Soil is Alive: Biological, Physical, and Chemical Analysis

SAVANAH SENN*[1], LES VION[1], MEIKA BEST[1], MATHEW KOSTOGLOU[1], KARU SMITH[1], ADRIANNA L. BOWERMAN[1], DAILA MELENDEZ[1,2], JILLIAN M. FORD[1], RAY E. ENKE[3], KAREN B. BARNARD-KUBOW[3], TONY DEVEYRA[4], ARIANNA BOZZOLO[4], BRUCE NASH[5]

[1]*Department of Agriculture, Los Angeles Pierce College, 6201 Winnetka Avenue, Woodland Hills, CA 91371 USA*

[2]*Horticulture Department, Oregon State University, Corvallis, OR, 97331*

[3]*Biology Dept., James Madison University, Harrisonburg, VA, 228407*

[4]*Rodale Organic Institute, Camarillo, CA, 93030*

[5]*DNA Learning Center, Cold Spring Harbor Labs, Brooklyn, NY 11201*

**stclais@piercecollege.edu*

Abstract: We analyzed soil samples from LA Pierce College and local farms. To assess life in the soil, 16S metabarcoding was performed. Industrial Hemp was grown traditionally at Western Fiber in Tulare CA; organic JinMa, a fiber variety, and CBD-type rootzone soil from The Rodale Institute at Camarillo CA was sampled. Noxious weeds in Marquis field at Pierce, soil from Pierce Arboretum, and a fallow proposed vineyard were also sampled. The purpose was to investigate the physical and chemical properties of soils from farms in California and to discover plant growth-promoting and antibiotic-producing bacteria from the rootzone. The hypothesis was that each environment would have a unique microbial community, and each of the fields would have similar fertility levels since they shared a prime farmland classification. We also hypothesized that the organic farm would have more diverse and abundant microbes performing ecological functions compared to the traditionally farmed soil. The main results indicated that there were significant differences in nitrogen, potassium, pH, and TDS between pairs of fields. The hypothesis was partially supported; organic matter and phosphorus levels were similar across farms. Potential plant growth-promoting bacteria and nitrogen fixers differed in proportion between sites. Greater than 80% of the *Chitinophaga* and 80% of the *Agromyces* reads identified in the study were from Camarillo, as well as more than 80% of the reads for *Pseudomonas* and greater than 60% of the *Massilia* sp. reads. The hypothesis that each environment would have a distinct community was partially supported. The hypothesis that the organic farm would have higher abundance and variety of microbes performing these functions was supported. The microbiome of hemp seemed to be more influenced by soil type and cropping method, rather than the genetic background of the cultivars. Soils need to be conditioned to retain their fertility. Periodic monitoring and monitoring before, during, and after cropping is recommended.

Keywords: metabarcoding, soil science, DNA sequencing, Agroecology, Microbial Ecology

Introduction

Soil Physical Properties

The physical properties of soil are an important factor in determining which crops to plant. Soil's physical properties are properties that can be seen and felt. Physical properties do not directly refer to chemical or biological properties but are still affected by both [1]. According to the Cornell University Agronomy Fact Sheet, "Soil texture determines the rate at which water drains through a saturated soil; water moves more freely through sandy soils than it does through clayey soils" [2]. From the different experiments, we expect to find the soil to have a moisture content between 10%- 45% and an organic matter content between 5%-10% according to previous research done on soils. Moisture is thought to be one of the most important, perhaps because it is the most immediately tangible characteristic of soil.

EC, pH, and TDS

pH represents the intensity of acidity in soil; calcium-saturated organic soil typically has a pH of 7.2 to 7.8, which is mild to moderately alkaline [3]. Regarding EC values, "Apparent soil electrical conductivity (EC) has shown promise as a soil survey tool" [3]. This means monitoring over time is necessary. TDS (Total Dissolved Solids) levels are important because, "Irrigation can contribute a substantial amount of salt to a field over the season" [4].

The accumulation of soluble salts may be an issue at 500 ppm for some sensitive plants, but is usually tolerated by most plants up to 1000 ppm. In these instances, soils are typically high in chlorides or sulfates. EC gives a more accurate measurement of soluble salts because electricity would not be conducted by pure water.

pH is important because it may ultimately lead to crop failure. The accumulation of salts is known to be associated with high pH. Excessive salts can lead to fertilizer burn. The pH of the soil is the most important factor for plants in terms of their ability to incorporate soil nutrients. Even if the necessary nutrients are present in the soil, they are not available to plants if the pH is not in the optimal range for that nutrient [5]. A good example is iron which is not available at high pH. Many soil elements change form because of reactions in the soil. These reactions, controlled by pH, alter the solubility of nutrients [1].

Soil Nutrients

The primary macronutrients, sometimes called fertilizer elements, are often not available in large enough amounts for optimal growth. The three primary nutrients— nitrogen, phosphorus, and potassium—are added to the soil by fertilization with synthetic mineral salts or organic amendments such as manure [1]. Visual cues may be used to detect excess or deficits but are unreliable for proper diagnosis. Many factors can cause issues such as chlorosis, and different species may take on different appearances. The fertilization regimen can become more efficient with soil testing, reducing the effects of contamination through runoff as well as reducing production costs. Keeping track of nitrogen, phosphorous, and potassium is important beyond the health of crops; it is also important in terms of protecting the environment.

Nitrogen and phosphorus pollution can both cause serious damage to surface waters. Nitrates cause eutrophication; phosphates destroy habitat. Potassium may create an environment that is too salty. Results from field studies indicate that rational use of manure and mineral fertilizers can help reduce the pollution problems arising from livestock farming practices [6]. In such systems, defining N rates where yields are maximized, and environmental harm is minimized could benefit both food production, human health, and the environment [7]. Testing the soil is the best way to find out how much fertilizer needs to be added to grow a productive crop without adding too much, which can contribute to input costs and environmental damage [5].

Fluctuations in nutrient levels within the soil can act as catalysts for proliferation of specific bacterial strains, shaping the microbial landscape. The interplay between nutrient availability and microbial diversity offers a promising avenue for understanding the relationships driving bacterial evolution. Recognizing the intricate connections between nutrient levels and microbial diversity provides insights into sustainable agricultural practices. Unraveling the impact of soil nutrients on bacterial discovery contributes not only to scientific understanding but also to the development of eco-friendly farming approaches.

Life in the Soil

Industrial hemp (Cannabis sativa) is a fiber crop with impressive carbon sequestration potential. For example, 1 ton of hemp stalks contains about 0.445 tons of carbon absorbed as a gas or 1.63 tons of CO2 [8]. The cannabidiol varieties of hemp are also purported to produce products that have radical scavenging capabilities. Thus, it may be useful for building materials or be therapeutically relevant. The rootzone microbes of industrial hemp have been little studied in a field crop setting. Previous research showed that rhizosphere microbes in cannabis are cultivar-specific. However, it was unknown to what extent soil types or cropping systems influence these microbes.

The technique used to assess this in our study was a comparative analysis of 16S soil metabarcoding results from 6 samples from each of two sites; in Camarillo, CA at Rodale Organic Institute where JinMa and CBD (cannabidiol) types of hemp were grown, and in Tulare, CA at Western Fiber's south side field, where the YuMa fiber variety was grown and harvested. The Tulare site had a history of traditional synthetic farming methods.

