eDNA Biodiversity Analysis — Executive Summary

Date: 2025-09-30

Audience: Non-technical stakeholders

Overview

We built a user-friendly system to analyze environmental DNA (eDNA) data and summarize biodiversity. The solution combines proven bioinformatics (BLAST) with modern machine learning to identify organisms in a sample and present results through an interactive web interface.

What we delivered

- End■to■end analysis pipeline that runs on Windows
- Web application with a Home page, Results Viewer, Run Browser, and Taxonomy Viewer
- Hybrid identification that leverages both ML-based similarity and BLAST database lookups
- Clean storage layout for datasets and runs for easy review and sharing

Key results (current run)

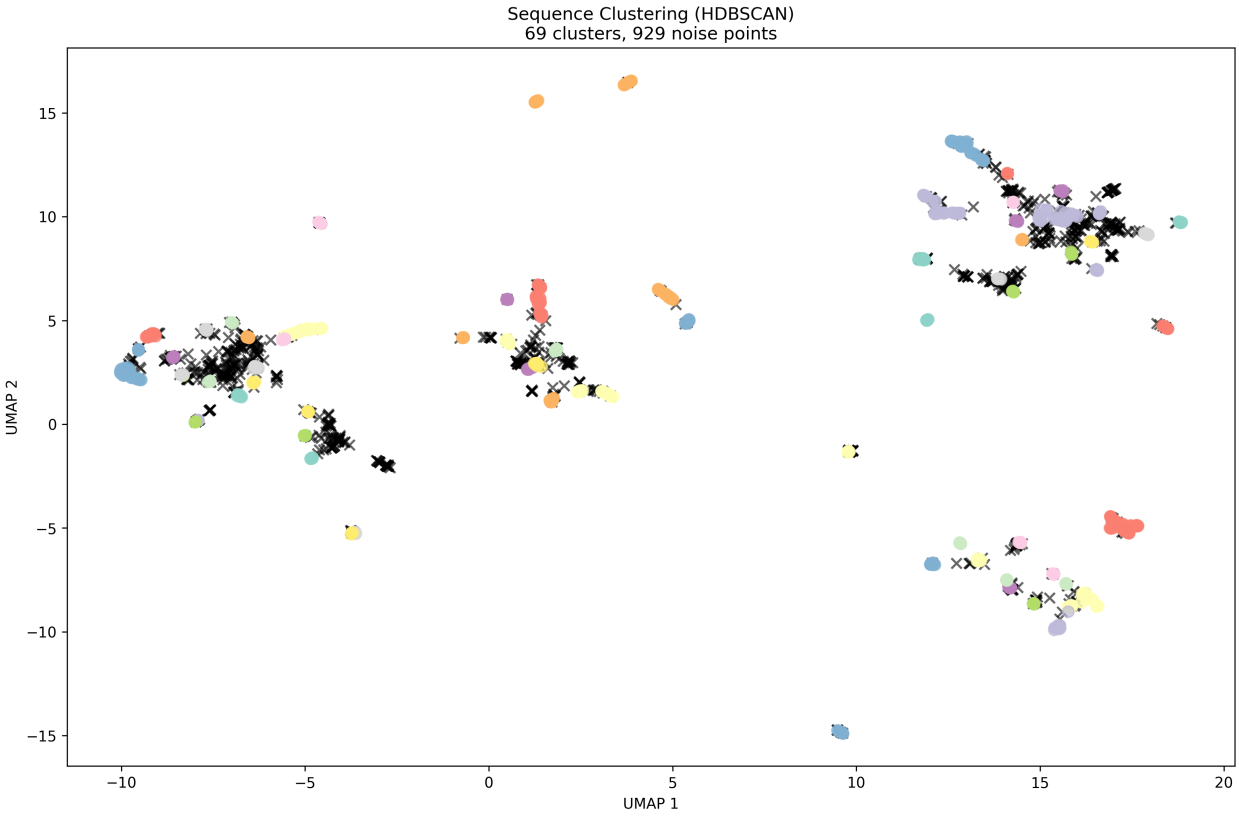
- Dataset: SRR35551197 (subset of 2,000 sequences processed for taxonomy)
- Time: Taxonomy step completed in ~31 seconds (subset)
- Findings: 61 unique taxa identified (with enriched scientific lineage)
- Reliability: BLAST taxonomic IDs prioritized when available; ML-only names used as fallback

Why this matters

- Faster, clearer biodiversity assessments for environmental samples
- Auditable, reproducible outputs with clear lineage information
- A web UI that allows exploration and sharing without command-line expertise

Visual snapshot

Figure 1. Clustering overview (representative layout)


Sequence Clustering (HDBSCAN)
69 clusters, 929 noise points
UMAP 2
UMAP 1
15
10
5
0
−5
−10
−15
−10
−5
0
5
10
15
20

How it works (in plain English)

- We transform each DNA sequence into a numeric “fingerprint” using a modern language model for DNA.

