Measuring the DNA Content of a Soil Sample from a Major American Research University using Nanopore and Illumina Sequencing

Zan Chaudhry Partners: Andres Parra & Anjan Singh
Methods in Nucleic Acid Sequencing Lab, Spring 2023 10:30 AM

Introduction

Agriculture faces a growing number of challenges today. Output must be increased to match the surge in global population, even as land availability decreases due to residential and commercial development. With increased globalization, pests and diseases flow more freely than ever, and the liberal use of various pesticides, herbicides, and antibiotics is driving the resistance of these organisms. Clearly, protecting crops and improving land usage are important topics of research to maximize yields. Particularly, investigating the properties of soil is essential for optimizing growth conditions, and recent developments in the biological sciences provide tools for more accurate and actionable soil analysis than ever before. The drastic decrease in nucleic acid sequencing costs make this tool accessible for agricultural use. This presents the potential for optimizing crop selection based on the soil microbiome, preventing losses due to disease by screening for pathogens, and protecting crops with early detection of pests.[1]

Today, several methods exist for sequencing. Two particularly widely adopted varieties are Nanopore and Illumina. Nanopore sequencing functions by measuring the ionic current within a protein pore complex embedded in a membrane. The DNA sequence of interest is fed through the pore by a motor protein, and each nucleotide disrupts the current, producing a characteristic signal as it passes through. The full sequence is recovered by assembling the individual current signals. Due to the simplicity of this method, Nanopore sequencing gives rapid results and can be packaged into inexpensive, portable devices. However, the method is relatively error prone, with recent devices achieving roughly 98% accuracy.[2, 3]

Illumina sequencing functions by synthesizing complementary strands to fragments of the genome using fluorescently labeled chain-terminating (dideoxy—) nucleotides. The DNA sample is first broken into small fragments a few hundred base pairs in length which are processed for analysis. The fragment solution is then poured over a flow cell coated with primer sequences, to which complementary fragments adhere. Nucleotides and polymerases are added to create local clusters of fragment replicates. Then, fluorescent chain-terminating nucleotides and polymerases are added, fluorescence is measured at each fragment to determine the nucleotide identity, and the nucleotide is removed while a chain-terminating nucleotide is added at the next position simultaneously. This process is repeated until the fragments have been completely sequenced. Illumina sequencing provides very high accuracy (as high as 99.9%), though these systems are large, expensive, and have potentially low yields/copies during the replication step.[4]

The investigation of these two sequencing methods for soil analysis is an important step in adapting this technology for agriculture. The following report utilizes both methods to sequence DNA found in a soil sample collected from the biomedical engineering department of a major American research university. Organisms in the soil sample are identified using the Kraken 2 database. By comparing the results of these methods, the researchers hope to shed light on sequencing as a next-generation tool for optimizing agriculture.

Methods

DNA Extraction

A 15 mL soil sample was collected from outside of the biomedical engineering department of a major American research university. The Qiagen PowerSoil Pro Kit was then used for extracting DNA from the soil sample. The soil sample was vortexed and centrifuged in the kit's PowerBead Pro Tube along with CD1 solution (a lysing agent). The supernatant (containing dissolved compounds and excluding the lysed cells which are in the pellet) was collected and CD2 solution (precipitates non-DNA compounds out of solution) was added, followed by another round of centrifugation. The supernatant was collected once more (which now contains a higher purity DNA solution after CD2 removed extraneous compounds). CD3 solution (a high salt solution) was then added and vortexed with the sample. This step prepares the DNA for binding to the silica membrane used for collection in the next step, since high salt concentrations promote DNA binding to silica. The DNA-CD3 solution was then added to the MB Spin Column, which contains a silica membrane to which the DNA adheres. The tube containing spin column was centrifuged to allow any non-DNA to flow through the membrane, and the flow-through was discarded. Solution EA (a wash buffer) was added to remove any additional non-DNA contaminants, and the column was centrifuged again with flow-through discarded. Next, CD5 (another wash agent) was added to remove additional contaminants, particularly ionic compounds such as salt attached to the membrane. After another centrifugation and flow-through discard, the final CD6 solution (DNA release solution) was added to the membrane to elute the purified DNA from the column. DNA purity and concentration were assessed using a Nanodrop spectrophotometer. The isolated DNA sample was stored for further use.[5]

