
CITE V.1: Interpretable RNA-Seq Clustering with an LLM-Based Agentic Evidence-Grounded Framework

Elias Hossain*
 University of Central Florida, USA
 Mdelias.hossain@ucf.edu

Mehrdad Shoeibi
 University of Central Florida, USA
 Me604598@ucf.edu

Ivan Garibay
 University of Central Florida, USA
 Igaribay@ucf.edu

Niloofar Yousefi
 University of Central Florida, USA
 Niloofar.Yousefi@ucf.edu

Abstract

We propose *CITE V.1*, an agentic, evidence-grounded framework that leverages Large Language Models (LLMs) to provide transparent and reproducible interpretations of RNA-seq clusters. Unlike existing enrichment-based approaches that reduce results to broad statistical associations and LLM-only models that risk unsupported claims or fabricated citations, *CITE V.1* transforms cluster interpretation by producing biologically coherent explanations explicitly anchored in the biomedical literature. The framework orchestrates three specialized agents: a Retriever that gathers domain knowledge from PubMed and UniProt, an Interpreter that formulates functional hypotheses, and Critics that evaluate claims, enforce evidence grounding, and qualify uncertainty through confidence and reliability indicators. Applied to *Salmonella enterica* RNA-seq data, *CITE V.1* generated biologically meaningful insights supported by the literature, while an LLM-only Gemini baseline frequently produced speculative results with false citations. By moving RNA-seq analysis from surface-level enrichment to auditable, interpretable, and evidence-based hypothesis generation, *CITE V.1* advances the transparency and reliability of AI in biomedicine.

1 Introduction

Interpreting RNA sequencing (RNA-seq) clusters remains a central challenge in transcriptomics. While clustering methods such as spectral clustering [1] and K-means [2] effectively group genes by expression, downstream analyses often rely on enrichment-based statistics [3] that reduce results to broad associations without offering cluster-specific explanations. This limitation restricts our ability to connect expression patterns to concrete biological functions and mechanisms. Without such links, a cluster may be labeled broadly as “metabolism-related”, but lack clarity on which pathways, regulators, or virulence factors are involved, details that are essential for designing downstream experiments such as drug target validation or dissecting host–pathogen interactions, both of which are high-impact yet experimentally challenging. Recent advances in Large Language Models (LLMs) have opened opportunities in biomedical text mining [4, 5], but without explicit domain grounding they risk generating inconsistent interpretations, unsupported claims, and fabricated citations.

This challenge is particularly critical in the study of *Salmonella enterica*, the causative agent of salmonellosis, the second most prevalent bacterial foodborne infection in the United States. The U.S. Centers for Disease Control and Prevention (CDC) estimates that nontyphoidal *Salmonella* causes about 1.35 million infections and 420 deaths annually in the United States [6], with roughly 99

Understanding the molecular mechanisms of *S. enterica* is therefore essential for unraveling host–pathogen interactions, virulence strategies, and environmental adaptation [7]. RNA-seq [8] provides transcriptome-wide views of these processes and has advanced with bulk and single-cell methods [9]. Yet, the interpretive gap persists: heterogeneous gene clusters are identified, but their functional roles remain only superficially explained.

To address this gap, we introduce *CITE V.1*, an agentic, evidence-grounded framework that leverages LLMs for interpretable RNA-seq clustering in *S. enterica*. Unlike single-model pipelines, *CITE V.1* orchestrates three specialized agents, a **Retriever** that gathers references from PubMed and UniProt, an **Interpreter** that formulates cluster-level hypotheses, and **Critics** that evaluate claims and qualify uncertainty through confidence and reliability indicators. By explicitly grounding outputs in biomedical literature, *CITE V.1* transforms cluster interpretation into an auditable, transparent, and reproducible process.

Our contributions are threefold:

1. We introduce *CITE V.1*, the first agentic LLM-based framework for interpretable RNA-seq clustering in *S. enterica*;
2. We design an orchestrated pipeline where agents integrate evidence retrieval, hypothesis generation, and reliability assessment; and
3. We demonstrate, through comparative evaluation, that *CITE V.1* produces biologically coherent, literature-supported insights while avoiding the speculative errors of LLM-only baselines.

