THIRD-GENERATION SEQUENCING TECHNOLOGY TO DEFINE MICROBIAL DIVERSITY IN WHEAT CULTIVATED SOILS IN TURKEY

by

JANA AL KHODOR

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfilment of
the requirements for the degree of
Master of Science

Sabanci University

July 2022

ABSTRACT

Jana al Khodor

Molecular Biology, Genetics, and Bioengineering, MSc, Thesis, July 2022

Thesis Supervisor: Assist. Prof. Stuart James Lucas

THIRD-GENERATION SEQUENCING TECHNOLOGY TO DEFINE MICROBIAL DIVERSITY IN WHEAT CULTIVATED SOILS IN TURKEY

Keywords: metagenomics, microbial consortium, DNA-seq, MinION, healthy bacteria

Microbial consortia or communities are the hallmark of a healthy soil rhizosphere and contribute to the wellbeing of domestic plants and animals, and consequently to the food production and food safety sectors. As the world population is constantly expanding, it is today a burden to meet their needs in food crops and cultivation. However, these are threatened by several environmental challenges including the excessive use of chemical fertilizers and climate change, which can harm healthy microorganisms residing within the rhizosphere. Instead, microbial communities are a natural and suitable option that can meet these challenges. The metagenomics approach overcomes the limitations of bacterial clonal cultures by studying the genomes of multiple bacterial genera or species in a given environmental area simultaneously. To decode the microbial diversity in wheat cultivated soils from five different regions in Turkey, we particularly used MinION Mk1B as a unique third generation nanopore sequencing technology. The 16S rRNA gene sequencing data, processed using Epi2ME Labs, reveals Proteobacteria as the predominant bacterial phyla (96-98%), as well as Bacteroidetes, Planctomycetes, Firmicutes, Verrucomicrobia, Actinobacteria, Acidobacteria, and Chloroflexi. We were able to identify community bacterial genera, *Ramlibacter* and *Massilia*, based on their co-occurrence pattern of distribution across the soil samples. Also, *Pseudoxanthomonas* was claimed as a bacterial genus that exhibit selective distribution pattern in certain soil samples, with several downstream species that were found to be involved in soil detoxification and recycling in the literature.

ÖZET

Jana al Khodor

Moleküler Bioloji, Genetik, ve Biomühendislik, Yüksek Lisans Tezi, Temmuz 2022

Tez Danışmanı: Dr. Öğr. Üyesi Stuart James Lucas

THIRD-GENERATION SEQUENCING TECHNOLOGY TO DEFINE MICROBIAL DIVERSITY IN WHEAT CULTIVATED SOILS IN TURKEY

Anahtar Kelimeler: metagenomik, mikrobial konsorsiyum, DNA-seq, MinION, sağlıklı bakteri

Mikrobiyal konsorsiyumlar veya topluluklar, sağlıklı bir toprak rizosferinin ayırt edici özelliğidir ve evcil bitki ve hayvanların refahına ve dolayısıyla gıda üretimi ve gıda güvenliği sektörlerine katkıda bulunur. Dünya nüfusu sürekli olarak artarken, günümüzde gıda ürünleri ve ekim alanlarındaki ihtiyaçlarını karşılamak bir yük haline gelmiştir. Bununla birlikte, bunlar, rizosferde yaşayan sağlıklı mikroorganizmalara zarar verebilecek aşırı kimyasal gübre kullanımı ve iklim değişikliği dahil olmak üzere çeşitli çevresel zorluklar tarafından tehdit edilmektedir. Bunun yerine mikrobiyal topluluklar, bu zorlukların üstesinden gelebilecek doğal ve uygun bir seçenektir. Metagenomik yaklaşım, belirli bir çevresel alanda aynı anda birden fazla bakteri cinsinin veya türünün genomlarını inceleyerek bakteri klonal kültürlerinin sınırlamalarının üstesinden gelir. Türkiye'nin beş farklı bölgesinden buğday ekili topraklardaki mikrobiyal çeşitliliği deşifre etmek için özellikle MinION Mk1B'yi benzersiz bir üçüncü nesil nano gözenek dizileme teknolojisi olarak kullandık. Epi2ME Labs kullanılarak işlenen 16S rRNA gen dizileme verileri, Proteobacteria'nın baskın bakteri filumunun (%96-98) yanı sıra Bacteroidetes, Planctomycetes, Firmicutes, Verrucomicrobia, Actinobacteria, Acidobacteria ve Chloroflexi olduğunu ortaya koymaktadır. Topluluk bakteri cinslerini, Ramlibacter ve Massilia'yı, toprak numuneleri boyunca birlikte oluşum dağılım modellerine dayanarak tanımlayabildik. Ayrıca, Pseudoxanthomonas'ın, literatürde toprak detoksifikasyonu ve geri dönüşüme dahil olduğu tespit edilen birkaç aşağı akış türüyle, belirli toprak örneklerinde seçici dağılım modeli sergileyen bir bakteri cinsi olduğu tercih edildi.