Illumina Sequencing Prep

A solution containing 500 ng of DNA with a volume of 30 μL was produced from the purified sample and nuclease-free water. This sample was then tagmented with bead-linked transposomes. In essence, the DNA present in the sample is bound to beads that attach adapter proteins while fragmenting the proteins into manageable lengths for Illumina sequencing. The adapter proteins are essential for future binding of the sequences to the flow cell used for sequencing. After the tagmentation reaction, all non-bead-bound compounds were removed by placing the tubes with beads suspended on a magnet (the beads are magnetic). As the beads adhere to the bottom of the tube, the solution clarifies and can be discarded. This process was repeated multiple times. Finally, the purified tagmented DNA was amplified using PCR and index adapters were added. Index adapters are complementary to the previously added adapters and contain primers on the ends. Thus, these adapters are incorporated into the replicated fragments, generating strands with primer ends. Primers are essential because they allow the fragments to bind to the Illumina flow cell during sequencing. After this amplification step, cleanup was performed to remove any remaining extraneous compounds in solution (such as unbound index adapters/polymerases/nucleotides). This involved a similar bead/magnet/wash cycle that was repeated multiple times before finally removing the purified, fragmented DNA (with primer ends) from the beads and storing at 4 °C until sequencing.[5]

Nanopore Sequencing

A solution containing 400 ng of DNA with a volume of 7.5 μL was produced from the purified sample and nuclease-free water. Fragmentation Mix (RB09) was then added to this adjusted sample to tagment the DNA. This step both breaks the DNA into manageable sequencing fragments and attaches barcode adapters, using a barcoded transposome complex. The fragment-containing solution was then purified using a similar bead/magnet/wash cycle as in the Illumina preparation (though this time using a 70% EtOH solution). Finally, the DNA was removed from the beads by resuspending the DNA-bound beads in 10 μL of a 10 mM Tris-HCl and 50 mM NaCl solution and pelleting the beads once more on a magnet. The eluate (containing purified, tagmented DNA) was collected and stored on ice until sequencing. The Nanopore flow cell was prepared by priming with a specialized Priming Mix. The sample was then added, and sequencing data was collected.[5]

Illumina Sequencing

The sample of Illumina prepped DNA was diluted to measure the correct input parameters after measuring concentration/purity with a Nanodrop. The sample was then denatured in a NaOH solution to produce single strands of DNA. The denatured sample was loaded onto a flow cell, which was then inserted into the Illumina sequencer, along with appropriate reagents (PR2). The sequencing run was then performed. The operating principle of Illumina sequencing is described in **Introduction**.[5]

Sequencing Analysis

Sequencing results were analyzed using FastQC in Python to determine sequencing quality. Taxonomic classification was performed using Kraken 2, with Pavian used for result visualization. [6,7]

Results

Illumina sequencing produced high-quality reads, with a per-base quality score of over 30 throughout the sample. However, the distribution of per sequence GC content deviated slightly from the theoretical distribution, possessing a similar mean and normal distribution shape but substantially smaller standard deviation. Nanopore sequencing produced markedly lower quality reads, with an average per-base quality score around 20 (from a minimum of 5 to a maximum of 48). The distribution of per sequence GC content deviated considerably from the theoretical distribution, with a similar mean for its main peak (61%), but also possessing a secondary peak (43%). Below, the data is presented for species content of the sample (Fig. 1 & 2 and Tables 1-4).

In the Illumina data, 1,981,489 reads were collected, with 71,617 (3.614%) classified and 1,909,872 (96.39%) unclassified. In the Nanopore data, 376,000 reads were collected, with 122,094 (32.47%) classified and 253,906 (67.53%) unclassified. The major bacterial phyla represented in both sets of sequencing data include: actinobacteria, bacteroidota, cyanobacteria, firmicutes, planctomycetota, proteobacteria, verrucomicrobia. One eukaryotic species was identified by both sequencing methods: *Homo sapiens* (Tables 3 & 4).