Purpose and Hypotheses

The purpose of this study was to discover plant growth-promoting and antibiotic-producing bacteria from the rootzones of plants. The hypothesis was that each soil environment would have a unique microbial community associated with it. This hypothesis was developed by considering the differences in geography and the fact that the soil is sandier at the Tulare site and contains more clay at the Camarillo site. We also hypothesized that the organic farm in Camarillo would have more diverse and abundant microbes performing these essential functions, when compared to the traditionally farmed soil in Tulare. This is because organic farming is generally considered to be more biodynamic than traditional or synthetic farming methods.

Overview and Importance

The soil metagenome consists of the 16S ribosomal genes detected in a soil sample by Next-Generation Gene sequencing techniques (Massive Parallel Sequencing, such as Illumina). The metagenome is a snapshot, both qualitatively and quantitatively, of the microbial community present in the soil. Biologic activity is an important indicator of the soil's health. The great multitude of functions of the soil's microbiota is impressive. Major planetary nutrient cycles are involved and dependent on the microbiota of the soil and are of vast importance to both natural ecosystems as well as agricultural endeavors. Experiments that assess the metagenome are of great importance in informing best practices associated with organic and regenerative farming.

The soil hosts billions of bacteria, but not all species can be readily cultured [9]. The number of culturable bacterial cells in soil is generally believed to be only about 1% of the total number of living cells present. However, in environmentally stressed soils, the number of culturable bacteria may be as low as 10^4 cells per gram [10]. This emphasizes the importance of metabarcoding and other cultivation-independent techniques. Metabarcoding as it now exists is novel, particularly in its ability to detect thousands of taxa quickly in a single sample and in an economically feasible fashion - opening gateways for research to further understand plant-microbe interactions, plant physiology, microbial characteristics, and anticipating potential threats from pathogens. There is also the potential to identify new species, which may be important to agricultural technologies.

The Rodale Institute has been involved in the research and development of organic regenerative agriculture and has not applied synthetic fertilizers, pesticides, or herbicides to any of the fields that have been sampled. Furthermore, the Rodale Institute has experimented with various cropping systems and policies of leaving crop residues post-harvest, cover-cropping, green manuring, and not allowing fields to lie fallow. The recent history of Western Fiber's field is in line with the current agricultural method of utilizing synthetic herbicides, insecticides, and fertilizers, with fields and orchards typically existing as monocultures and often left fallow during non-growing seasons. Conventional tillage systems have been shown to break down soil structure and diminish the quantity of soil bacteria [1].

Methods

Soil Physical and Chemical Properties

During Fall 2021- Spring 2022, 51 soil samples from Los Angeles Pierce College farm and local farms were analyzed for soil chemical and physical properties. Industrial Hemp YuMa plots in Tulare grown at Western Fiber were sampled for soil. In Camarillo, JinMa fiber variety and CBD type rootzones were sampled for soil. Noxious weeds growing in Marquis field at Pierce College, soil from the Pierce Arboretum, and soil from fallow ground from a proposed vineyard site were also sampled.

A B

Fig. 1.The Google Earth images of the sampling locations at Rodale Organic Institute (A) and LA Pierce College (B) are pictured.

Moisture, organic matter, EC, TDS, pH, N, P, and K were determined using the methods described in the St. Clair Soil Science Lab Manual. Moisture % was determined by the oven dry method. Organic matter percentage was determined by dry combustion. A 1:5 soil-to-water ratio was used for soil pH measurements. N, P, and K were determined using Hach color-changing reagents, and %T was measured using a Genesys 20 spectrophotometer. Soil physical and chemical properties from the sites were also measured, including EC, TDS, and texture by touch. Colony forming units (CFUs) were also quantified. Comparisons between fields were carried out in the R stats package. Since there was an indication of non-normality and unequal variance, the nonparametric Kruskal-Wallis regression method was used to analyze the data.

Soil Metabarcoding

During Fall 2021, 22 soil metabarcoding samples were collected from Los Angeles Pierce College farm and local farms, DNA was extracted, and samples were sequenced on Illumina. The samples and their associated locations are referred to as follows: JinMa and CBD are from Camarillo, Ventura County, and Industrial Hemp N_side, and S_side samples are from Lemoore, Tulare County.

The data consists of 16S amplicons. DNA was extracted with the Qiagen (Hilden, Germany) Power Soil DNA kit. Quantification of the DNA extracted was achieved with spectrophotometry using the Thermo Fisher MultiSkan SkyHigh Microplate Spectrophotometer. DNAs were then sent to James Madison University for 16S amplification and library preparation, and 16S amplicon NGS was pooled on the Illumina (San Diego, CA, USA) MiniSeq platform. The V4 region of the bacterial 16S rRNA gene was amplified and barcoded for each sample using the primers developed by Kozich et al [11]. Samples were pooled. A double-sided bead cleanup was carried out. The quality and concentration of the pooled library were checked using a Bioanalyzer (Agilent, Santa Clara, CA, USA) and NEB's Library Quant Kit for Illumina. The library was sequenced on a MiniSeq using a mid-output reagent cartridge. Before loading, the library was combined with Illumina's PhiX control (30:70 16s: PhiX) to ensure a high-quality run despite the problem of relatively low diversity in the 16S library. Metabarcoding data analysis was carried out using the DNA Subway Purple Line [12] using QIIME 2 [13] and Emperor [14]. DNA Subway was used for bioinformatics; the Purple Line analysis implemented DADA2 and QIIME2 for quality control, alpha rarefaction, and the output of the ASV table for

taxonomic diversity analyses. Using MS Excel pivot tables, the relative frequency of reads from different functional categories of microbes was visualized [15].

Results and Discussion

Soil Physical and Chemical Properties

Data showed that the conditions between the fields differed significantly in values for organic matter, N, K, TDS, and pH ($p<0.05$); this offered a diverse panel of substrates for the discovery of bacteria performing beneficial functions to plants and humans. A visual representation of the correlation matrix is given in Figure 2. It is apparent that Potassium concentrations are making the highest contribution to EC and TDS from the elements measured, based on the Pearson correlation matrix. It also appears that the CFU concentrations for the 10^-5 dilution were mildly associated with higher N and P levels. As expected, EC and TDS values were 100% correlated.

According to the results of nonparametric Kruskal-Wallis regression [16]. Field was associated with OM ($p=0.02$). There was a trend toward the Arboretum and Marquis C samples exhibiting the highest OM percentages. However, the results of the pairwise Wilcoxon rank sum test indicated that there was no significant difference in OM between the groups, based on the adjusted p-values using the BH correction (BH correction= Benjamini-Hochberg false discovery rate).

According to regression results, Field was associated with N ($p=0.02$). There was a trend toward higher nitrogen levels at the proposed vineyard site, and the lowest N levels were from Rodale Institute. Most samples had low values for Nitrogen in ppm. No significant pairwise differences were detected when using the BH-corrected p-values and a cutoff of 0.10.

Field was associated with K ($p=3.24*10^{-5}$). In the pairwise Wilcoxon rank sum test, the Potassium levels of Marquis A soil samples varied from the Arboretum soils, and Western Fiber soils from Tulare differed from Rodale soil K levels, using a marginally significant p-value cutoff of 0.10. According to the results of nonparametric Kruskal-Wallis regression, Field was associated with pH ($p=0.002$). In the pairwise Wilcoxon rank sum test, the pH of Marquis A soil samples varied from the Arboretum soils, and Western Fiber soils from N. side Tulare differed from Rodale soil pH levels, using a marginally significant adjusted p-value cutoff of 0.10.

