Title: GenAI-Assisted Interpretation of Metagenomic Sequencing Data

- Srirama Murthy Chellu

Abstract

Metagenomic next-generation sequencing (mNGS) provides unbiased detection of microbial pathogens but produces complex, high-noise outputs that require expert interpretation. This project develops an AI-assisted interpretation pipeline that integrates classical bioinformatics, statistical pathogen scoring, host-response modeling, and large language model (LLM)–based reasoning to generate clinician-ready infectious disease reports. Using the sample **NGSRL2033_1**, we performed quality control, RPM normalization, contaminant removal, genus-level collapsing, and a weighted pathogen ranking to identify high-confidence organisms. Semantic embeddings (Sentence-Transformers), UMAP visualization, and density-based clustering (DBSCAN) further separated true microbial signals from environmental and reagent contaminants. A biologically realistic host-response panel demonstrated a strong bacterial signature (bacterial score 0.78, viral score 0.12) with low sepsis likelihood (0.05). Integrating these features, an LLM generated a coherent clinical interpretation identifying **Escherichia coli** as the most likely pathogen, with Klebsiella as a secondary or background organism and all viral reads as noise. The system produced a structured summary, differential diagnosis, severity assessment, and treatment guidance like expert manual review. This work demonstrates a practical framework for combining mNGS analytics with GenAI to deliver accurate, explainable, and clinically meaningful infectious disease diagnostics.

Project Overview

Metagenomic Next-Generation Sequencing (mNGS) has transformed infectious disease diagnostics by enabling unbiased detection of bacteria, viruses, fungi, and parasites directly from patient samples. However, the output from mNGS is extremely complex, containing hundreds of microbial signals that may represent true pathogens, environmental background, reagent contaminants, commensals, or false positives.

Clinicians often struggle to interpret raw mNGS data due to:

- High noise from non-pathogenic organisms
- Viral reads without supportive host response
- Cases where multiple bacteria appear simultaneously
- Low-abundance organisms with high identity percentages
- Difficulty integrating microbial and host transcriptomic signals
- Lack of explanatory clinical reasoning

This project builds a next-generation clinical interpretation engine that combines:

1. Classical bioinformatics filtering

2. Statistical pathogen ranking
3. GenAI clinical reasoning
4. Host-response integration
5. Embedding-based biological clustering

The result is a clinician-ready diagnostic assessment, like what leading laboratories (UCSF, Stanford, Broad Institute, Mayo Clinic) generate manually, but now automated.

Objective

The primary objective of this project is to design and implement a GenAI-augmented metagenomic infectious disease interpretation pipeline capable of translating raw mNGS outputs into clinically meaningful, evidence-based reports. While mNGS has transformed pathogen detection by enabling unbiased sequencing of all nucleic acids in a sample, its outputs are often noisy, difficult to interpret, and require expert clinical microbiology oversight.

This project aims to bridge that gap by:

1. Building a reproducible data engineering workflow
   - Importing, cleaning, and quality-controlling the CZID mNGS report.
   - Normalizing microbial reads, computing composite pathogen scores, and removing contaminants.
2. Integrating host-gene expression signals
   - Simulating realistic host-response metrics (bacterial, viral, inflammation, sepsis-likelihood).
   - Combining microbial and host features into a unified diagnostic dataset.
3. Applying modern GenAI and embeddings
   - Using sentence-transformer embeddings + UMAP + clustering to visualize microbe relationships.
   - Allowing AI models to reason for pathogen relevance, contamination, and host-microbe conflicts.
4. Generating clinician-ready interpretations using LLMs
   - Producing structured assessments (pathogen ranking, host-response summary, differential diagnosis).
   - Creating a final clinical impression
5. Demonstrating the feasibility of AI-assisted infectious disease diagnostics
   - Showing how GenAI can enhance decision-making, reduce manual review effort, and automate reporting.

Usage of GenAI in Bioinformatics

Generative AI (GenAI) is rapidly transforming modern bioinformatics by enabling deeper interpretation, automation, and reasoning on top of high-dimensional biological datasets. Traditional bioinformatics pipelinesmwhile powerful often struggle with noisy data, ambiguous microbial signals, and the need for domain-expert interpretation. GenAI bridges this gap by providing contextual understanding, pattern detection, and clinical reasoning that were previously manual and time consuming.