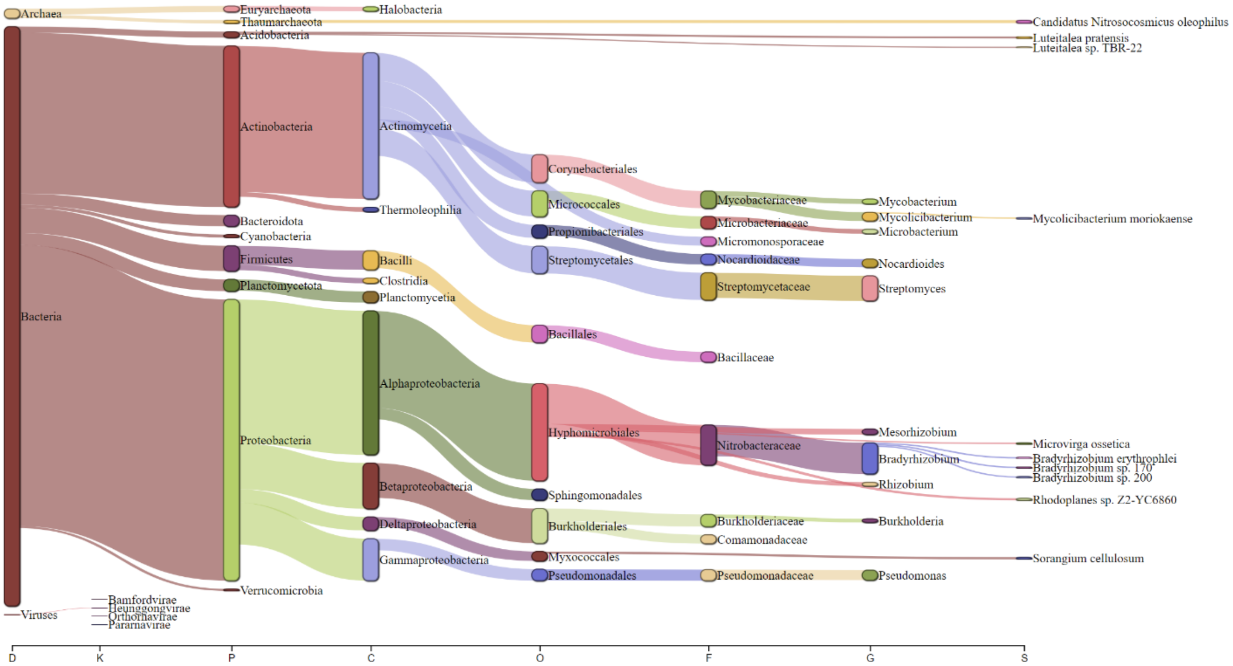

Fig 1. Sankey visualization of Illumina sequencing results. Displays the microbial content of the soil sample with taxonomic classifications. The top 10 identified taxonomic classes for each node are included. Notice the dominance of bacteria in the sample, especially proteobacteria.