According to the results of nonparametric Kruskal-Wallis regression, Field was associated with TDS ($p=8.49*10^{-6}$). In the pairwise Wilcoxon rank sum test, the TDS of the Arboretum soils was significantly different than N. side Tulare field soils ($p\text{-adj}<0.05$). Furthermore, the Marquis C Drain soil samples varied significantly in TDS from the Marquis A soils ($p\text{-adj}=0.03$), and Marquis A was significantly different in TDS than N. Side Tulare fields. Marquis B and Rodale samples were also significantly different in TDS when compared with Tulare N. Side samples ($p\text{-adj}=0.03$). The highest TDS values were from Marquis B and the proposed vineyard site (Figure 2). The BH correction was applied.

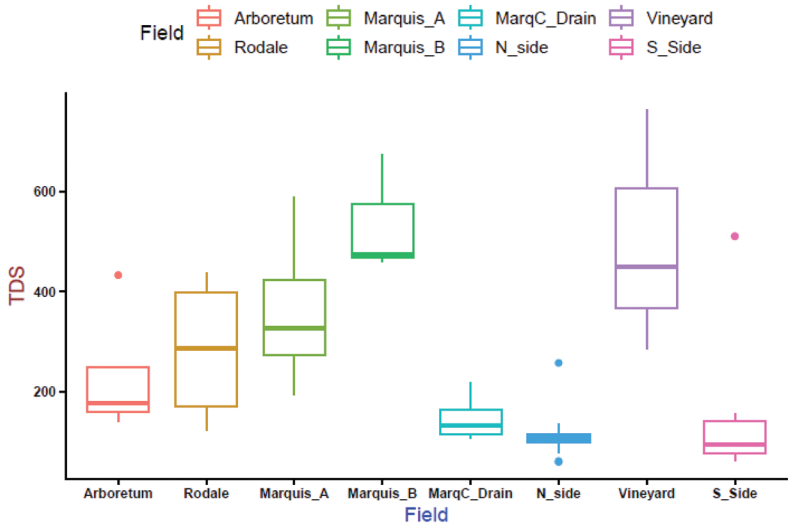

Fig. 2. A comparison between the TDS values of the fields is shown.

The tested sites, characterized by the lowest nitrogen levels among the fields, present an intriguing scenario. Nitrogen scarcity often acts as a selective pressure on microbes, encouraging the emergence of bacteria with specialized adaptations for nitrogen acquisition. Microbial communities, being highly adaptable, strategically adjust their physiology based on nutrient availability. Potassium influences microbial dynamics, and the correlation between potassium concentrations and Electrical Conductivity (EC) raises the question of whether microbial abundance may respond to varying potassium levels. Colony Forming Unit (CFU) concentrations for the 10^-5 dilution exhibit a subtle association with higher nitrogen (N) and phosphorus (P) levels. This suggests a microbial response to varying nutrient levels, hinting at a potential correlation between nutrient availability and microbial abundance.

Some bacteria can increase iron availability in conditions of high pH. *Bacillus*, which was detected in our Tulare samples, is a genus known to catabolize and exude siderophores (compounds acting as iron chelators) into the rhizosphere, which create Fe availability in soils that would otherwise have too alkaline a pH for sufficient iron mobility in the soil, thus increasing the upper limit of the pH range tolerable to any specific crop; iron availability is naturally limited at high pH [17].

Beta Diversity Analysis

In the beta diversity analysis using the unweighted UNIFRAC distance for a principal coordinates visualization of the 16S data, the Camarillo samples from both CBD and JinMa varieties appeared to be very similar based on their genetic distances (Figure 3). Meanwhile, the cluster of samples from Camarillo appeared to separate from the Tulare samples. The Tulare industrial hemp-planted and postharvest plots also appeared very similar, based on their projected distance on the PrinCoA graph.

Principal coordinates analysis is a method commonly employed in the analysis of beta diversity, which is the similarity or dissimilarity between samples or groups of samples. The UNIFRAC distance has been shown to be an effective distance metric in simulation studies [18]. The first three axes of the principal coordinates accounted for 42.54% of the variation. Industrial hemp associated soil samples from Tulare tended to be low on Axis 2 and high on Axis 3, whereas the soil samples from Camarillo were high on Axis 2 and high on Axis 1. The postharvest samples from S. Side 6 appear to be like the industrial hemp plots. These samples, six samples from Camarillo and six samples from Tulare will be the focus of the remainder of the metabarcoding analysis, due to this interesting observation and the lack of sufficient replicates for the Pierce farm samples, some of which failed sequencing.

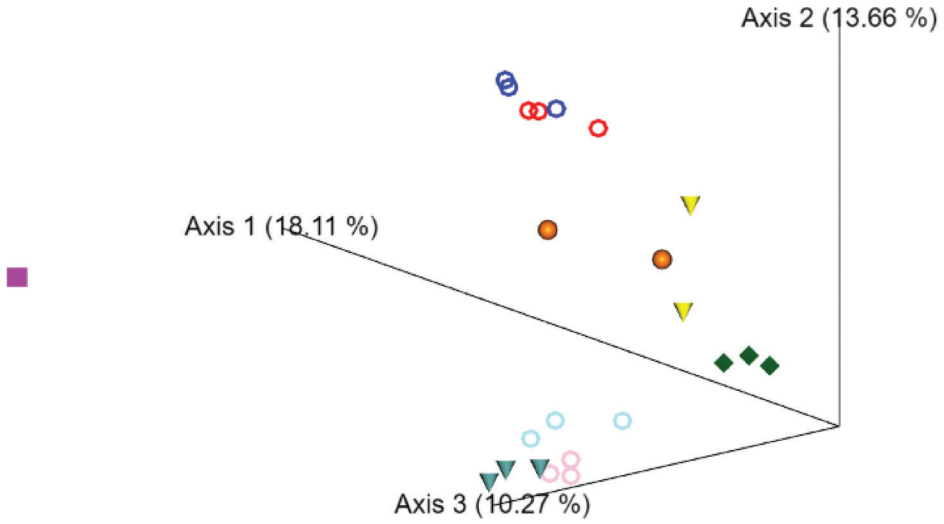

Fig. 3. The beta diversity comparison involved visualizing the first three axes of Principal Coordinates Analysis. The distance between sample groups reflects the genetic distance (between group diversity). Pink rings: Tulare S. Side Plot 3 Yu-Ma Industrial Hemp; Light Blue rings: Tulare S. Side Plot 2 Yu-Ma Industrial Hemp; Green cones: Tulare S. Side Plot 6 Postharvest; Dark Green diamonds: Tulare N. Side Plot 6 Fallow; Orange spheres: Los Angeles Marquis A fallow; Yellow cones: Los Angeles Ridge Vineyard Site Fallow; Red rings: Ventura CBD Trial; Blue rings: Ventura JinMa trial; Purple square: positive control.

