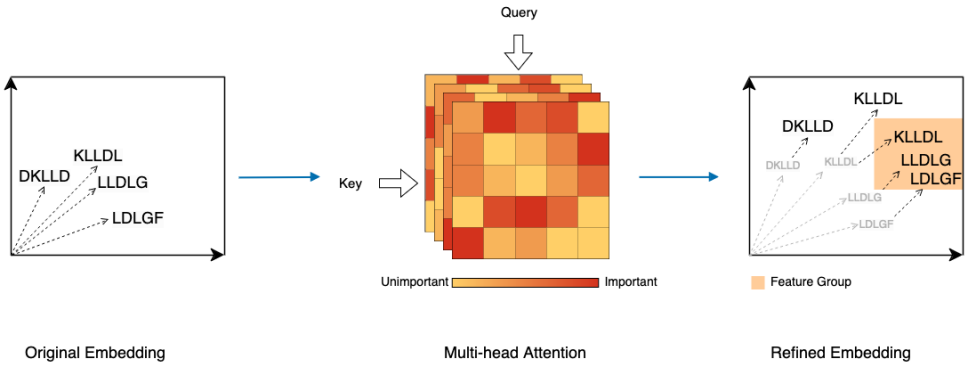

Figure 1. Schematic of the embedding refinement process, where attention scores group k-mers into feature sets that facilitate classification.

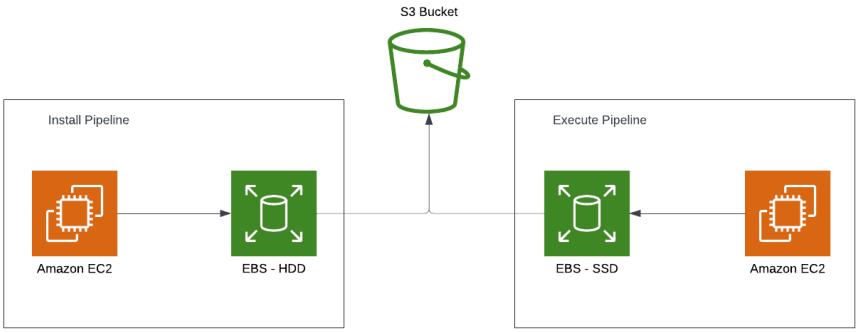

Figure 2. Resources deployed on the AWS cloud to install (left) and execute (right) PhytoPipe.

space and applies a transformer-based architecture with multi-head self-attention (MHA) to capture contextual relationships within sequence fragments (Fig. 1). This enables the model to identify discriminative patterns without explicit alignment or dependence on extensive reference libraries, allowing classification with reduced memory and storage requirements.

From a systems perspective, the key advantage lies in its role as a lightweight pre-filtering stage. By retaining only relevant reads prior to downstream processing, the model reduces the volume of data entering computationally intensive stages such as assembly and alignment. This directly translates into lower memory usage, reduced compute time, and diminished data movement. The classifier is evaluated as a proof of concept and is not yet integrated into the deployed pipeline.

4. Cloud Deployment and AI-Based Read Classification Results

The RNA-seq pipeline was successfully deployed as a cloud service on Amazon Web Services (AWS), demonstrating the feasibility of executing phytosanitary workflows without local HPC infrastructure. Elastic compute and storage enabled flexible resource allocation for database construction and sample analysis (Fig. 2). However, reference-dependent steps—particularly database generation and read classification—dominated runtime and memory consumption, requiring high-memory instances and extended execution times.

The deployed pipeline required significant resources: database construction on an r5a.12xlarge EC2 instance (48 vCPUs, 384 GB RAM) took approximately five days, even after reducing the Kraken2 database size to 256 GB and reusing pre-built Kaiju databases. Sample analysis used the same instance type with cost-optimized storage and automated resource shutdown. The average cost per sample was approximately USD 61, with peaks up to USD 100, and storage costs slightly exceeding compute due to intensive I/O. Further gains are achievable through cloud-level optimizations such as storage tiering and cost-aware instance selection.