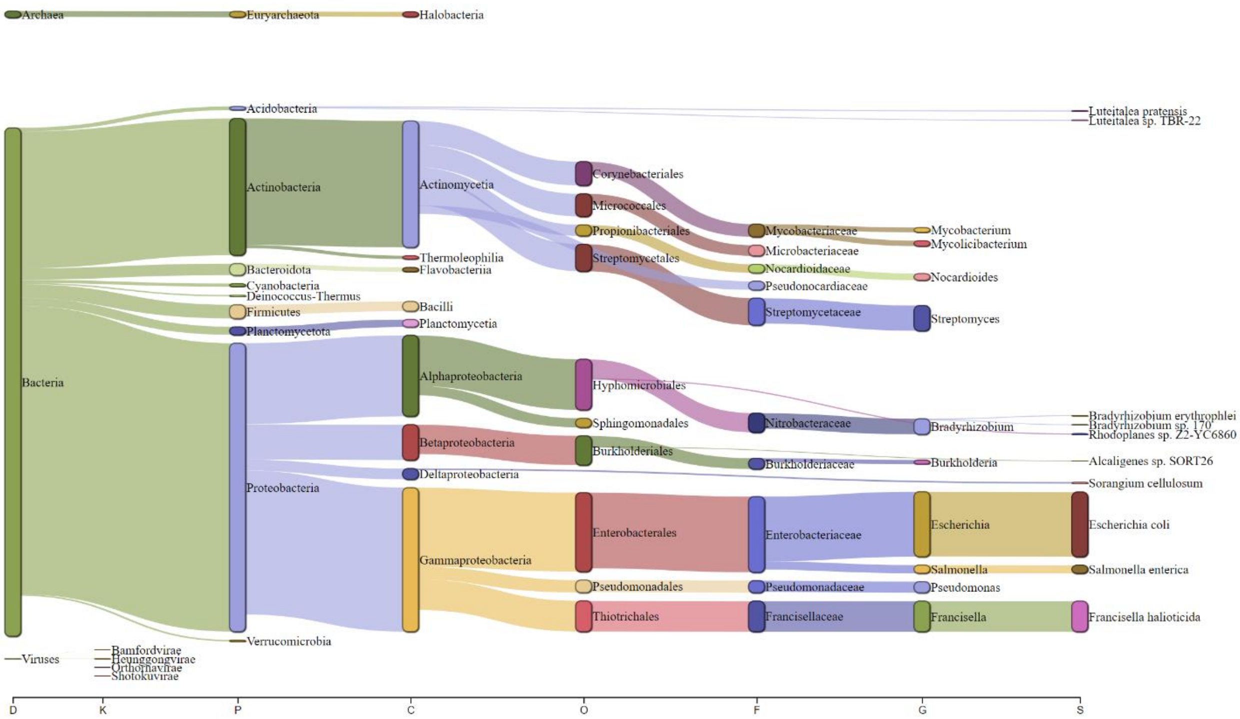

Fig 2. Sankey visualization of Nanopore sequencing results. Displays the microbial content of the soil sample with taxonomic classifications. The top 10 identified taxonomic classes for each node are included. Notice the dominance of bacteria in the sample, especially proteobacteria.

Table 1. Top 15 bacterial species identified by Illumina sequencing. Species are ranked according to read count.

Name	Rank	TID	Read Count	Percentage of Classified Reads
Rhodoplanes sp. Z2-YC6860	1	674703	264	0.3686275605
Sorangium cellulosum	2	56	258	0.3602496614
Luteitalea pratensis	3	1855912	228	0.3183601659
Bradyrhizobium sp. 170	4	2782641	163	0.2275995923
Bradyrhizobium erythrophlei	5	1437360	163	0.2275995923
Microvirga ossetica	6	1882682	159	0.2220143262
Mycolicibacterium moriokaense	7	39691	158	0.2206180097
Bradyrhizobium sp. 200	8	2782665	151	0.2108437941
Luteitalea sp. TBR-22	9	2802971	139	0.1940879959
Parageobacillus caldoxylosilyticus	10	81408	126	0.1759358811
Rhodopseudomonas palustris	11	1076	125	0.1745395646

Pseudorhodoplanes sinuspersici	12	1235591	125	0.1745395646
Usitatibacter palustris	13	2732487	125	0.1745395646
Mycolicibacterium thermoresistibile	14	1797	116	0.161972716
Pseudolabrys taiwanensis	15	331696	101	0.1410279682

Table 2. Top 15 bacterial species identified by Nanopore sequencing. Species are ranked according to read count. Notice the gammaproteobacteria (top three species), which are absent from the Illumina data.