2 CITE V.1 Framework

CITE V.1 is an LLM-based Agentic framework for interpretable RNA-seq clustering, as illustrated in Figure 1. Starting from an expression matrix $M \in \mathbb{R}^{s \times g}$, clustering partitions genes into modules $\{G_1, \dots, G_m\}$. Each module is analyzed through three coordinated agent roles: (i) a *Retriever*, which queries PubMed and UniProt (with supplementary search when necessary) to collect both specific references at the gene or protein level and broader context references; (ii) an *Interpreter*, which synthesizes the retrieved evidence into cluster-level hypotheses covering functional themes, pathways, transcriptional links, distinctiveness, and citations; and (iii) a panel of *Critics*, including Evidence-Strict, Semantic, and Adversarial evaluators, whose assessments are integrated by a Consensus Critic to yield a reliability flag and confidence score. By combining retrieval, interpretation, and multi-perspective evaluation, *CITE V.1* moves beyond enrichment-based pipelines, producing structured, auditable, and evidence-grounded interpretations.

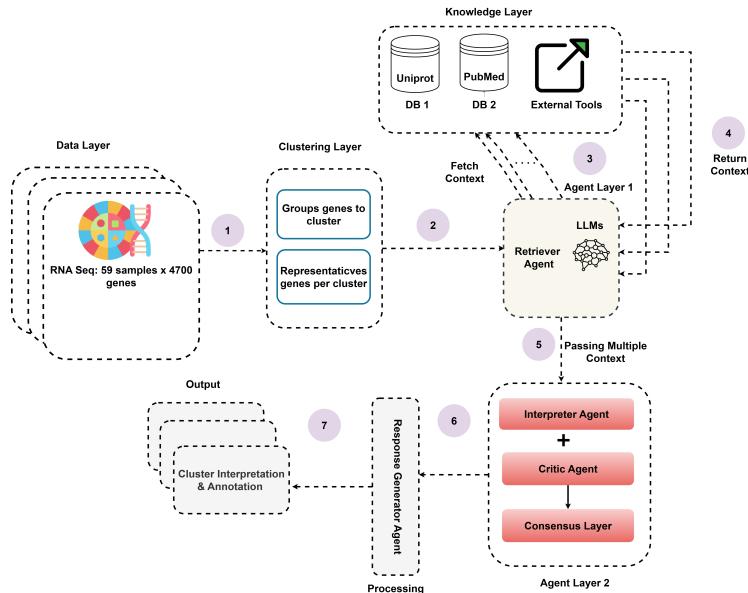
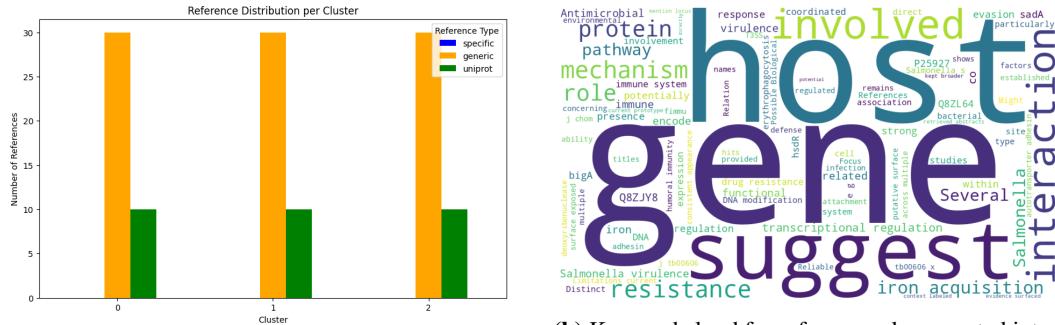


Figure 1: Overview of the *CITE V.1* framework for interpretable RNA-seq analysis. Clusters are enriched with PubMed and UniProt evidence, interpreted by the Interpreter Agent, and validated through a panel of Critic Agents, producing transparent literature-grounded outputs.



(a) Reference distribution per cluster. The framework retrieved UniProt and PubMed-generic references in this dataset, while the PubMed-specific category is retained to reflect the full layered evidence design.

Figure 2: Supporting analyses from CITE V.1. (a) Distribution of references (specific, generic, and UniProt) demonstrates the framework’s layered evidence grounding, with the absence of PubMed-specific hits in this dataset reflecting dataset-specific coverage. (b) Keyword extraction reveals recurrent biological themes and functional linkages across clusters.