ACKNOWLEDGEMENTS

To my support system, mother, father, siblings, and sister-in-law. To whom birth brought happiness into my life, my beloved nephew, my cupcake, Hamza. Those years couldn't have passed without your encouragement and trust in me. I am here today thanks to your patience and love. Though I cannot stress enough the countless hard days I spent without you by my side, away from home… I shall never take you for granted.

To whom I owe this and every success henceforth, Dr. Stuart Lucas, you have been the best supervisor a graduate student could ever wish for. A massive thank you for your support, your invaluable guidance in my research, and your incredible kindness towards a student who started in the laboratory with null experience to a junior, not so bad, independent researcher like me. Words will never be enough…

I acknowledge TUBITAK for sponsoring and funding this unique project in which we are the very first to use this technology for soil microbiome DNA sequencing in Turkey. Many thanks as well to Prof. Yelda Özden and colleague Sena Nur Acet, PhD, from Gebze Technical University for their collaboration and contribution to this work.

I am very grateful for all the friendships I made here that helped this foreigner survive on campus. Yomna, Elif, Aybüke, Gülşah, Cemile, İlknur, Zülal, and Sümeyra, I am so lucky to have met you all. Never missing on my dearest friends in my homeland, what would I have done without your funniest meme messages and jokes at midnight…

And yet to my most beloved, future daughter and son, I hope you know this hard work has meant to strengthen me to achieve more in life, for you. I hope one day you'll stand proud of the mother and father you have, the ones who promise to love and cherish you unconditionally. To you, I shall offer this success…

“In this space between the earth and the wildflower crowns

the air thickens to balm with nectar infused winds…

I inhale deeply allowing my body to meld

with the sweetness of nature

giving me strength to carry on”

Courtney L. Smith

TABLE OF CONTENT

INTRODUCTION

1. Metagenomics

1.1 History

Launched in 1998 by Handelsman *et al.*, Metagenomics arose as a novel approach to overcome the limitations of clonal culture and single-whole-genome analysis that researchers have long faced with the classical genetic engineering approaches, given the necessity to manipulate and unravel the microbiome world, especially with microorganisms that cannot be cultivated (Nora et al., 2019). The concept of clonal cultures simply refers to culturing only one type or species of organisms, in which a single culture medium that contains multiple species would be termed as "contaminated". Hence, this limitation has always prevented the microbiologists from decoding the real treasure of our microbial planet. Handelsman *et al.* describes that the nearly 99.9% uncultured soil microflora represents the emerging "stunning", novel genetic diversity that could not be discovered at that time (Jo Handelsman et al., 1998); which has led to introducing the Greek term *meta* for "transcendent". In fact, the metagenome refers to the ensemble of the genetic material of different organisms in a given environmental sample, which is in our study the entire soil microbial (precisely bacterial) genome. Indeed, little has been known about how the genes of these soil microbes would contribute to their collective functions as community partners (J. Handelsman & Tiedje, 2007). The giant world of microbes further shows that even if cultivability is resolved to an extent, diversity remains a major issue due to the large number of microbial species in all environments. Given a particular field of study, microbial genomics can only access 1% of the genetic resources using traditional cultivation methods, whereas metagenomics offers a full access to the data for application through direct isolation of the DNA from the environment.