Name	Rank	TID	Read Count	Percentage of Classified Reads
Escherichia coli	1	562	14849	12.1619408
Francisella halioticida	2	549298	6985	5.721001851
Salmonella enterica	3	28901	1861	1.524235425
Sorangium cellulosum	4	56	349	0.2858453323
Rhodoplanes sp. Z2-YC6860	5	674703	246	0.2014841024
Luteitalea pratensis	6	1855912	233	0.1908365685
Luteitalea sp. TBR-22	7	2802971	195	0.159713008
Bradyrhizobium erythrophlei	8	1437360	185	0.1515225973
Alcaligenes sp. SORT26	9	2813780	160	0.1310465707
Bradyrhizobium sp. 170	10	2782641	159	0.1302275296
Rhodopseudomonas palustris	11	1076	145	0.1187609547
Conexibacter woesei	12	191495	143	0.1171228725
Capillimicrobium parvum	13	2884022	139	0.1138467083
Microvirga ossetica	14	1882682	136	0.1113895851
Bradyrhizobium sp. 200	15	2782665	134	0.1097515029

Table 3. Top eukaryotic species identified by Illumina sequencing. Notice that there is only one species. This was the only eukaryotic species classified in the sample.

Name	Rank	TID	Read Count	Percentage of Classified Reads
Homo sapiens	1	9606	545	0.7609925

Table 4. Top eukaryotic species identified by Nanopore sequencing. Notice that there is only one species. This was the only eukaryotic species classified in the sample.

Name	Rank	TID	Read Count	Percentage of Classified Reads
Homo sapiens	1	9606	2961	2.4251806

Discussion

The percentage of classified reads can be found in the second paragraph of **Results**. Kraken 2 was used to assign classifications to reads in the data, and the default database was implemented. Two criteria must be satisfied for a read to be classified: first, the read must contain a sequence that is unique to some taxonomy; second, the sequence/taxonomy must be present in the database. The unclassified reads thus failed to satisfy both criteria. Most of the unclassified reads likely arise from the limited size of the database. To illustrate this point, the database only contains one eukaryotic species (as mentioned in **Results**), severely limiting the scope of classification.

As seen in **Results** (Tables 3 & 4), the Illumina and Nanopore data presented different lists for the most abundant bacterial species. Thus, the top three species (for further discussion) were chosen by selecting the three species that appeared first in both lists. These were: *Sorangium cellulosum, Rhodoplanes sp. Z2-YC6860*, and *Luteitalea pratensis*. *Sorangium cellulosum* is a soil-dwelling bacterium that anaerobically metabolizes cellulose and is known for its production of a vast range of metabolites aimed at reducing soil competition, including antibacterial, antifungal, and even anti-mammalian compounds. Bacteria of the genus *Rhodoplanes* are phototrophic (generating nutrition/metabolic compounds by capturing energy from light) soil-dwellers. Bacteria of the genus *Luteitalea* are soil-dwelling chemoorganotrophs with a very widespread distribution. All of these bacteria are expected in the sample, considering their soil-dwelling natures. [8,9,10]

Both sequencing methods had different bacteria identified as least abundant, with a minimum of 10 reads. Thus, a representative bacterium was chosen by sorting through all bacteria with a minimum of 10 reads from either method. Several species with only 10 reads were found. The first species for which information on the contig number could be found was selected. This was *Bradyrhizobium symbiodeficiens*. The genus *Bradyrhizobium* consists of nitrogen-fixing bacteria often associated with legumes, however interestingly, *symbiodeficiens* lacks these key nitrogen-fixation genes and is non-symbiotic, though still associated with legumes. The genome contains 1 contig and is unsurprising in this sample, since these bacteria are also soil-dwelling.[11]

One eukaryotic sample was found in the analysis: *Homo sapiens*. *Homo sapiens* is a land-dwelling mammal known for its characteristically large brain. *Homo sapiens* is also known for its extensive ability to design and construct complex tools, including the recent development of Nanopore and Illumina sequencing. This species is particularly abundant at the location from which the soil was extracted (Clark Hall; Baltimore, MD). The most likely cause for the high read counts (545 in Illumina and 2,961 in Nanopore) of *Homo sapiens* was manipulation of the soil sample by the experimenters (who are individuals of this species), which lead to incorporation of some DNA (perhaps from stray skin cells) in the sample.