Analysis of Bacterial Taxonomic Abundance

To gain further insights into which genera may be performing essential ecological functions at the sites where industrial hemp was grown, Excel pivot table analysis was performed using the methods described in the Pipeline for Undergraduate Microbiome Analysis [15]. The bacterial genera found in the different samples were compared and visualized. The functional classification was based on their putative ecological functions from the literature.

In terms of potential antibiotic producers, the genera Actinoplanes, Lysobacter, Bacillus, and Pseudomonas were well represented in the Camarillo samples. However, in Tulare samples, Actinomadura, Bacillus, and Streptomyces sp. had the highest abundance of reads. Low counts were also present for Lysinibacillus. The genus Lysinibacillus includes entomopathogenic bacteria that produce antifungal and antibacterial compounds [19].

Serratia species such as S. marcescens have been shown to promote growth in wheat seedlings and have a full suite of active nitroreductases [20]. Streptomyces are known for denitrification abilities, which is part of its suite of plant growth promotion characteristics [21]. *Vibrio* is a pathogen that uses nitrate reduction as a strategy. *Vibrio cholerae* uses nitrate reduction to determine how much it should grow its population under anaerobic and changing pH conditions [22]. Similarly, *Vibrio* sp. collected from sediments of Scottish estuaries have been shown to produce nitrite and ammonia from nitrate [23]. *Pseudomonas* such as *P. aeruginosa* are able to denitrify soil under anaerobic conditions [24]; a new nitrate-reducing species, *Pseudomonas oligotropha*, was recently discovered [25]. *Achromobacter* has denitrifying bacteria such as *A. denitrificans*, which lives only in aerobic conditions and was isolated from soil [26]. Bacilli such as *B. subtilis* are valued for their biofertilizer applications, including their ability to reduce nitrogen in agricultural soils, which reduces the loss of nitrate that typically occurs readily following a precipitation event [27]. Nitrogen reducers were represented by *Achromobacter, Bacillus, Pseudomonas, Serratia, Streptomyces*, and *Vibrio* in the Camarillo samples. In the Tulare samples, nitrogen-reducing bacteria were represented by *Bacillus* and *Streptomyces*.

In terms of potential plant growth-promoting bacteria and nitrogen fixers, 100% of the *Brevibacillus* reads identified in the study were in the Camarillo soil samples. Greater than 80% of the *Chitinophaga* and 80% of the *Agromyces* reads identified in the study were from Camarillo, as well as more than 80% of the reads for *Pseudomonas* and greater than 60% of the *Massilia* sp. reads. Furthermore, more than 70% of the reads for *Flavobacterium* sp. belonged to the Camarillo samples (Figure 4).

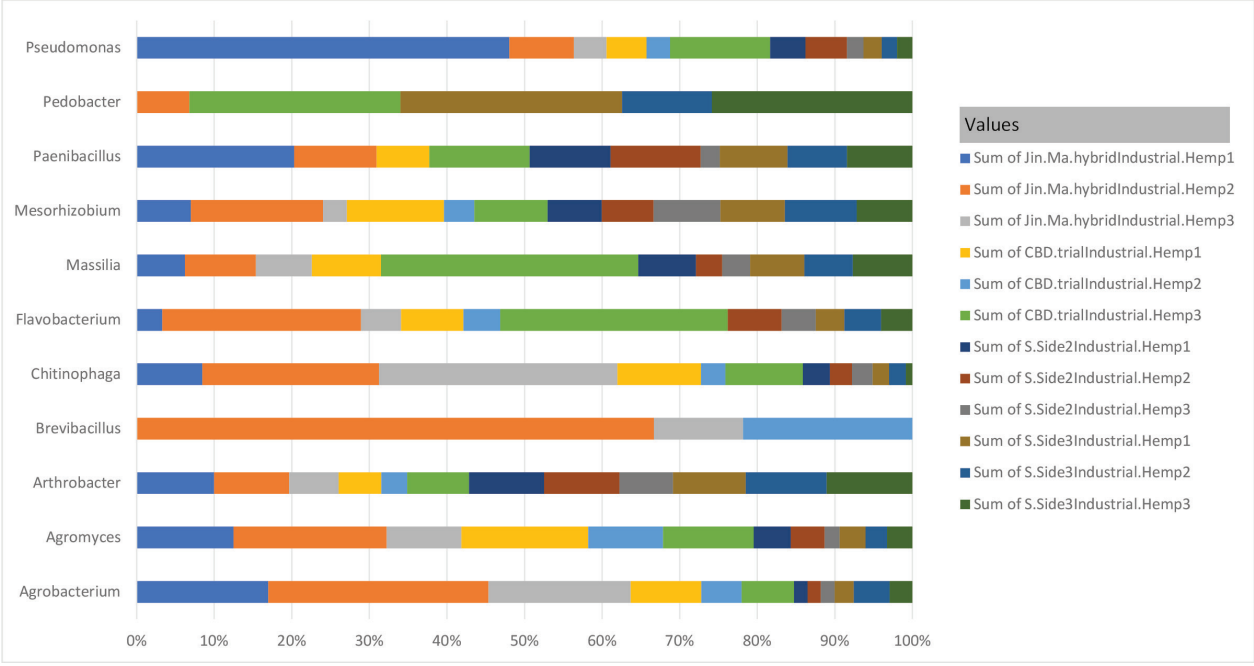

Fig. 4. The relative abundances of several potential plant growth promoting bacteria are shown for Tulare and Camarillo samples.

More than 60% of the reads from *Pedobacter* were from Tulare. It is interesting because *Pedobacter* and *Flavobacterium* are related, and known for antioxidant production but were not found to be equally abundant at both sites. Greater than 55% of *Arthrobacter* sp. reads were from Tulare. *Paenibacillus* reads were split evenly between sites; *Mesorhizobium* reads were also found in similar proportions at both soil sampling sites. Among the remaining suspected nitrogen fixing genera, reads were more prevalent in the Tulare samples for *Arthrobacter, Bradyrhizobium*, and *Devosia*. Conversely, *Sinorhizobium* sequences were only found in Camarillo soil samples, and of the sequences that were assigned to *Rhizobium*, 60% belonged to samples from Camarillo.

Potential prokaryotic contributors to essential ecological functions in the industrial hemp rootzone were identified. Although there were differences in the most prevalent species at each site, all communities had players in each of the functional roles studied here: Antibiotic production, nitrogen fixation, denitrification, plant growth promotion, and pesticide and insecticide degradation. Variations in the community may be explained by both the plants grown at the location and their associated rhizosphere, as well as the soil conditions and current cultural practices of agriculture, performed at those specific locations.

Bacteria from the plant rhizosphere alter plant growth, recruiting or repelling certain species. Soil biota from bulk soil provides the seed bank for the rhizosphere biota, creating a cycle. In this consideration, simply growing hemp or any other plant may act as a bioinoculant, which may modify edaphic soil conditions, which will, in turn, become an inoculant to future rhizospheres. This stresses the importance of not allowing a field to lay fallow for an extended period. As performed in another study [28], future work could involve the collection of bulk soil samples (edaphic soil), rhizosphere samples, and endosphere samples. A bulk soil sample would be useful to serve as a control by holding the fallow condition constant and reflecting the land use and soil factors. Future planned studies should work on the isolation of beneficial microbes that were unclassified to the species level from the organic farm in order to provide more tangible evidence of the claims made in this study. This would also lend itself to the potential discovery of novel soil bacteria. Including fungi in the next round of sequencing using an ITS (internal transcribed spacer) region [29] is also of interest.