(b) Keyword cloud from framework-generated interpretations. Prominent terms such as *gene*, *host*, and *interaction* highlight recurrent biological themes including resistance, iron acquisition, and transcriptional regulation.

3 Results and Comparative Evaluation

CITE V.1 produced interpretable outputs by integrating three layers of evidence: (i) *PubMed-specific references*, which are articles explicitly mentioning the target gene, protein, or locus in their titles or abstracts; (ii) *PubMed-generic references*, which discuss broader *Salmonella* biology (e.g., host-pathogen interactions, virulence, or antimicrobial response) without directly focusing on the studied gene; and (iii) *UniProt entries*, which contribute curated protein-level annotations. While the framework supports all three evidence layers, for the *Salmonella* dataset analyzed here the PubMed-specific layer did not yield results for the top clustered genes, leading to its absence in Figure 2a. This reflects dataset-specific coverage rather than a limitation of the framework. Supporting analyses (Figure 2a, Figure 2b) nevertheless illustrate balanced grounding in generic and UniProt evidence, with the keyword cloud in Figure 2b highlighting recurrent biological themes such as host interaction, iron acquisition, and antimicrobial resistance.

A detailed structured interpretation for Cluster 1 is shown in Figure 3. The output brings together three evidence layers: PubMed-specific articles, broader PubMed-generic context, and UniProt annotations, to connect virulence-associated genes (*sadA*, *bigA*), iron uptake mechanisms, and resistance factors such as *hsdR*. At the same time, the framework reports limitations: because no transcriptional regulation evidence was retrieved, the interpretation was flagged as *unreliable* with a confidence score of 0.45. By making both strengths and gaps explicit, the figure shows how CITE V.1 avoids overstated certainty and instead provides transparent, critic-qualified interpretations grounded in biomedical references. For comparison, Figure 3(b) illustrates the Gemini baseline, which produced speculative outputs, misclassified the organism as *Streptomyces*, and listed only hypothetical references marked with “[Citation Needed],” underscoring the absence of verifiable evidence.

To contextualize these outputs, Figure 4 compares *CITE V.1*'s Cluster 1 interpretation with Gemini-only baselines, with both panels corresponding to Cluster 1 to enable a direct comparison. While *CITE V.1* produced structured and auditable results grounded in references, Gemini frequently misclassified organisms, relied on vague functional terms, and failed to provide verifiable citations. Its confidence scores (0.3–0.4) were self-reported estimates without critic validation. In contrast, the *CITE V.1* framework combined reference diversity with critic-based evaluation, demonstrating a clear advantage over single-model pipelines in producing trustworthy and interpretable biological insights.

However, details of how reliability and confidence were assessed for both *CITE V.1* and the Gemini baseline are described in Appendix C.

Response: Cluster 1 Example Interpretation (Generated by CITE V.1)	Response: Cluster 1 Example Interpretation (Generated by Gemini)
<p>1. Functional Themes: Virulence genes (<i>sadA</i>, <i>bigA</i>), iron uptake (<i>hsdR</i>), surface proteins.</p> <p>2. Pathways: Host-pathogen interaction (SadA, BigA); iron acquisition (erythrophagocytosis); antimicrobial resistance (<i>hsdR</i>).</p> <p>3. Regulation: No direct transcriptional regulation found.</p> <p>4. Distinctiveness: Genes co-occur around virulence, iron, resistance.</p> <p>5. References: PubMed IDs 40373749, 32038650; UniProt Q8ZL64, P25927, Q8ZY8.</p> <p>Confidence: 0.45 (unreliable).</p> <p>Limitations: No locus names in titles; regulation missing.</p>	<p>1. Functional Themes: Speculative roles in secondary metabolism, stress response, regulation (organism misclassified as <i>Streptomyces</i>).</p> <p>2. Pathways: Hypothetical secondary metabolite biosynthesis; signal transduction; stress response.</p> <p>3. Regulation: Suggests a regulator may control expression, but no evidence provided.</p> <p>4. Distinctiveness: Interpreted as a “functional module,” largely hypothetical.</p> <p>5. References: Hypothetical Ref 1: [Citation Needed] (from Gemini output) <i>Regulation of secondary metabolite biosynthesis in Streptomyces</i>. Hypothetical Ref 2: [Citation Needed] (from Gemini output) <i>Stress response mechanisms in Streptomyces</i>.</p> <p>Confidence: 0.3 (low).</p>

(a) Interpretation generated by CITE V.1.