1.2 Contributions and applications

The rise of metagenomics in research has opened the door for various contributions in different domains, that were not applicable with clonal microbial genomics. These potential applications involve biotechnology, bioremediation, earth sciences and ecology, biomedical sciences, vaccines, bioenergy, sustainability, and agriculture. Particularly in agriculture, metagenomics helps understand the essential role of beneficial microbial communities and their effect on domestic plants and animals, and allows for the development of more effective detection methods for diseases that threaten the food production and food safety sectors (J. Handelsman & Tiedje, 2007). Due to the spontaneous increase in the world population, still

to date an estimate of 7.884 billion people living on five of the world's six continents with the highest records in Asia (60%) and Africa (17%) (Sadigov, 2022), there is a tremendous pressure to improve the current agricultural systems to satisfy the needs of the populations and provide larger quantities and better quality in food production (Priya et al., 2021). In other words, serious steps must be taken to reduce the use or dependence on chemicals as fertilizers, instead rely on beneficial microorganisms that have the potential to achieve this goal. Eventually, abiotic and biotic factors constitute the major stress caused by many environmental events, which surely affect the downstream soil microorganisms and hence the crops living in the field. Therefore, combining molecular techniques and bioinformatics tools as a metagenomic approach would reveal the complexity of the rhizosphere as an ecosystem and its microbial components.

2. Microbial consortium

We live in a microbial planet. And yet the idea of microbial consortium or microbiota started not long ago. The American molecular biologist Joshua Lederberg was the first scientist to suggest or introduce the term *microbiome* in 2001, after being one of the three recipients of the Nobel Prize in Medicine in 1958 awarded for their research in bacterial genetics (Sebastián-Domingo & Sánchez-Sánchez, 2018). Lederberg has focused his research on the microbiome in the human systems precisely, stating that symbiotic microorganisms and human together establish a strong metabolic unit, and that these are in fact healthy bacteria that protects the human body. Eventually, *microbiota* is defined as the community in which microorganisms share life in a determined ecological niche, such as soil rhizosphere, water, or human intestine. While the *microbiome* term refers to the entity or ensemble formed by these microbial species, their genomes, and metabolites in a determined ecological niche. Though, the interactions between the different species of microbes, especially bacteria, residing together within the same soil rhizosphere have not been studied comprehensively.

2.1 Role of soil bacteria in plant health

Bacteria are the most predominant form of life on planet Earth, and are found in every habitat like soil, rock, humans, animals, plants, oceans, and even arctic snow. For decades, people have thought that bacteria are primarily related to a disease or infection. On the contrary, this empire represents the hallmark of a healthy rhizosphere and plays a pivotal role in the nutrient cycles, soil formation, detoxification, decomposing organic matter, and more (Fatima et al., 2014). Consequently, the secondary metabolites generated by the bacterial degradation of organic matter attract more microorganisms to self-culture in carbon-enriched soil, thus

leading to increased fertility of the root ecosystem and reducing the need for chemical fertilizers and pesticides (Lugtenberg & Kamilova, 2009). Such beneficial bacteria are termed as Plant-Growth-Promoting-Rhizobacteria (PGPR). These may be applied to plants as whole microorganisms, microbial metabolites, or by seed inoculation. The two main categories of PGPR are known as rhizosphere bacteria that inhabit the soil around the plant root, as well as endophytic bacteria which are present inside the root tissues. Additionally, PGPR can indirectly promote plant growth and health by acting as biocontrol agents. Indeed, a variety of bacteriocins and antibiotics are produced by PGPR to cease the deleterious activities of harmful plant pathogens (Priya et al., 2021). Nevertheless, different types of bacteria and other microorganisms benefit the soil collectively for plant growth and productivity. These include: nitrogen-fixing bacteria, phosphate-solubilizing bacteria, protozoa, mycorrhizal fungi, soil-borne pathogens, and macroinvertebrates (Mendes et al., 2015). In other words, the metagenomic approach is designed to investigate the complex functional diversity and phylogenetics of the microbiome in the rhizosphere, together with metatranscriptomics and metaproteomics which target the functional genes and metabolic activities of proteins, to overall construct plant growth promoting bioformulations (Priya et al., 2021).