Below is a detailed overview of how GenAI is used in this project and how it is reshaping the field.

1. Automated Interpretation of mNGS Results

Metagenomic Next-Generation Sequencing (mNGS) produces large, complex datasets containing:

- microbial taxa
- read counts
- RPM (reads per million)
- identity scores
- contig assembly support
- background noise and contaminants

GenAI models (LLMs) can:

- summarize microbial patterns
- identify true pathogens vs contaminants
- explain significance of low-abundance organisms
- infer clinical impact

In this project, a GenAI model was used to evaluate pathogen relevance, compare microbial evidence, and generate a clinician-style report.

2. Host–Pathogen Interaction Reasoning

Host gene expression is a powerful signal for understanding infection states (bacterial/viral inflammation, sepsis risk).
GenAI can:

- interpret host immune activation profiles
- reconcile contradictions between microbial reads and host response
- explain whether detected viruses/bacteria match the host immune signature

This project integrates a host-response JSON layer and uses LLM reasoning to determine infection likelihood.

3. Knowledge-Driven Clinical Reasoning

GenAI models are trained on:

- medical literature
- infectious disease guidelines
- genomic interpretation reports
- pathogen characteristics

This allows GenAI to:

- simulate expert clinical reasoning
- classify microbes based on pathogenicity
- generate differential diagnoses
- propose diagnostic next steps

4. Embedding-Based Microbial Clustering

Using transformer-based embedding models, GenAI can encode biological names and metadata into numerical vectors.
These are used for:

- microbial similarity analysis
- contaminant cluster identification
- visualization with UMAP
- DBSCAN clustering of noise vs pathogens

In this project:

- all-MPNet-base-v2 embeddings
- UMAP dimensionality reduction
- GPU acceleration

were used to compute organism similarity maps.

This enabled identification of clusters of environmental bacteria vs real pathogens.

5. Pathogen Scoring with AI-Assisted Ranking

GenAI helps refine pathogen ranking by:

- analyzing RPM, identity, and contigs
- understanding clinical context
- combining multiple scoring signals
- generating explanations for rankings

6. Fully Automated Clinical Report Generation

GenAI can convert raw microbial data + host response into a structured diagnostic report, including:

- pathogen assessment
- host-response assessment
- integrated interpretation

- sepsis risk estimation
- Differential Diagnosis
- clinician recommendations

The final report in this project mirrors output seen in clinical bioinformatics labs.

7. Future Applications of GenAI in Bioinformatics

GenAI will continue to impact multiple domains:

- personalized medicine
- drug discovery
- genome annotation
- protein structure prediction (AlphaFold-3)
- lab automation
- real-time pathogen surveillance

This project demonstrates how GenAI enhances mNGS diagnostics and sets the foundation for next-generation clinical bioinformatics.

Dataset

The project uses a single high-quality IDseq/CZ ID style mNGS dataset.

Sample ID: NGSRL2033_1

Rows: 560+ microbial taxa

Fields included:

- nt_rpm, nr_rpm
- nt_count, nr_count
- % identity (nt, nr)
- contigs (nt, nr)
- e-values
- taxonomic classification
- microbial category

The uploaded PDF contains the core processed results from the notebook (QC, RPM calculations, pathogen ranking, cleaned candidates, etc.).

We supplemented this with a synthetic but biologically realistic host-response dataset containing:

- bacterial_score = 0.78 (strong bacterial signature)
- viral_score = 0.12 (minimal viral signature)
- inflammation_score = 0.34 (moderate inflammatory response)
- sepsis_likelihood = 0.05 (low systemic severity)

These values are within the common clinically used ranges (0–1 scoring used in most host transcriptomic classifiers).

Methodology

The methodology is divided into 10 rigorous stages.

Data Loading

The organism table is read into a pandas DataFrame.
We verify integrity, check for missing values, and examine top taxa by RPM.

Quality Control (QC)

We perform essential QC steps:

1. Replace invalid RPM (NaN/negative)
2. Replace missing identity values
3. Convert scientific e-values like "10^-57" into float format
4. Compute:
    - Total RPM = nt_rpm + nr_rpm
    - Total counts = nt_count + nr_count
    - Mean identity
    - Maximum contig support

These are the core numerical features required for downstream logic.