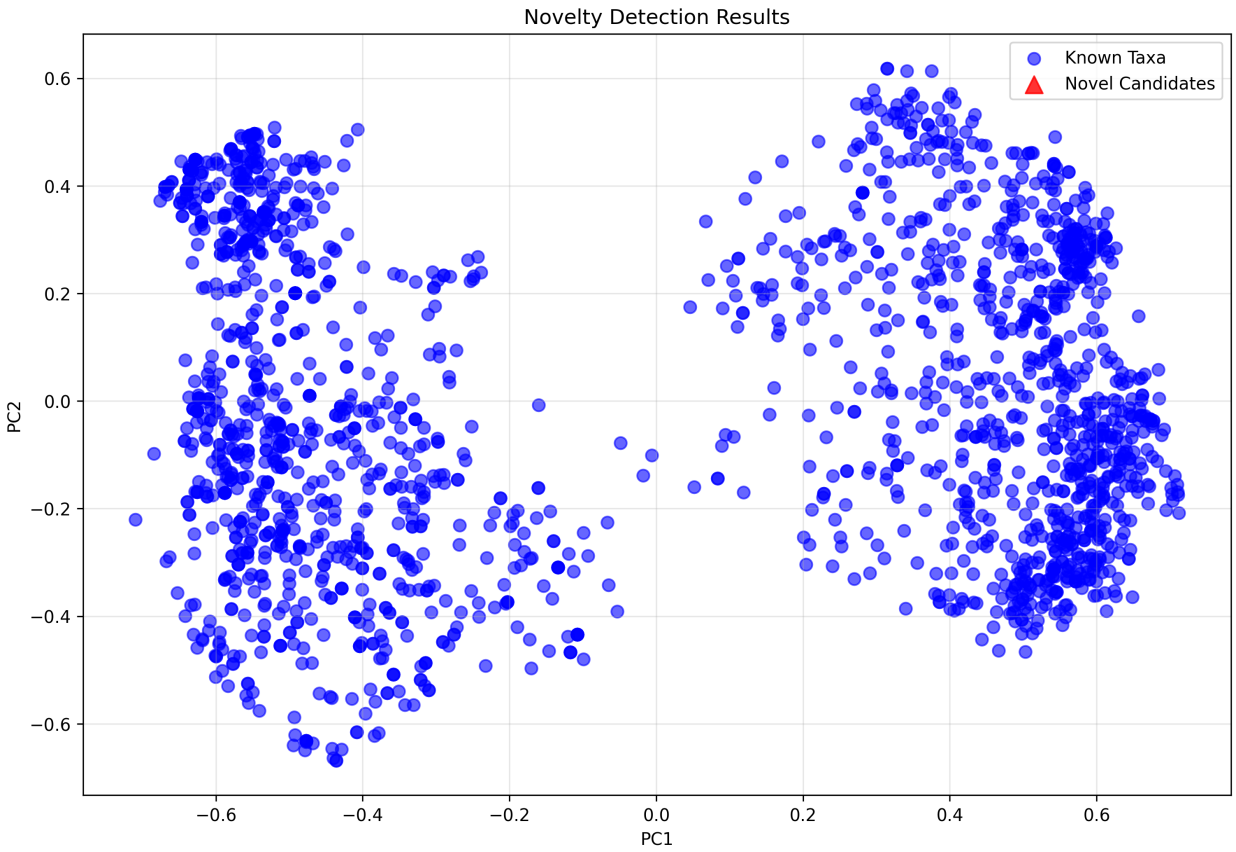
- We compare these fingerprints to known references and also run BLAST lookups.

- When BLAST provides a direct taxonomic ID, we use it to fetch full scientific lineage (species up to kingdom).

- When BLAST is uncertain or missing, we rely on the model’s nearest neighbors and take a conservative consensus.

Additional snapshot

Figure 2. Novelty view (no novel candidates under current thresholds)



Practical uses

- Rapid biodiversity summaries for environmental surveys

- Triage of large datasets—quickly see what’s common vs. rare

- Shareable, interactive results for collaborators and stakeholders

Limitations (current setup)

- Large, full■scale runs are slow on CPU; a GPU will make embedding much faster

- Advanced clustering (HDBSCAN) may require extra dependencies on some systems

What’s next

- Speedups for large datasets (GPU use or CPU-side optimizations)
- Run-to-run comparisons in the UI
- Optional filters and confidence thresholds for taxonomic calls

Where to find things

- Runs and reports: F:\AvalancheData\runs
- Web UI (when launched): http://localhost:8501
- Quick taxonomy CSV for the UI: C:\Volume D\Avalanche\results\taxonomy\taxonomy_predictions.csv