2.2 Why the microbial consortia approach?

As we discussed the importance of beneficial bacteria like PGPR in soil and plant health, it is however more valuable to rely on and apply the microbial consortia approach for several reasons. These microbial communities comprise member organisms of different taxa classification that, collectively, are much more robust and resistant to damaging, sometimes deleterious, environmental changes. In addition, they cooperate together to reduce the metabolic burden thanks to the division of labor (DOL) and resources exchange, therefore exhibit more powerful metabolic capabilities compared to monocultures or single-species fashion of agriculture (McCarty & Ledesma-Amaro, 2019). Not to mention that they establish chemical and physical communication patterns between different species (Bassler & Losick, 2006; Stenuit & Agathos, 2015). Overall, microbial consortia are pivotal to food production, recycling of micronutrients, and maintaining the health of humans, animals, and plants. These hallmark characteristics make the microbial consortia an attractive approach over bacterial monocultures to be incorporated in various biotechnological applications, including productive agriculture, fermentation, wastewater treatment, bioeconomy, and more.

In fact, this thesis is part of a larger project funded by TÜBİTAK (1001) entitled “Construction of wheat-specific microbial consortium and design of new biofertilizer

formulation via its immobilization to PolyHIPE polymer". Indeed, wheat is the most important cereal for Turkey, yet wheat production is facing several challenges. Most importantly, the relatively high production costs due to the usage of exported chemical fertilizers and uncontrolled water consumption. Hence, utilization of microbial fertilizers that have the ability to fix nitrogen, solubilize phosphorous and produce siderophore is preferred. However, the most preferred logical way to enhance yield is to identify plant-specific microbiome and design specific and new microbial consortium instead of usage of general known bacteria, given the advantages we discussed earlier. Ultimately, the project workflow is designed to isolate new microbial consortia from the wheat growing fields that have higher yield and will be combined with the PGB-invit; which is identified as endophytic, new, and specific bacterium (Şeker et al., 2017) in the previous TÜBİTAK project (117R002). Then this constructed microbial consortia will be immobilized to PolyHIPE polymer, which has the ability to trap and make controlled release of water and will be tested on two different wheat cultivars in greenhouse conditions.

3. Third-generation sequencing

3.1 The NGS revolution

The DNA sequence carries the genetic makeup of the cell and transmits it to the following generations by DNA replication. By decoding this sequence, scientists would be able to define the prevalence of threatening diseases within a population, unravel genetic mutations, process in the discovery of production of novel antibiotics and vaccines, establish new tools for food production and agriculture, and a lot more. In other words, the DNA sequence confers genetic information that the cell uses to make RNA molecules and proteins, that is essential to understand how the genomes of different organisms cooperate. Initially, the Sanger method was the first DNA sequencing platform to be developed, and the only available technology between 1975 and 2005. Although it produces relatively long reads (500 to 1000 bp), this technology still requires a prior amplification step which downsides involve copying errors and loss of information (Shendure et al., 2017). Also, it requires each individual sample to be resolved separately by gel or capillary electrophoresis, which was a major limitation on throughput in terms of cost and labor. For instance, the first Human Genome Project (1990 – 2003) which used entirely Sanger sequencing to interpret the 3.2 billion nucleotide base pairs of the human genome, cost $2.7 billion.

The "Next-Generation Sequencing" (NGS) revolution began in 2005, as 454 Life Sciences (Branford, CT, USA) first introduced the pyrosequencing platform as a high-throughput short

read technology (Margulies et al., 2005), that reduced the cost down to a fraction of the cost of Sanger sequencing and also the sequencing time. In fact, NGS has overcome Sanger sequencing in multiple ways (Shendure et al., 2017). Most importantly multiplexing, as a complex DNA library was immobilized onto a two-dimensional surface where all the templates were accessible to one reagent volume instead of using one tube per reaction. Second, *in vitro* amplification replaced bacterial cloning to generate several copies of each template to be sequenced. Also, instead of measuring fragment lengths, "sequencing-by-synthesis" detection method was introduced through polymerase-mediated binding of fluorescently labelled nucleotides. Though in order to maximize the detection of the light signals, a prior emulsion PCR amplification step is needed. Overall, these technologies are also known as second-generation sequencing platforms. The most famous platform in the NGS market today is Illumina, that can generate up to a billion bases in a single run using fluorescently labelled reversible terminator nucleotides (Bentley et al., 2008; Pareek et al., 2011). Interestingly, Illumina has taken great steps regarding the sequence accuracy to considerably minimize the error rates of their sequencers, hence making the short read length their primary downside.