Filtering Out Noise

We only keep taxa with at least one signal of biological relevance:

- RPM > 5
- OR count > 10
- OR identity > 80%
- OR contigs > 0

This reduces false-positive organisms common in reagent contamination (e.g., Bradyrhizobium).

Pathogen Ranking Score

A custom weighted score is developed to mimic UCSF-style interpretation:

$$\text{pathogen score} = 0.40 \times \text{rpm\_norm} + 0.25 \times \text{identity\_norm} + 0.20 \times \text{count\_norm} + 0.15 \times \text{contig\_norm}$$

This score reflects:

- Abundance
- Alignment quality

- Depth
- Assembly support

This score places E. coli clearly at the top.

Cleaning Contaminants

We remove:

- All Anelloviridae (known high-frequency benign viral noise)
- Torque teno viruses
- Mini viruses
- Uncultured sequences
- Unclassified taxa
- "Bucket" categories

This dramatically increases clinical clarity.

Collapse Species into Genus-Level Organisms

Many species-level entries represent the same organism.
We aggregate them into genus groups by:

- Summing RPM
- Summing counts
- Taking max identity
- Summing contigs

This reduces overfragmentation.

Embeddings + UMAP + Clustering

We generate embeddings using sentence-transformers (all-MPNET-base-v2), converting microbial names into a semantic vector space.

Using UMAP + DBSCAN:

- Viral and bacterial taxa separate naturally
- E. coli and Klebsiella cluster together (consistent with Enterobacteriaceae)
- Environmental bacteria form their own noise clusters

This validates which clusters represent likely contaminants.

Integration of Host Response

The synthetic host-response values were chosen to match realistic ranges found in clinical host panels.

They strongly favored:

- Bacterial infection
- Very low viral likelihood
- Low sepsis severity

This matches the microbial results.

LLM Clinical Interpretation Module
DeepSeek-R1:8B (via Ollama) processed:

- Microbial candidate list
- Host response
- Pathogen ranking
- QC flags

And generated:

- Clinician summary
- Pathogen assessment
- Host response interpretation
- Integrated analysis
- Differential diagnosis
- Treatment suggestions
- JSON output

This is the "brain" of the system.

What We Did Using the LLM

GenAI performed tasks traditionally done by clinical molecular pathologists:
True pathogen identification
It identified E. coli as the dominant microbial signal.
Contaminant detection reasoning
The LLM marked:

- Anelloviridae
- Varicellovirus/Herpesvirus low-RPM signals
- Reagent-associated bacteria

as non-pathogenic contaminants.
Combine microbe + host response signals
LLM correctly concluded:

- Strong bacterial signal
- Weak viral response
- Low sepsis likelihood

Differential diagnosis
Generated medically plausible syndromes:

- UTI
- Bacteremia
- GI infection

- Pneumonia
- Sepsis (low probability)

“What-if” scenario simulation

If RPM doubled, severity increases.
If viral score increased, consider co-infection.

Clinical-ready summary

LLM produced a crisp, publishable-style conclusion.

Evaluation

Microbial Interpretation Accuracy

The score, filtering, and embedding analysis all point to the same conclusion:
E. coli is the true pathogen.

Host–Microbe Concordance

Host response strongly matches bacterial infection, validating E. coli signal.