To assess AI-based alternatives, we evaluated an attention-based read classifier against established tools using multiple RNA-seq datasets [Silva et al. 2025]. For high-confidence predictions, the model substantially enriched viral reads, with up to 70% of predicted viral reads confirmed despite low baseline abundance, while maintaining high accuracy for host reads. These results demonstrate that accurate classification of short HTS reads is feasible using attention-based models. Notably, inference can be executed on standard personal computers, in contrast to the high-memory cloud requirements of reference-based methods. As an efficient pre-filtering stage, this approach can reduce both the computational cost of read pre-selection and the data volume entering downstream, computationally intensive steps, thereby supporting more efficient cloud-based RNA-seq workflows. However, although the classifier demonstrates that relevant viral information is encoded in short reads, a non-negligible fraction of viral reads is still discarded during filtering, indicating that further improvements in sensitivity are required before productive integration into RNA-seq pipelines. The implementation of the MHA classifier is publicly available as a Python module.[1]

5. Conclusions and Outlook

This work demonstrates the feasibility of deploying an RNA-seq analysis pipeline as a cloud service. While cloud platforms provide elasticity and accessibility, migrating conventional workflows does not resolve their inherent computational inefficiencies. Reference-dependent steps such as read classification and database construction remain dominant cost and resource drivers.

We explored attention-based neural networks as a low-resource alternative for early-stage read classification. The results show that compact models can extract meaningful information from short sequencing reads and enable selective data reduction, lowering memory usage and compute time. These findings suggest a shift in pipeline design: rather than relying exclusively on large reference databases, future workflows may combine cloud scalability with AI-driven data reduction. Challenges remain, including integration into production pipelines, trade-offs between sensitivity and efficiency, and improving model generalization across diverse datasets, potentially leveraging pre-trained RNA-seq embedding models.

In summary, this work advocates a resource-aware hybrid design paradigm for RNA-seq analysis in the cloud, where scalability is complemented by intelligent reduction of computational demand.

[1] `https://github.com/elissonlima/CassBERT`

Appendix: AWS Services Used

This work relied on several Amazon Web Services (AWS) components for both pipeline execution and model development.

Pipeline Deployment (PhytoPipe). The phytosanitary pipeline was deployed using the following services:

- **Amazon EC2:** Provided compute resources for both database construction and sample analysis, using high-memory instances (e.g., r5a.12xlarge).
- **Amazon S3:** Used for persistent storage of input datasets, reference data, and output results, enabling efficient data exchange between processing stages.
- **Amazon EFS:** Provided shared file storage for intermediate data and reference databases during pipeline execution.

Model Training. The attention-based classifier was developed and trained using:

- **Amazon SageMaker:** Used to execute Python-based training workflows on GPU-enabled instances (e.g., ml.p3.2xlarge with 8 vCPUs, 61 GB RAM, and 16 GB VRAM), supporting scalable model training.
- **Amazon S3:** Used to store training and validation datasets, as well as serialized model artifacts after training.

In the model development workflow, training data were first uploaded to S3, followed by execution of training notebooks within SageMaker-managed environments. The resulting trained models were then stored back in S3 for subsequent inference.

Statement on the use of Artificial Intelligence

The authors declare that generative artificial intelligence (AI) tools were used in a limited and supportive role during the preparation of this manuscript. Specifically, such tools were employed for language refinement and English editing, including improvements in grammar, clarity, and overall readability of the text. No AI tools were used to generate original scientific content, results, figures, or data presented in this work.

All scientific ideas, methodologies, analyses, and conclusions remain the sole responsibility of the authors. The authors have carefully reviewed and validated the final manuscript to ensure its accuracy, originality, and compliance with ethical standards, including the avoidance of plagiarism. Generative AI tools are not listed as authors and did not contribute intellectually to the research.

Funding

This work was supported by AWS and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq 400049/2023-6). SB and RB received CNPq productivity grants. RB acknowledge funding from the Brazilian Federal Agency for Post-graduate Education (CAPES-PROBRAL 88881.895122/2023-01).

References

Bohnsack, K. S., Kaden, M., Abel, J., and Villmann, T. (2023). Alignment-Free Sequence Comparison: A Systematic Survey From a Machine Learning Perspective. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(1):119–135.

Cinaglia, P., Vázquez-Poletti, J. L., and Cannataro, M. (2023). Massive Parallel Alignment of RNA-seq Reads in Serverless Computing. *Big Data and Cognitive Computing*, 7(2):98.