In general, the Nanopore and Illumina sequencing data present similar results (see Figs. 1 & 2). However, a few key differences are present. First, Illumina performed much better on sequencing quality metrics. In particular, the per base quality of Illumina remained above 30 throughout the sample, while Nanopore per base quality began below 10 and remained roughly between 18 and 20 for most of the sample. Second, Illumina contained a much higher total read count (nearly 2 million), though with a small proportion classified (3.614%), while Nanopore contained a much lower read count (nearly 400,000), though with a much higher proportion classified (32.47%). Finally, the top three bacterial species identified in Nanopore were found in the Illumina data at negligible levels (10 reads each). Addressing the first point, Illumina is much less error prone due to its use of localized replicates, leading to higher quality scores. Second, Illumina fragments the DNA into much smaller sequences (a couple hundred base pairs) than Nanopore (known for its long-sequence potential), leading to a larger number of reads. This also contributes to the lower proportion of classified reads, since the total read count is much higher. Finally, the missing species could be accounted for due to inability to match particular reads to these bacteria in the Illumina data. Since Illumina fragments the given DNA, there is a chance that the fragments do not contain sequences unique to the species of interest (such as the three species discussed).

Conclusion

In the presented report, a soil sample from a major American research university was sequenced using both Illumina and Nanopore sequencing methods. Both methods produced data with taxonomic classifications that match the expectation, consisting of largely soil-dwelling microbial DNA and contaminant *Homo sapiens* DNA from the experimenters. However, Illumina produced much higher quality reads, though lacking some species that were highly represented in Nanopore sequencing. Ultimately, the study illustrates the utility of nucleic acid sequencing as a tool for characterizing soil samples, with potential applications in improving agriculture through more advanced crop optimization.

References

1. Esposito, A., Colantuono, C., Ruggieri, V. *et al.* (2016) Bioinformatics for agriculture in the Next-Generation sequencing era. *Chem. Biol. Technol. Agric.* **3**, 9. https://doi.org/10.1186/s40538-016-0054-8
2. "How Nanopore Sequencing Works." Oxford Nanopore Technologies, 5 Apr. 2023, https://nanoporetech.com/support/how-it-works.

3. Timp, W. *Methods in Nucleic Acid Sequencing Lab: Introduction*, 2023.
4. "What Is the Illumina Method of DNA Sequencing?" @Yourgenome · Science Website, 21 July 2021, https://www.yourgenome.org/facts/what-is-the-illumina-method-of-dna-sequencing/.
5. Timp, W. *Methods in Nucleic Acid Sequencing Lab: Microbiome Sequencing Protocol*, 2023.
6. "Kraken." Johns Hopkins Center for Computational Biology, https://ccb.jhu.edu/software/kraken/.
7. Breitwieser, F.P., and Salzberg, S.L. "Pavian: Interactive Analysis of Metagenomics Data for Microbiome Studies and Pathogen Identification." Bioinformatics, vol. 36, no. 4, 2019, pp. 1303–1304., https://doi.org/10.1093/bioinformatics/btz715.
8. "Sorangium Cellulosum." Wikipedia, Wikimedia Foundation, 29 Jan. 2023, https://en.wikipedia.org/wiki/Sorangium_cellulosum.
9. Srinivas, A., Sasikala, C., & Ramana, C. V. (2014). Rhodoplanes oryzae sp. nov., a phototrophic alphaproteobacterium isolated from the rhizosphere soil of paddy. International journal of systematic and evolutionary microbiology, 64(Pt 7), 2198–2203. https://doi.org/10.1099/ijs.0.063347-0
10. Vieira, S., Luckner, M., Wanner, G., & Overmann, J. (2017). Luteitalea pratensis gen. nov., sp. nov. a new member of subdivision 6 Acidobacteria isolated from temperate grassland soil. International journal of systematic and evolutionary microbiology, 67(5), 1408–1414. https://doi.org/10.1099/ijsem.0.001827
11. Bromfield, E. S. P., Cloutier, S., & Nguyen, H. D. T. (2020). Description and complete genome sequences of Bradyrhizobium symbiodeficiens sp. nov., a non-symbiotic bacterium associated with legumes native to Canada. International journal of systematic and evolutionary microbiology, 70(1), 442–449. https://doi.org/10.1099/ijsem.0.003772