(b) Interpretation generated by Gemini baseline.

Figure 3: Comparison of interpretations for Cluster 1. (a) CITE V.1 provides structured, reference-grounded outputs with explicit limitations. (b) Gemini baseline yields speculative, less verifiable interpretations with placeholder references explicitly returned as “[Citation Needed]” in the original Gemini output.

CITE V.1: Cluster 1 Example	Gemini Baseline: Cluster1 1 Example
<p>Strengths: Clear, evidence-grounded themes. – Virulence genes (<i>sadA</i>, <i>bigA</i>). – Iron acquisition and resistance (<i>hsdR</i>). – Specific PubMed/UniProt references. – Reliability flag = Unreliable, confidence = 0.45 (transparent).</p> <p>Limitations: No transcriptional regulation detected.</p>	<p>Weaknesses: Speculative and misclassified. – Incorrect organism (<i>Streptomyces</i> vs. <i>Salmonella</i>). – Vague terms (“signal transduction,” “stress response”). – No references (placeholders only). – Arbitrary confidence scores (0.3–0.4), no critic validation.</p>

(a) CITE V.1 output grounded in references with transparent reliability.

(b) Gemini outputs were speculative, unreferenced, and unreliable.

Figure 4: Comparative visualization of cluster interpretations: (a) CITE V.1 produced structured, reference-grounded interpretations with critic-based reliability for Cluster 1, while (b) Gemini baseline on Cluster 1 was speculative, misclassified, and lacked evidence.

4 Limitations and Future Work

While CITE V.1 highlights the advantages of LLM-based agentic orchestration in RNA-seq cluster interpretation, this study was evaluated on a relatively small dataset and requires expert validation to confirm robustness. The framework also has room for refinement, particularly in retrieval coverage and critic evaluation. Limitations include reliance on biomedical databases, which may introduce bias, and the computational cost of multi-agent orchestration. Future work will scale evaluation to larger datasets, integrate systematic expert validation, and extend the framework to broader bacterial genomics applications.

5 Conclusion

This research introduced CITE V.1, an LLM-based agentic framework that integrates retrieval, interpretation, and critique for interpretable RNA-seq cluster analysis in *Salmonella enterica*. By grounding outputs in PubMed and UniProt, the framework advances interpretation beyond statistical associations toward literature-backed hypotheses. Our evaluation showed that CITE V.1 consistently linked genes to functional themes and pathways while qualifying uncertainty with confidence scores and reliability flags. In contrast, the Gemini baseline often produced speculative, misclassified interpretations with unverifiable references, underscoring the risks of relying solely on general-purpose LLMs. These findings demonstrate the value of agentic orchestration for improving accuracy and transparency in biomedical interpretation. Beyond addressing a key transcriptomics bottleneck, the system provides a scalable foundation for evidence-grounded analyses. Future work will extend the framework to additional pathogens, larger datasets, and expert validation.

Acknowledgment

The authors would like to express their sincere gratitude to the Complex Adaptive Systems Laboratory at the University of Central Florida for providing support and resources that made this research possible.