Deshpande, D., Chhugani, K., Chang, Y., Karlsberg, A., Loeffler, C., Zhang, J., Muszyńska, A., Munteanu, V., Yang, H., Rotman, J., Tao, L., Balliu, B., Tseng, E., Eskin, E., Zhao, F., Mohammadi, P., Łabaj, P. P., and Mangul, S. (2023). RNA-seq data science: From raw data to effective interpretation. *Frontiers in Genetics*, 14:997383.

Hu, X., Hurtado-Gonzales, O. P., Adhikari, B. N., French-Monar, R. D., Malapi, M., Foster, J. A., and McFarland, C. D. (2023). PhytoPipe: a phytosanitary pipeline for plant pathogen detection and diagnosis using RNA-seq data. *BMC Bioinformatics*, 24(1):470.

Kica, P., Lichołai, S., Orzechowski, M., and Malawski, M. (2025). Accelerating Cloud-Based Transcriptomics: Performance Analysis and Optimization of the STAR Aligner Workflow. ICCS - International Conference on Computer Science, pages 257–265.

Lambert, C., Braxton, C., Charlebois, R. L., Deyati, A., Duncan, P., Neve, F. L., Malicki, H. D., Ribrioux, S., Rozelle, D. K., Michaels, B., Sun, W., Yang, Z., and Khan, A. S. (2018). Considerations for Optimization of High-Throughput Sequencing Bioinformatics Pipelines for Virus Detection. *Viruses*, 10(10):528.

Li, Y., Wang, H., Nie, K., Zhang, C., Zhang, Y., Wang, J., Niu, P., and Ma, X. (2016). VIP: an integrated pipeline for metagenomics of virus identification and discovery. *Scientific Reports*, 6(1):23774.

Maree, H. J., Fox, A., Rwahnih, M. A., Boonham, N., and Candresse, T. (2018). Application of HTS for Routine Plant Virus Diagnostics: State of the Art and Challenges. *Frontiers in Plant Science*, 9:1082.

Mateos, P. A., Balboa, R. F., Easteal, S., Eyras, E., and Patel, H. R. (2021). PACIFIC: a lightweight deep-learning classifier of SARS-CoV-2 and co-infecting RNA viruses. *Scientific Reports*, 11(1):3209.

Menzel, P., Ng, K. L., and Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, 7(1):11257.

Mock, F., Kretschmer, F., Kriese, A., Böcker, S., and Marz, M. (2022). Taxonomic classification of DNA sequences beyond sequence similarity using deep neural networks. *Proceedings of the National Academy of Sciences*, 119(35):e2122636119.

Ren, J., Song, K., Deng, C., Ahlgren, N. A., Fuhrman, J. A., Li, Y., Xie, X., Poplin, R., and Sun, F. (2020). Identifying viruses from metagenomic data using deep learning. *Quantitative Biology*, 8(1):64–77.

Silva, E., Margaria, P., Blawid, R., Oliveira, E. J., Winter, S., and Blawid, S. (2025). Hardware-Aware RNA-seq Diagnostics: Plant Virus Detection via Cloud and AI, PREPRINT (Version 1). *available at Research Square*.

Sukhorukov, G., Khalili, M., Gascuel, O., Candresse, T., Marais-Colombel, A., and Nikolski, M. (2022). VirHunter: A Deep Learning-Based Method for Detection of Novel RNA Viruses in Plant Sequencing Data. *Frontiers in Bioinformatics*, 2:867111.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. *arXiv*.

Vazquez-Iglesias, I., Santala, J., Vossenberg, B., Gaafar, Y., and Massart, S. (2022). Considerations for the use of high throughput sequencing in plant health diagnostics. *EPPO Bulletin*, 52(3):619–642.

Villamor, D. E. V., Ho, T., Rwahnih, M. A., Martin, R. R., and Tzanetakis, I. E. (2019). High Throughput Sequencing For Plant Virus Detection and Discovery. *Phytopathology*, 109(5):716–725.

Wichmann, A., Buschong, E., Müller, A., Jünger, D., Hildebrandt, A., Hankeln, T., and Schmidt, B. (2023). MetaTransformer: deep metagenomic sequencing read classification using self-attention models. *NAR Genomics and Bioinformatics*, 5(3):1–16.

Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1):257.

Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46.

Wright, R. J., Comeau, A. M., and Langille, M. G. I. (2023). From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools. *Microbial Genomics*, 9(3):000949.