Conclusion

The main results indicated that there were significant differences in nitrogen, potassium, pH, and TDS between pairs of fields. There was a subtle association between higher N and P levels and CFUs for the 10^-5 dilution. Therefore, the hypothesis was partially supported, particularly for organic matter and phosphorus levels, which were similar across farms. The hypothesis that each environment would have a distinct community was partially supported. The hypothesis that the organic farm would have a higher abundance and more variety of microbes performing these essential functions was supported. A surprising result was that the microbiome of industrial hemp seemed to be more influenced by soil type and cropping method rather than the genetic background of the cultivars.

This comparative analysis contributes to the literature a novel investigation of soil microbes in the industrial hemp rootzone. The limitations of the study involve the need to have a more rigorous experimental design and allow for more statistical analysis and biochemical testing. The sites with the unique combination of low nitrogen, significant potassium, and high pH have likely shaped a distinctive microbial community. This discovery emphasizes the intricate interplay between soil nutrient levels and microbial adaptations, paving the way for sustainable agricultural practices and microbial-based solutions for soil management. Soil needs to be conditioned, with cover crops and amendments over time to retain their fertility. Periodic monitoring before, during, and after cropping are necessary. In order to draw a strong conclusion about the significance and interaction between soil conditions and cultivar types, a design that tests soils from the same cultivar at multiple sites would be helpful in controlling for some of the variables.

Acknowledgments. This work was supported by training and resources from the National Science Foundation IUSE grant, NSF DUE-1821657, awarded to Bruce Nash and Ray Enke in part to support Community College faculty. This material is based upon work supported by the National Science Foundation under Grant Nos. DBI-0735191, DBI-1265383, and DBI-1743442, which supported iPlant Collaborative and DNA Subway development. Metabarcoding reads are from public project 7929 on Cyverse.org DNA Subway Purple Line.

Disclosures. The authors declare no conflicts of interest.

References

[1] E. J. Plaster, *Soil Science and Management International Edition*, Clifton Park: Delmar Cengage, 2009.

[2] Nutrient Management Spear Program, "Agronomy Fact Sheets. (2005-2020)," Cornell University. Accessed: Dec. 20, 2023. [Online]. Available: http://www.nmsp.cals.cornell.edu/guidelines/factsheets.html

[3] R. E. Lucas, and J. F. Davis, "Relationships between pH. values of organic soils and availabilities of 12 plant nutrients", *Soil Science* vol. 92, pp. 177-182, 1961, doi: 10.1097/00010694-196109000-00005.

[4] S. Grattan, "*Irrigation Water Salinity and Crop Production*", publication 8066, University of California, 2002. Available: https://www.waterboards.ca.gov/waterrights/water_issues/programs/bay_delta/california_waterfix/exhibits/docs/Islands/II_8.pdf

[5] S. St. Clair, M. Saraylou, and E. Maine, *Introduction to Soil Science Lab Manual*, 3rd ed. Pierce College, 2021. [Online]. Available: https://www.studocu.com/en-us/document/pierce-college/intro-to-soils/soils-lab-manual-8/83932958

[6] V. Hooda and S. K. Tehlan. "Effect of biofertilizers, FYM and nitrogen levels on seed yield and seed quality of coriander (Coriandrum sativum L.)," *Annals of Agri-Bio Research* vol. 19 no. 1, pp. 121-123, 2014, doi: 10.5555/20143063983.

[7] T. B. Sapkota *et al*., "Identifying optimum rates of fertilizer nitrogen application to maximize economic return and minimize nitrous oxide emission from rice–wheat systems in the Indo-Gangetic Plains of India," *Archives of Agronomy and Soil Science*, vol. 66, no. 14, pp. 2039–2054, Jan. 2020, doi: 10.1080/03650340.2019.1708332.

[8] K. Karamova, G. Galieva, N. Pronovich, P. Kuryntseva, and P. Galitskaya, "Use of superabsorbent plants for urban greening as a tool to sequester atmosphere carbon," E3S Web of Conf., vol. 463, p. 02008, 2023, doi: 10.1051/e3sconf/202346302008.

[9] L. L. Ling *et al*., "A new antibiotic kills pathogens without detectable resistance," *Nature*, vol. 517, no. 7535, pp. 455–459, Jan. 2015, doi.org/10.1038/nature14098.

[10] B. R. Glick, "Plant Growth-Promoting Bacteria: Mechanisms and Applications," *Scientifica*, vol. 2012, pp. 1–15, 2012, doi.org/10.6064/2012/963401.

[11] J. J. Kozich, S. L. Westcott, N. T. Baxter, S. K. Highlander, and P. D. Schloss, "Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform," *Applied and Environmental Microbiology*, vol. 79, no. 17, pp. 5112–5120, Jun. 2013, doi.org/10.1128/aem.01043-13.

[12] U. Hilgert, S. McKay, M. Khalfan, J. Williams, C. Ghiban, and D. Micklos, "DNA Subway: Making Genome Analysis Egalitarian," in *Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment*, New York, NY,: Association for Computing Machinery, Jul. 2014, pp. 1–3, doi: 10.1145/2616498.2616575.

[13] Bolyen, Evan et al., "Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2." *Nature biotechnology* vol. 37 pp. 852-857. Jul. 24. 2014, doi: 10.1038/s41587-019-0209-9.

[14] Y. Vázquez-Baeza, M. Pirrung, A. Gonzalez, and R. Knight, "EMPeror: a tool for visualizing high-throughput microbial community data," *GigaScience*, vol. 2, no. 1, Nov. 2013, doi.org/10.1186/2047-217x-2-16.

[15] K. Mitchell et al., "PUMAA: A Platform for Accessible Microbiome Analysis in the Undergraduate Classroom," *Frontiers in Microbiology*, vol. 11, Oct. 2020, doi.org/10.3389/fmicb.2020.584699.

[16] A. Kassambra. "Kruskal-Wallis Test in R." sthda.com. Accessed 28 Dec. 2023. [Online.] Available: http://www.sthda.com/english/wiki/kruskal-wallis-test-in-r

[17] A. A. Pourbabaee, S. Soleymani, M. Farahbakhsh, and E. Torabi, "Biodegradation of diazinon by the Stenotrophomonas maltophilia PS: pesticide dissipation kinetics and breakdown characterization using FTIR," *International journal of environmental science and technology*, vol. 15, no. 5, pp. 1073–1084, Jul. 2017, doi.org/10.1007/s13762-017-1452-6.

[18] C. Lozupone et al., "UniFrac: an effective distance metric for microbial community comparison," *The ISME Journal*, vol. 5, no. 2, pp. 169–172, Sep. 2010, doi.org/10.1038/ismej.2010.133.

[19] Q. M. S. Jamal and V. Ahmad, "Lysinibacilli: A Biological Factories Intended for Bio-Insecticidal, Bio-Control, and Bioremediation Activities," *Journal of Fungi*, vol. 8, no. 12, p. 1288, Dec. 2022, doi.org/10.3390/jof8121288.

[20] M. A. Hamada and S. Soliman, "Characterization and genomics identification of key genes involved in denitrification-DNRA-nitrification pathway of plant growth-promoting rhizobacteria (Serratia marcescens OK482790)," *BMC Microbiology*, vol. 23, no. 1, Aug. 2023, doi.org/10.1186/s12866-023-02941-7.