Declaration Of Competing Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- [1] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007. [1](#)
- [2] Kristina P Sinaga and Miin-Shen Yang. Unsupervised k-means clustering algorithm. *IEEE access*, 8:80716–80727, 2020. [1](#)
- [3] Dongmei Li. Statistical methods for rna sequencing data analysis. *Exon Publications*, pages 85–99, 2019. [1](#)
- [4] Simona Emilova Doneva, Sijing Qin, Beate Sick, Tilia Ellendorff, Jean-Philippe Goldman, Gerold Schneider, and Benjamin Victor Ineichen. Large language models to process, analyze, and synthesize biomedical texts: a scoping review. *Discover Artificial Intelligence*, 4(1):107, 2024. [1](#)
- [5] Baqer M Merzah, Tania Taami, Salman Asoudeh, Saeed Mirzaee, Amir Ali Bengari, et al. Biopars: A pretrained biomedical large language model for persian biomedical text mining. *arXiv preprint arXiv:2506.21567*, 2025. [1](#)
- [6] Bibek Lamichhane, Asmaa MM Mawad, Mohamed Saleh, William G Kelley, Patrick J Harrington, Cayenne W Lovestad, Jessica Amezcuia, Mohamed M Sarhan, Mohamed E El Zowalaty, Hazem Ramadan, et al. Salmonellosis: an overview of epidemiology, pathogenesis, and innovative approaches to mitigate the antimicrobial resistant infections. *Antibiotics*, 13(1):76, 2024. [1](#)
- [7] Lu Tan, Zhihao Guo, Yanwen Shao, Lianwei Ye, Miaomiao Wang, Xin Deng, Sheng Chen, and Runsheng Li. Analysis of bacterial transcriptome and epitranscriptome using nanopore direct rna sequencing. *Nucleic acids research*, 52(15):8746–8762, 2024. [1](#)
- [8] Francisco García-del Portillo and M Graciela Pucciarelli. Rna-seq unveils new attributes of the heterogeneous salmonella-host cell communication. *RNA biology*, 14(4):429–435, 2017. [1](#)
- [9] Xinmin Li and Cun-Yu Wang. From bulk, single-cell to spatial rna sequencing. *International journal of oral science*, 13(1):36, 2021. [1](#)
- [10] Jiabin Zhang, Lei Zhao, Wei Wang, Quan Zhang, Xue-Ting Wang, De-Feng Xing, Nan-Qi Ren, Duu-Jong Lee, and Chuan Chen. Large language model for horizontal transfer of resistance gene: From resistance gene prevalence detection to plasmid conjugation rate evaluation. *Science of The Total Environment*, 931:172466, 2024. [7](#)
- [11] Yujie You, Kan Tan, Zekun Jiang, and Le Zhang. Developing a predictive platform for salmonella antimicrobial resistance based on a large language model and quantum computing. *Engineering*, 2025. [7](#)
- [12] Moses B Ayoola, Athish Ram Das, B Santhana Krishnan, David R Smith, Bindu Nanduri, and Mahalingam Ramkumar. Predicting salmonella mic and deciphering genomic determinants of antibiotic resistance and susceptibility. *Microorganisms*, 12(1):134, 2024. [7](#)
- [13] Nicole E Wheeler, Paul P Gardner, and Lars Barquist. Machine learning identifies signatures of host adaptation in the bacterial pathogen salmonella enterica. *PLoS genetics*, 14(5):e1007333, 2018. [8](#)
- [14] European Nucleotide Archive. European nucleotide archive (ena), 2025. URL <https://www.ebi.ac.uk/ena>. Accessed: September 21, 2025. [8](#)

- [15] Virginie Stévenin, Claudia E Coipan, Janneke W Duijster, Daphne M van Elsland, Linda Voogd, Lise Bigey, Angela HAM van Hoek, Lucas M Wijnands, Lennert Janssen, Jimmy JLL Akkermans, et al. Multi-omics analyses of cancer-linked clinical salmonellae reveal bacterial-induced host metabolic shift and mtor-dependent cell transformation. *Cell Reports*, 43(11), 2024. 8

Appendix

Contents

1	Introduction	1
2	<i>CITE V.1</i> Framework	2
3	Results and Comparative Evaluation	3
4	Limitations and Future Work	4
5	Conclusion	4
	Appendix	7
A	Related Work	7
B	Experimental Setup	8
B.1	Data Collection and Cleaning	8
B.2	Computational Resources and Frameworks	9
B.3	Training Configuration	10
C	Evaluation Metrics	10

A Related Work

This section provides an overview of the representative works that emphasize the combination of LLMs, machine learning methods, and genomics for both host adaptation analysis and resistance prediction.

Zhang et al. [10] utilized plasmid and chromosomal sequence datasets with annotated ARGs, together with experimental data on plasmid conjugation rates. Their method applied DNABERT for plasmid versus chromosome classification and ARG detection, combined in an ensemble with k-mer features and random forest for conjugation rate prediction under different conditions. The results demonstrated improved ARG detection and identified key factors such as repression status, cell density, and temperature that influence plasmid conjugation. The strength of this study lies in the integration of genomic LLMs with biological experiments, effectively linking gene prevalence to plasmid transfer and expanding AMR surveillance.