However, sequencing technologies that require template amplification remained a challenge as they result in massive copying errors, sequence-dependent biases, information loss like methylation, excessive preparation time and complexity. Whereas researchers were aiming for more accurate DNA sequencing without read-length limitations. PacBio joined the race as the first approach of third-generation sequencing initiated by Webb and Craighead, and later developed by Korlach, Turner and Pacific Biosciences, with the aim of testing real-time PCR (Eid et al., 2009; Levene et al., 2003). PacBio uses a zero-mode waveguide by which only fluorescently tagged bases would be called, through a tiny hole less than half the wavelength of light where a single polymerase and the template can fit. Remarkably, this technology gives long reads ranging from 10 kb up to 100 kb, high yet random error rate, and tolerance of GC content.

3.2 Nanopore sequencing: ONT

Nonetheless, it took years of work to apply a newer concept regarding third-generation sequencing, that is nanopore sequencing. Basically, the nanopore is formed when the α-hemolysin protein is embedded into the biological membrane shaping a narrow hole or pore within which the "cyclodextrin" is covalently bound, given that cyclodextrin is the nucleotide binding site when the DNA molecule traverses the pore (Pareek et al., 2011) (figure 1).

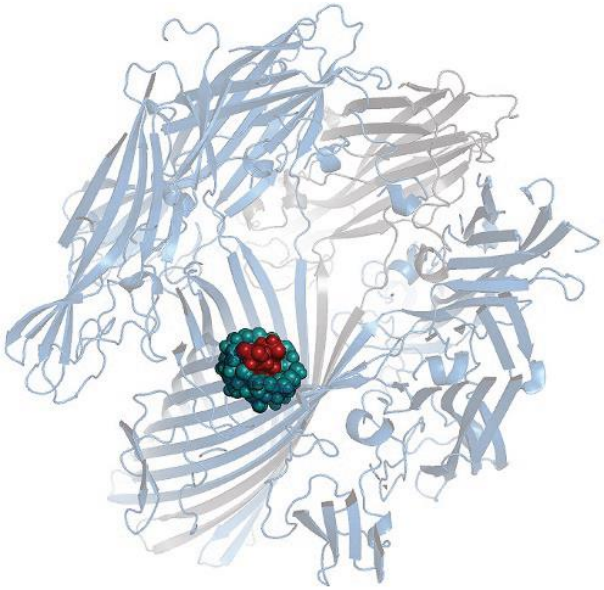

Figure 1. The nanopore structure showing the α-hemolysin protein (ribbon diagram) with covalently bound cyclodextrin as the nucleotide binding site (teal), and a nucleotide base traversing through the pore (red) (Oxford Nanopore Technologies; Rusk, 2009).

In this novel approach, one single-stranded DNA molecule passes through the nanopore causing ionic change across the membrane. As the DNA molecule traverses, a motor protein attached to the DNA strand at the barcoded adapter sequence pushes it through the nanopore at a constant speed (figure 2 A) and blocks the ionic current for a certain amplitude that is specific to each nucleotide basecalled A, G, C, and T (figure 2 B), which finally reveals the primary sequence of the DNA molecule (Astier et al., 2006; Rusk, 2009). Using adaptive sampling, it is possible to reject sequencing reads that are not of interest, leaving the nanopore solely available for the targeted DNA region of interest. In addition, the nanopore is uniquely designed to be capable of sequencing the methylcytosine base without bisulfite conversion, a feature that added a huge interest in the epigenetics research. In particular, Oxford Nanopore Technologies (ONT) is currently leading the third-generation sequencing community most famously for its real-time data analysis, relatively cheap, and rapid devices generating short to ultra-long reads of any DNA or RNA fragments. “Nanopore sequencing has the advantage that it does not require any labeling of the DNA, no expensive fluorescent reagents or really expensive CCD (charge-coupled device) cameras to record from optical chips,” says Hagan Bayley from ONT (Rusk, 2009). Besides, the scalability of ONT’s devices from portable to ultra-high throughput formats are designed and catered to the needs of researchers and their studies. These include NVIDIA DGX Station A100, PromethION, GridION, Flongle, and the most famous MinION.