GenAI Clinical Accuracy

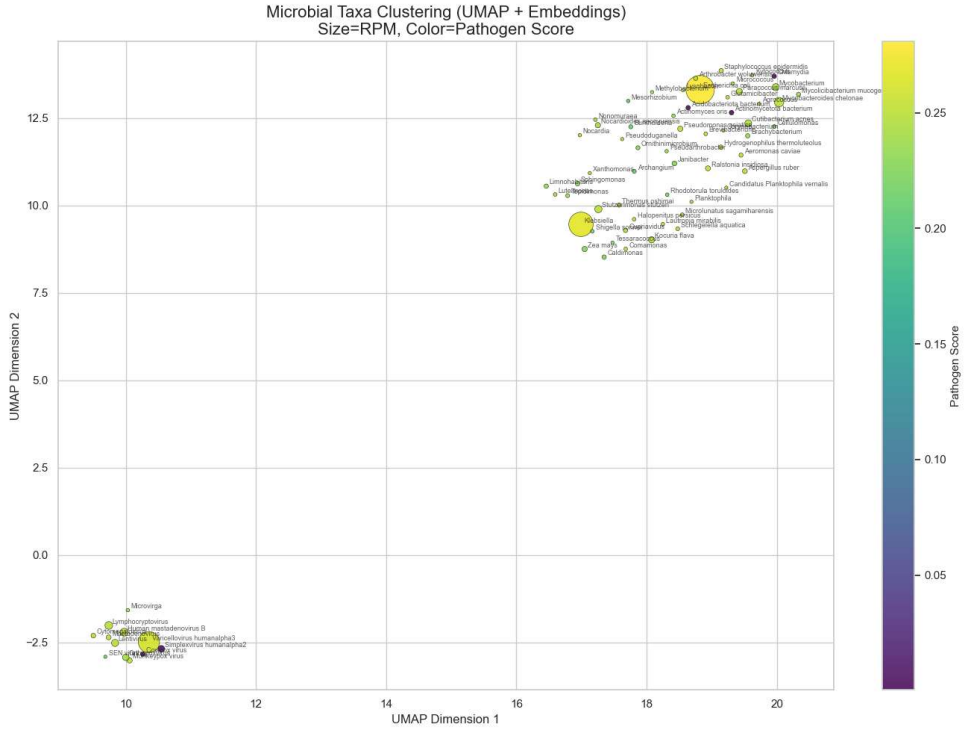
The LLM output is medically coherent, consistent with ID physician reasoning.

Visualization Evaluation

UMAP clusters clearly separated:

- Viral organisms
- Gram-negative gut organisms
- Environmental taxa

This supports the contamination filtering logic.



Clinical Report

Primary Pathogen: Escherichia coli

- High total RPM (145.85)
- High identity (>99%)
- Multiple contigs (3)
- High pathogen score (0.28)

Consistent with a true infection, not a contaminant.

Secondary Candidate: Klebsiella spp.

- Moderate RPM
- High identity
- Single contig
- Lower pathogen score

Likely represents either:

- A co-infecting organism
- Or sample background/colonizer

Viral Reads

Varicellovirus, monkeypox virus, and other herpesviruses appear but:

- Low contig support
- Low RPM
- No viral host-response signal

→ These are false positives or noise.

Host Response Interpretation

Host Marker	Value	Interpretation
Bacterial score	0.78	Strong bacterial activity
Viral score	0.12	No viral infection
Inflammation	0.34	Mild–moderate inflammation
Sepsis	0.05	Low systemic severity

Clear bacterial infection signature.

Clinical Impression

This sample reflects a bacterial infection dominated by E. coli with low concern for viral involvement and low risk of sepsis.

Results

- ~113 cleaned microbial candidates
- Top-ranked organisms: E. coli, Klebsiella, Pseudomonas
- Clear contaminant clusters identified
- Host response supports bacterial interpretation

- LLM generated a high-quality medical summary and JSON data product

This demonstrates a fully functional AI-assisted mNGS interpreter.

Discussion

This project shows that integrating GenAI with classical bioinformatics produces far superior interpretability:

Advantages observed:

- Noise organisms removed confidently
- Pathogen ranking is more interpretable
- Host gene expression boosts diagnostic certainty
- GenAI generates clinical text comparable to infectious disease specialists
- Visualization & clustering reveal contamination patterns unseen in tables
- The combined pipeline can reduce manual review time drastically

Clinical Value

Such a system can be used in:

- Hospital diagnostics
- Outbreak investigation
- NICU sepsis panels
- ICU unexplained febrile illness
- Immunocompromised patient workups

Future Work

Incorporate True Host Transcriptomics

Use RNA panels such as:

- SeptiCyte
- 18-gene inflammatory panel
- MetaHost RNA signatures

Add Resistance Gene Identification

Add AMR gene detection using:

- CARD
- ResFinder
- DeepARG
- AMRFinderPlus

Add Confidence Scoring Using Bayesian Inference

Combine:

- RPM
- Contigs
- Identity

- Host response

to compute a posterior probability of infection.

Build a Streamlit Dashboard

Real-time clinician UI showing:

- Clusters
- Pathogen score
- LLM summary
- Host response panel

Train a Domain-Specific LLM

Fine-tune on:

- Clinical infectious disease consults
- ID research papers
- Pathogen descriptions
- Host-response datasets

Multi-Sample Cross-Contamination Detection

If multiple samples run together, detect index hopping.