You et al. [11] used *Salmonella* genomic data focused on resistance gene sequences and related features. Their method introduced the SARPLLM platform, which combined Qwen2 LLM (with LoRA fine-tuning) and the quantum algorithm QSMOTEN, employing a two-step feature selection process for key resistance genes. The results showed higher accuracy and robustness compared to classical machine learning models, with strong performance in predicting antimicrobial resistance and addressing class imbalance. The strength of this work lies in integrating LLMs with quantum computing, ensuring both interpretability and scalability for AMR prediction.

Ayoola et al. [12] analyzed *Salmonella* whole-genome sequencing data paired with minimum inhibitory concentration (MIC) values for multiple antibiotics. Their method employed a machine learning framework based on gradient boosting models to predict MIC values from genomic features, with SHAP analysis applied to interpret the contribution of specific genes and mutations to resistance or susceptibility. The results demonstrated high accuracy in predicting resistance profiles and identified key genomic determinants associated with antibiotic response, providing insights into mechanisms of antimicrobial resistance in *Salmonella*. The strength of this work lies in combining

predictive modeling with interpretable feature attribution, enabling both accurate resistance prediction and biological insight into genomic drivers of AMR.

Wheeler et al. [13] analyzed a large dataset of *Salmonella enterica* genomes from human, animal, and environmental sources. Their method applied machine learning to detect genomic variations and classify strains by host specificity, combined with comparative genomics. The results identified genetic signatures separating host-adapted strains and candidate genes linked to adaptation, thereby improving knowledge of evolution and host-pathogen interactions. The strength of this study lies in the integration of machine learning and comparative genomics, which uncovered host-adaptive features not easily detected by traditional methods.

Even though these research show how effective machine learning and LLMs are at predicting AMR and analyzing host-pathogen interactions, the majority of them concentrate on feature attribution or prediction accuracy rather than interpretability, trustworthiness, or integration with well chosen biological evidence. This gaps highlights the need for frameworks such as *CITE V.1* that combine multi-agent reasoning, evidence retrieval, and critic-driven validation to deliver transparent and trustworthy biological interpretations.

B Experimental Setup

B.1 Data Collection and Cleaning

Raw RNA-seq data were obtained from the European Nucleotide Archive (ENA) [14] under accession PRJEB67574, originally reported in [15]. The dataset included 59 *Salmonella enterica* isolates from colon cancer patients and control samples, offering a useful resource for studying pathogen-driven host transformation. Paired-end FASTQ files were available for each isolate. Since all samples came from a single study, inter-study batch effects were minimized, while residual technical variation was later addressed through normalization.

Several preprocessing and quality control steps were applied to maintain data integrity. FASTQ files were downloaded using `wget`, and any incomplete files were re-downloaded. Quality control was carried out with *FastQC* to evaluate per-base quality, GC content, adapter contamination, and sequence duplication levels. Results were then summarized with *MultiQC*. Figure 5 presents sequencing depth and duplication rates across all samples, confirming that sufficient unique reads were available for downstream analysis. Likewise, the GC content distribution (Figure 6) showed a unimodal peak near 50%, consistent with the expected genomic profile of *Salmonella enterica*.

At the read level, *fastp* trimming was used to remove adapters, low-quality bases, and very short sequences, ensuring that only reliable reads were retained for quantification. The cleaned reads were aligned to the *Salmonella enterica* Typhimurium LT2 reference genome (RefSeq: GCF_000006945.2) using *BWA-MEM*. Reference FASTA and GFF annotation files were obtained from NCBI RefSeq to support genome alignment and feature quantification. After alignment, *samtools* was employed for sorting, indexing, and BAM file management. Finally, *featureCounts* generated a complete sample-by-gene expression matrix with non-zero counts for most annotated genes.

Following preprocessing, alignment, and gene-level quantification, a clean dataset was obtained for downstream unsupervised learning. The final expression matrix included 58 samples and 4,679 genes, representing normalized counts across isolates. The main properties of this dataset are summarized in Table 1.

Table 1: Overview of the final processed dataset used for clustering and interpretation.