[21] S. St. Clair *et al*., "Analysis of the Soil Microbiome of a Los Angeles Urban Farm," *Applied and Environmental Soil Science*, vol. 2020, pp. 1–16, Feb. 2020, doi.org/10.1155/2020/5738237.

[22] E. Bueno *et al*., "Anaerobic nitrate reduction divergently governs population expansion of the enteropathogen Vibrio cholerae," *Nature Microbiology*, vol. 3, no. 12, pp. 1346–1353, Oct. 2018, doi.org/10.1038/s41564-018-0253-0.

[23] G. T. Macfarlane and R. A. Herbert, "Nitrate Dissimilation by Vibrio spp. Isolated From Estuarine Sediments," *Microbiology*, vol. 128, no. 10, pp. 2463–2468, Oct. 1982, doi.org/10.1099/00221287-128-10-2463.

[24] S. R. Sias and J. L. Ingraham, "Isolation and analysis of mutants of Pseudomonas aeruginosa unable to assimilate nitrate," *Archives of Microbiology*, vol. 122, no. 3, pp. 263–270, Sep. 1979, doi.org/10.1007/BF00411289.

[25] M. Zhang, A. Li, Q. Yao, B. Xiao, and H. Zhu, "Pseudomonas oligotrophica sp. nov., a Novel Denitrifying Bacterium Possessing Nitrogen Removal Capability Under Low Carbon–Nitrogen Ratio Condition," *Frontiers in Microbiology*, vol. 13, May 2022, doi.org/10.3389/fmicb.2022.882890.

[26] L. C. Reimer *et al*., "BacDive in 2022: the knowledge base for standardized bacterial and archaeal data," *Nucleic Acids Research*, vol. 50, no. D1, pp. D741–D746, Oct. 2021, doi.org/10.1093/nar/gkab961.

[27] B. Sun *et al*., "Application of biofertilizer containing Bacillus subtilis reduced the nitrogen loss in agricultural soil," *Soil Biology and Biochemistry*, vol. 148, p. 107911, Sep. 2020, doi.org/10.1016/j.soilbio.2020.107911.

[28] M. E. Winston *et al*., "Understanding Cultivar-Specificity and Soil Determinants of the Cannabis Microbiome," *PLoS ONE*, vol. 9, no. 6, p. e99641, Jun. 2014, doi.org/10.1371/journal.pone.0099641.

[29] C. G. Olds *et al*., "Applying a modified metabarcoding approach for the sequencing of macrofungal specimens from fungarium collections," *Applications in Plant Sciences*, vol. 11, no. 1, Jan. 2023, doi.org/10.1002/aps3.11508.

The Annotation of the Complete Genome of the Mycobacterium phage Inverness

NICHOLAS SHELTRA[1], RAND DEARDORFF[1], SHARELLE BAILEY[1], EMMA DASILVA-MARTINEZ[1], SHANNON DICKEY[1], BETTINA EVANS[1], ADDIE MEHL[1], SHARON GUSKY[1]*

[1]*STEM Department Connecticut State Community College: Northwestern Campus, Winsted, CT*

**sharon.gusky@ctstate.edu*

Abstract: The gene annotation of the Mycobacterium phage Inverness was performed to establish certain genetic characteristics and qualities of the phage. Our research was designed to investigate the potential usefulness of this phage for medical purposes as part of the SEA-PHAGES project. Due to the decline in effectiveness of antibiotics when treating bacterial diseases, demand for alternative therapies and treatment has grown substantially. Phages have the potential to meet this demand. The genome of Inverness was found to be 68,264 base pairs in length, to possess a GC content of 66.5%, and to contain 99 protein-coding genes. Based on nucleotide similarities, the Mycobacterium phage Inverness was placed into cluster B and subcluster B1. The 33 genes with identifiable functions had an almost even split between rightward (49.49%) and leftward (50.51%) oriented genes. No putative function could be identified for 66 genes. No tRNAs were found to be present within the genome for this specific phage. It should also be noted that, among those genes with identified functions, two codes for lysins, which are proteins that kill bacteria cells. This suggests potential future opportunities to use this phage as a treatment for certain cases of antibiotic resistant infections.

Keywords: bacteriophage, gene annotation, bioinformatics

Introduction

The Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES) Program is a course-embedded research program where students search for new bacteriophages in soil samples, perform a variety of laboratory techniques, and complete a complex genome annotation. The goals of the project are twofold: first, to increase undergraduate student interest and retention in the biological sciences, and second, to identify phages that can be used to treat antibiotic resistant infections [1]. The SEA-PHAGES project is jointly run by the Howard Hughes Medical Institute's Science Education division and Graham Hatfull's laboratory at the University of Pittsburgh. Participation in the SEA-PHAGES project has been shown to increase student retention and to influence career choices [1]. Students at Connecticut State Community College, Northwestern campus, participated in the SEA-PHAGES program while taking a Molecular Genetics course. As part of this course, students performed the structural and functional annotation of the Mycobacterium phage Inverness to determine its potential to treat antibiotic resistant infections.

Antibiotic resistant infections occur when mutation of the infecting bacterium prevents its destruction by antibiotics. Infections with antibiotic resistant organisms can be difficult or impossible to treat. According to the World Health Organization, 1.27 million people died in 2019 due to infections with antibiotic resistant organisms [2]. Bacteriophages have proven successful as a last resort in treating antibiotic resistant infections and are being tested in clinical trials [3].

Phages, also known as bacteriophages, are viruses that infect and replicate within bacterial cells. They are the most prevalent biological agent on Earth and are found everywhere in the environment [4]. The size, shape, and genetic structure of phages exhibit remarkable diversity [4]. All phages are made up of a nucleic acid genome covered in a capsid protein shell which protects the genetic information and facilitates its transfer to host cells. Many phages have tails that are used to deliver the genome into the host cell. The ability of phages to lyse bacterial cells makes them potentially effective in treating patients with antibiotic resistant infections [5]. To understand the potential for a phage to be used to treat patients, the genome of the phage must be

annotated. Gene annotation involves the comprehensive process of detecting and characterizing genes within a genome. It starts by using computational gene prediction tools to identify potential genes based on genomic features. Structural annotation involves identifying the boundaries of the genes and functional annotation involves assigning putative functions to gene products. Gene annotation is crucial for understanding a phage's genetic traits, which must be considered when evaluating a phage for the possible treatment of antibiotic resistant infections. For a phage to be used as treatment "the phage genomes should not include any genes known or suspected to be toxic" [3].

Methods

Obtaining the Sequence of the Mycobacterium phage Inverness

The work to identify, isolate, purify, extract, and sequence DNA was not performed as part of this research project. However, a brief description of those processes is described here.

The Mycobacterium phage Inverness was collected from a bag of Miracle-Gro® potting soil obtained in Fort Collins, Colorado, USA. It was isolated, purified, and amplified by Sean Anderson of Rocky Mountain High School in Colorado, as part of the Phage Hunters Integrating Research and Education (PHIRE) program [6], using *Mycobacterium smegmatis mc*^{2} *155* as a host [7].