Property	Value
Study accession	PRJEB67574
Organism	<i>Salmonella enterica</i>
Number of samples	58
Number of genes	4,679
Matrix dimensions	58 × 4679
Data type	Gene-level normalized expression counts
Processing pipeline	<i>FastQC</i> , <i>fastp</i> , <i>BWA-MEM</i> , <i>samtools</i> , <i>featureCounts</i>

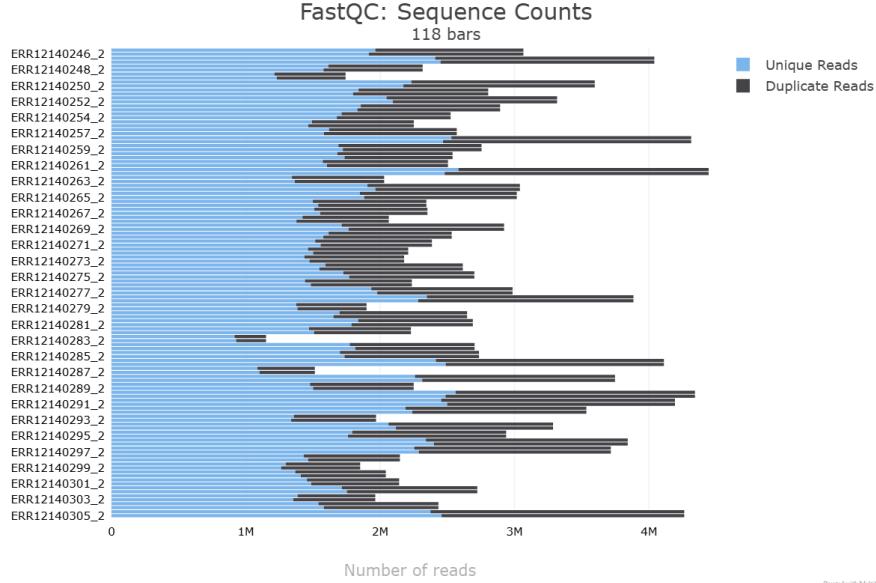


Figure 5: MultiQC summary of sequence counts across samples. Blue bars indicate unique reads, while black bars denote duplicate reads.

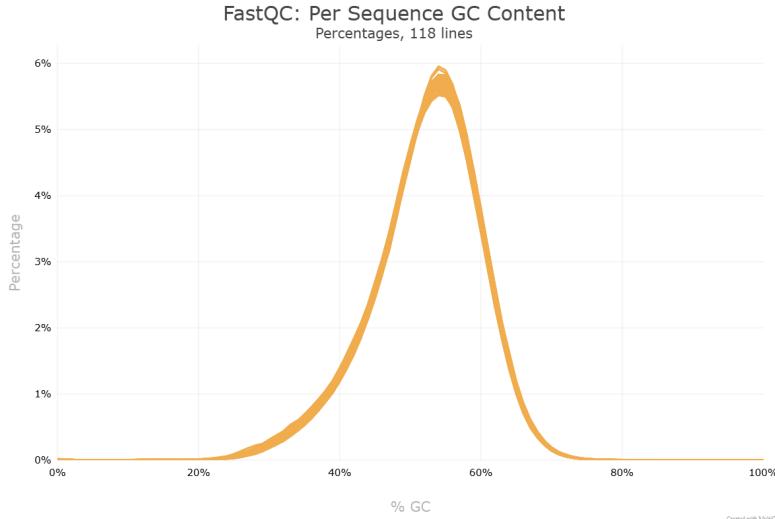


Figure 6: MultiQC summary of per-sequence GC content distribution across samples. The centered peak around 50% GC indicates expected bacterial genomic composition.

This matrix served as the input to the unsupervised clustering step and subsequent LLM-driven interpretation framework.

B.2 Computational Resources and Frameworks

All experiments were performed on Google Colab Pro, which provided stable access to GPUs (NVIDIA Tesla T4 and V100) and high-memory CPUs. This environment ensured consistent performance across runs and supported the training and evaluation of both classical and transformer-based models.

The framework relied on different components, each serving a distinct purpose. Data processing libraries such as *pandas* and *numpy* were used to clean, organize, and prepare RNA-seq inputs, with *tqdm* enabling efficient progress monitoring. Classical machine learning methods, including KMeans clustering, were implemented using *scikit-learn*. For biological knowledge retrieval, *BioPython*

accessed PubMed and UniProt via Entrez APIs, supplemented by *requests* for REST queries and *googlesearch* for additional literature evidence. The core reasoning engine was Google Gemini (1.5 Flash), accessed through the *google-generativeai* SDK, which powered transcriptional module interpretation. On top of this, a custom agentic pipeline was developed, consisting of a *ClusterAgent* for unsupervised module discovery, a *RetrieverAgent* for evidence collection, an *InterpreterAgent* for LLM-based biological interpretation, and a *CriticAgent* for evaluating reliability. All configuration parameters, including dataset paths, number of clusters, top gene selection, and API keys, were organized through a modular *config/settings* system.

B.3 Training Configuration

Clustering and downstream LLM-based interpretation were carried out in a fully unsupervised manner, without any supervised training or fine-tuning. In particular, *scikit-learn*'s *KMeans* was used with `random_state=42` and the number of clusters defined in `settings.N_CLUSTERS`. We adopted *KMeans* because it provides a simple and widely used baseline for partitioning high-dimensional gene-expression data, making it easier to evaluate how much added value the multi-agent interpretation layer contributes beyond standard clustering. In this study, we set the number of clusters to $k = 3$ and selected the top 10 variable genes for interpretation to focus on the most informative expression signals. The algorithm was applied to the gene-expression matrix containing all columns except `Sample`, and cluster labels were generated using `fit_predict` and added as a new `Cluster` column in the *pandas* DataFrame. No scaling, normalization, or imputation was applied—the matrix was used as originally loaded.

For interpretation, the backbone LLM was Google Gemini 1.5 Flash (via *google-generativeai*). The *RetrieverAgent* first queried PubMed through *BioPython* Entrez (with `retmax=3`), then fell back to the UniProt REST API (with `size=3`), and finally *googlesearch*. The retrieved evidence, combined with the gene list, was embedded into a structured prompt. The *InterpreterAgent* then produced structured biological interpretations, while the *CriticAgent* evaluated their validity. Reliability was not assigned as a fixed value but was dynamically assessed through the CriticAgent and consolidated in a consensus layer, ensuring that interpretations reflected agreement across multiple evaluation criteria. All run-time parameters, including dataset paths, API keys, number of clusters, and top gene settings, were handled through a modular *config/settings* system.

This workflow followed the architecture in Figure 1, where the RetrieverAgent supplies evidence, the InterpreterAgent generates hypotheses, and the CriticAgent moderates outputs before aggregation in the consensus layer.

C Evaluation Metrics

of critic-based

The evaluation strategy differed between *CITE V.1* and the Gemini baseline, reflecting their design differences.

Evaluation of *CITE V.1*:

The trustworthiness of *CITE V.1* outputs was assessed through a consensus-driven critic method that combines rule-based and LLM-based evaluations. First, the *Consensus Critic* aggregated the outputs of specialized critics who reviewed each cluster interpretation produced by the *Interpreter Agent*. To detect inconsistencies, unsupported claims, or weak reasoning, the *Adversarial LLM Critic* employed a large language model and returned a reliability score (0.0–1.0) together with a binary reliability flag.

The *Evidence-Strict Critic* quantified evidence support, assigning weights of 0.6 to PubMed-specific references, 0.2 to PubMed-generic, 0.15 to UniProt, and 0.05 to supplementary web sources. Interpretations with scores above 0.7 were labeled as reliable. The *Semantic Critic* verified alignment between retrieved evidence and interpretation. All outputs were then aggregated by the *Consensus Critic*, which applied majority voting for reliability and averaged scores across critics.

Evaluation of the Gemini Baseline:

Gemini was evaluated as a raw general-purpose LLM without the agentic pipeline. The dataset was clustered using *KMeans* ($k = 3$), and the top 10 high-variance genes from each cluster were provided

as input. Gemini was prompted with a structured template requesting: (i) functional themes, (ii) possible pathways, (iii) transcriptional regulation, (iv) reasons for distinctiveness, (v) references, and (vi) a final confidence score (0.0–1.0). Outputs were then assessed for:

- Correctness of organism and biological context,
- Plausibility of functional assignments,
- Presence or absence of references (noting fabricated or placeholder citations),
- Consistency of self-reported confidence scores.

Unlike *CITE V.1*, Gemini did not include critic-based validation or weighted evidence checks, making evaluation reliant on post-hoc inspection of outputs.

Overall, these measures ensured a fair comparison: *CITE V.1* was evaluated through LLM-based Agentic orchestration with critic-driven validation, while Gemini was benchmarked as a raw LLM prompted for cluster interpretation. This distinction highlights the benefits of agentic critic orchestration in generating transparent, auditable interpretations.