DNA extracted from the phage was sent to the Pittsburgh Bacteriophage Institute for sequencing. The Pittsburgh Bacteriophage Institute completed sequencing on December 20th, 2020, using Illumina Sequencing, with an approximate shotgun coverage of 511. The shotgun method of sequencing involves breaking the genome up into pieces, sequencing each piece and then reconstructing the entire genome. The shotgun coverage number, in this case, 511, describes the average number of reads that align to the reference database.

The sequence information was used to create the FASTA file. The FASTA file containing the text-based sequence of the nucleotides for the phage genome was uploaded into the PhagesDB database by the SEA-PHAGES project administrators [8].

Annotation Process

The FASTA file was obtained from the PhagesDB database and loaded into DNA Master v5.0.2, a gene exploration and annotation tool used to predict the probable genes in the sequence [9]. The Mycobacterium phage Inverness genome sequence was also run through the evidential programs contained within The Phage Evidence Collection And Annotation Network (PECAAN) version 20221109 [10]. PECAAN was utilized to compile data from several other databases for comparison [10].

Within PECAAN, gene start and stop recommendations came from the Gene Locator and Interpolated Markov ModelER (Glimmer) system v3.02b, along with the GeneMark v4.28, and Starterator v1.2 systems [11,12]. GeneMark was also utilized for determining coding capacity [11].

This information along with the Z-score, gap or overlap between genes, the final score, and coding capacity were used to select the best starting position for each gene.

The Z-score provides the standard deviation of a score when compared to the best scores from all possible start positions in the genome. The Z-score provides the standard deviation of a score when compared to the best scores from all possible start positions in the genome. DNA Master and PECAAN produce Z-scores for the various possible starting positions of each predicted gene. While the exact values of the Z-scores vary for each gene, the best Z-score is one that is closest to 2. The starting position selected for each gene is based on selecting a position that has a calculated Z-score that is closest to 2.

In addition, the Genemark map was used as an effective visual reference for the regional coding potential of prospective gene candidates, and Actinobacteriophage Phamerator, version 567, was used to compare related phages in subcluster B1 [13].

Gene functions were determined by comparing the protein sequences for each gene to previously annotated genomes. PECAAN provided protein analysis recommendations from BLASTp v2.13.0., the Protein database and Non-Redundant Protein Sequences database, and HHPred v2.08 as well as NCBI_Conserved_Domains (CD) databases [14,15]. These databases are used by the PECAAN algorithms to detect similarities between proteins. The evidence selected was based on evaluating probability ratings and e-values, which measure the significance between sequences.

The Transmembrane Helices: Hidden Markov Model, TMHMM, provided evidence on transmembrane protein predictions [16]. Transmembrane proteins play a role in controlling the lysis of bacteria after infection by a bacteriophage. Phages with transmembrane proteins have the potential to kill bacteria and, therefore, to be used to treat antibiotic resistant infections.

TRNAscan and Aragorn programs were used to look for the presence of tRNAs [17,18]. The full annotation from PECAAN was run through DNA Master to create the minimal file suitable for submission to the PhagesDB website following SEA-PHAGES protocols.

A quality control check was performed before the complete genome was submitted to GenBank by the SEA-PHAGES' administrators who check to make sure that the guiding principles for gene annotation were followed and that all the functions call are allowable functions [19].

Results and Discussion

The Mycobacterium phage Inverness has a 68,264 base pair long genome with a GC content of 66.5% with a circularly permuted genome end character. Genomes with circularly permuted genomes form circular molecules upon injection into the host. These circular molecules can be used for replication of the phage. The genome annotation identified 99 protein-coding genes, with a nearly equal distribution between the rightwards (49.49%) and leftwards (50.51%) genes.

Thirty-three genes were assigned putative functions, but no functions could be identified for the remaining 66 genes.

Proteins involved in capsid structures are clustered between genes #9 and #13. Proteins involved in the tail structure are located between genes #18 and #41. The tape measure protein was identified as being coded for by gene #28. This gene determines tail length. Long-tailed phages have a tape-measure protein gene consisting of 2,000 or more base pairs.

The tape measure protein gene in the Mycobacterium phage Inverness is 5,976 base pairs long and codes for 1,991 amino acids, indicating that it has a long tail. Figure 1 shows the size of the tape measure protein gene when compared to the genes for the tail assembly chaperone and the minor tail protein.

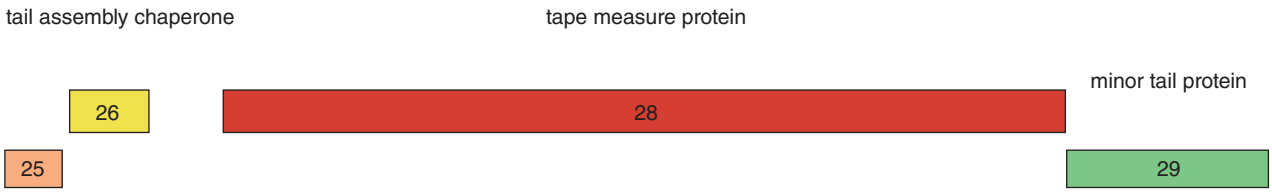

Fig. 1: This figure, taken from the Phamerator Database [13], shows the relative size of the tape measure protein gene (#28) when compared to the tail assembly chaperone genes #25 and #26 and the minor tail protein #29. The reverse gene #27 is not shown.

Genes assigned putative functions include those coding for structural proteins like RuvC-resolvase and for enzymes involved in DNA replication and packaging like DNA helicase, DNA primase, and HNH endonuclease.

DNA Helicase is the enzyme responsible for the break of the hydrogen bonds between DNA strands during the process of DNA replication. DNA primase is also used during DNA replication and is responsible for adding primers or starting sequences to the strands. This allows DNA polymerase to add new nucleotides to the growing nucleotide strands. HNH endonuclease is a small DNA-binding protein that can cleave the covalent bonds between nucleotides in a strand.

Genes that code for Lysin A and Lysin B were identified as #48 and #49, respectively. Lysins are enzymes that destroy bacteria cell walls. This leads to the death of the bacteria cells and indicates that this phage has the potential to be used to treat antibiotic resistant infections.

No tRNAs were found, indicating that the phage uses the host tRNAs in the translation process.

The complete genome annotation can be found in the National Library of Medicine's GenBank Database under accession number OR159656 [20].

The full list of genes with identified functions can be found in Table 1.

Table 1 Genes with Identified Function

Gene Number	Length in basepairs	# of Amino Acids	Product
1	567	188	Adenylate Kinase
2	1788	595	Terminase
6	555	184	RUVC-like Resolvase
8	1905	634	Portal Protein
9	2574	857	Major Capsid and Fusion Protein
10	435	144	HNH Endonuclease
12	1746	581	Major capsid Hexamer Protein
13	804	267	Major Capsid Pentamer Protein
15	309	129	Holin
18	801	266	Major Tail Protein
20	777	258	Queuine tRNA Ribosyltranferase
22	738	245	Head-to-Tail Adapter
25	423	140	Tail Assembly Chaperone
26	564	187	Tail Assembly Chaperone
28	5976	1991	Tape Measure Protein
29	1434	477	Minor Tail Protein
30	1113	370	Minor Tail Protein
31	2256	751	Minor Tail Protein
32	1347	448	Minor Tail Protein
33	1161	386	Minor Tail Protein
41	408	135	Tail Fiber
45	216	71	Helix-turn-Helix Binding Domain Protein
46	525	174	Helix-turn-Helix Binding Domain Protein
48	1329	442	Lysin A
49	1356	451	Lysin B
51	1404	476	Exonuclease
52	1704	567	DNA Helicase
57	2748	915	DNA Primase
59	1860	619	DNA Replicase
66	180	59	Ribbon Helix-turn-Helix Binding Domain Protein
68	699	232	DNA Binding Protein
82	303	100	HNH Endonuclease

Conclusion

The Mycobacterium phage Inverness was assigned to cluster B and subcluster B1 based on the nucleotide similarities to other phages in the clusters. Like other subcluster B1 phages, Mycobacterium phage Inverness infects mycobacterium and has a GC content of 66.5% and a base pair length of 68,264 [21]. Based on its gene content, it is predicted to be of the siphovirus morphotype. Phages with this morphotype have long, flexible tails that are non-contractible and have heads that are hexagonal and icosahedral.

The capsid size, head, and tail length of Mycobacterium Phage Inverness are unknown since electron microscopy was not performed. However, it was determined that gene #12 codes for a hexamer major capsid protein, confirming that the head is hexagonal. The tape measure gene (#28) was found to be 5,976 base pairs long, confirming that the tail is long.

The Mycobacterium Phage Inverness contains two genes that code for lysins. Lysin A is coded for in gene #48, and Lysin B is coded for in gene #49. Lysins disrupt the complex structures of bacteria cells, leading to rapid cell lysis [22]. This action kills the bacteria cells and offers a promising solution to combat infections caused by antibiotic-resistant strains of Mycobacterium.

Acknowledgements. We acknowledge support from the Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES), Howard Hughes Medical Institute, Chevy Chase, MD, including support from S.M.R. Gurney, Claire Reinhart, D.A. Russell, Deborah Jacobs-Sera, Vic Sivanathan, and Grahm Hatfull. The National Science Foundation (NSF) Advanced Technology Education Grant under award #1081062 provided research support for the authors.

Disclosures. The authors declare no conflicts of interest.

References

[1] D. I. Hanauer, et al., "An inclusive Research Education Community (iREC): Impact of the SEA-PHAGES program on research outcomes and student learning," *Proc Natl Acad Sci*, vol. 114, no. 51, pp. 13531-13536, Dec. 2017, doi: 10.1073/pnas.1718188115

[2] World Health Organization, "Antibiotic resistance," WHO Fact Sheets. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance

[3] G. F. Hatfull, R. M. Dedrick, and R.T. Schooley, "Phage Therapy for Antibiotic-Resistant Bacterial Infections," *Annu Rev Med*, vol. 73, pp. 197–211, Jan. 2022, doi: 10.1146/annurev-med-080219-122208.

[4] L. M. Kasman, and L.D. Porter, "Bacteriophages," in StatPearls [Internet], StatPearls Publishing, 2022. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK493185/

[5] G. F. Hatfull, "Mycobacteriophages: From Petri Dish to patient," *PLoS Pathog*. Jul. 2022, doi: 10.1371/journal.ppat.1010602.

[6] Hatfull Lab, "Phage Hunters Integrating Research and Education (PHIRE)." [Online]. Available: https://hatfull.org/courses

[7] M. Poxleitner, W. Pope, D. Jacobs-Sera, V. Sivanathan, and G. Hatfull, "Phage discovery guide," Howard Hughes Medical Institute. [Online]. Available: https://seaphagesphagediscoveryguide.helpdocsonline.com/home

[8] D. A. Russell, and G. F. Hatfull, "PhagesDB: The actinobacteriophage database," *Bioinformatics*, vol. 33, no. 5, pp. 784–786, Mar. 2017, doi: 10.1093/bioinformatics/btw711.

[9] W. H. Pope, D. Jacobs-Sera, "Annotation of bacteriophage genome sequences using DNA Master: an overview," *Methods Mol Biol*, vol. 1681, pp. 217–229, 2018, doi: 10.1007/978-1-4939-7343-9_16.

[10] C. A. Rinehart, B. L. Gaffney, J. R. Smith, and J. D. Wood, "PECAAN: Phage Evidence Collection and Annotation Network," Western Kentucky University Bioinformatics and Information Science Center. [Online]. Available: https://discover.kbrinsgd.org/login

[11] A. V. Lukashin, and M. Borodovsky, "GeneMark.hmm: new solutions for gene finding," *Nucleic Acids Res*, vol. 26, no. 4, pp. 1107–1115, Feb. 1998, doi: 10.1093/nar/26.4.1107.

[12] M. Pacey, "Starterator guide," University of Pittsburgh. [Online], Available: https://seaphages.org/media/docs/Starterator_Guide_2016.pdf

[13] S. G. Cresawn, M. Bogel, N. Day, D. Jacobs-Sera, R.W. Hendrix, and G. F. Hatfull, "Phamerator: a bioinformatic tool for comparative bacteriophage genomics," *BMC Bioinformatics*, vol. 12, no. 1, p. 395, Oct. 2011, doi: 10.1186/1471-2105-12-395.

[14] A. Marchler-Bauer et al.,"CDD: NCBI's conserved domain database," *Nucleic Acids Res* vol. 43, no. D1, pp. D222–D226, Jan. 2015, doi: 10.1093/nar/gku1221.

[15] J. Söding, A. Biegert, and A. N. Lupas, 2005, "The HHpred interactive server for protein homology detection and structure prediction," *Nucleic Acids Res*, vol. 33, no. suppl_2, pp. W244–W248, Jul. 2005, doi: 10.1093/nar/gki408.

[16] N. Chaturvedi, S. Shanker, V. K. Singh, D. Sinha, P.N. Pandey, "Hidden markov model for the prediction of transmembrane proteins using MATLAB," *Bioinformation*, vol. 7, no. 8, pp. 418–421, 2011, doi: 10.6026/97320630007418.

[17] D. Laslett, and B. Canback, "ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences," *Nucleic Acids Res*, vol. 32, no. 1, pp. 11–16, 2004, doi: 10.1093/nar/gkh152.

[18] T. M. Lowe, and S.R. Eddy, "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence," *Nucleic Acids Res*, vol. 25, no. 5, pp. 955–964, Mar. 1997, doi: 10.1093/nar/25.5.955.

[19] M. Poxleitner, W. Pope, D. Jacobs-Sera, V. Sivanathan, and G. Hatfull, "Guiding Principles of Bacteriophage Gene Annotation," SEAPHAGES Bioinformatics Guide, Howard Hughes Medical Institute. [Online]. Available: https://seaphagesbioinformatics.helpdocsonline.com/home

[20] S. A. Bailey, et al., "Mycobacterium phage Inverness: The Complete Genome," National Library of Medicine: GenBank. [Online]. Available: https://www.ncbi.nlm.nih.gov/nuccore/OR159656

[21] G. F. Hatfull, "The secret lives of mycobacteriophages," *Adv Virus Res*, vol. 82, pp. 179–288, 2012, doi: 10.1016/B978-0-12-394621-8.00015-7.

[22] C. Ghose, C.W. Euler, "Gram-Negative Bacterial Lysins," *Antibiotics*, vol. 9, no. 2, Art. no. 2, Feb. 2020, doi: 10.3390/antibiotics9020074.