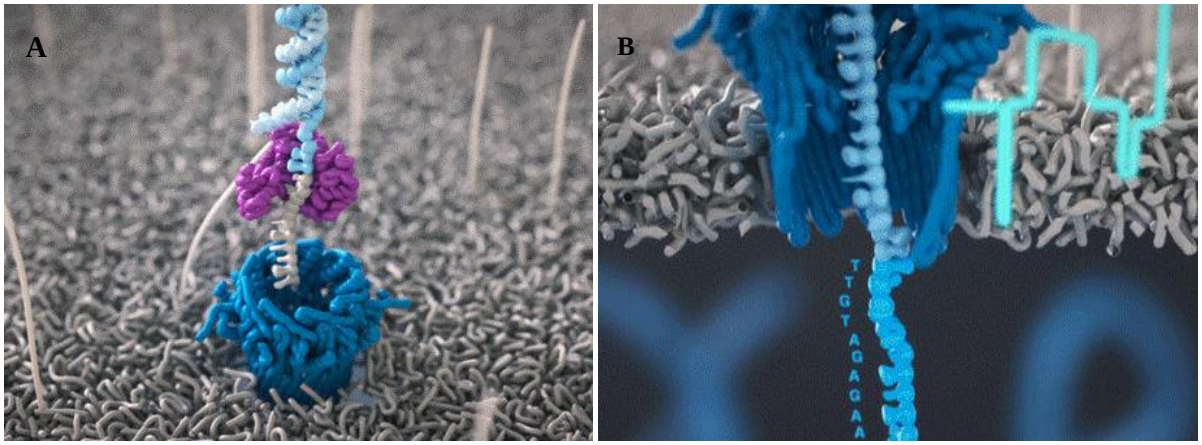

Figure 2. Illustration of the nanopore. **A:** a single-stranded DNA molecule (light blue) barcoded with the adapter sequence (grey) traverses through the nanopore with the help of a motor protein (purple). **B:** each nucleotide base disrupts the ion current across the membrane producing a different electrical signal that, consequently, reveals the sequence of the DNA strand (Oxford Nanopore Technologies).

3.3 Why MinION?

In fact, MinION was the first commercial product using nanopore technology. It is the only portable sequencing device with a size of a thumb that can plug in a USB port 3 of the computer, with the simplest configuration and low hardware requirement (Besser et al., 2018; Stoddart et al., 2009). Besides, it would be an astonishing achievement to sequence the human genome for cheap, provided that MinION is affordable starting from 1000$ only to generate short to ultra-long reads (>4 Mb) with data sets up to 50 Gb from a single flow cell. Following ONT's protocols, the library is easily prepared and loaded into the flow cell. Once plugged in, the sequencing run is ready to start and, shortly, the data can be accessibly analyzed in real-time while basecalling is done locally on the computer using MinKNOW program that first produces a Fast5 file format unique to ONT traces then converts it to Fastq file format for compatibility with other platforms, to be analyzed downstream in the cloud-based platform Epi2ME Labs, such as aligning the sequence reads to genome databases. A detailed illustration of the MinION Mk1B used in this study is discussed further below in the materials and methods section. Suitable applications of MinION include metagenomics, whole genomes/exomes, epigenetics, targeted sequencing, whole transcriptome (cDNA), and small transcriptomes (direct RNA).

3.4 MinION for soil microbiome sequencing

Given the fundamental characteristics mentioned above, MinION has been widely used in almost every research area, such as biomedical sciences, genomics, epigenetics, precision medicine, infectious diseases, agricultural studies, microbial metagenomics, and more. Indeed, researchers have studied the soil microbiome sequencing in different biogeographical areas using MinION.

For instance, Srivastava et al. recently analyzed the predominant bacterial communities present in the wheat rhizosphere of the Ghazipur regions of Eastern Indogangatic Plain in India (Srivastava et al., 2020). Like in any metagenome approach, wheat rhizosphere soil was collected, followed by DNA isolation from the soil, then amplification of the 16S RNA gene, sequencing using ONT MinION flow cell, and finally analysis of the dataset using Epi2ME platform. The sequencing resulted in 44,125 classified reads out of 51,909 reads in total. In terms of microbial diversity, the data analysis revealed the most dominant phyla as follows: Proteobacteria (68%), Firmicutes (13%), Bacteroidetes (3%), Actinobacteria (3%), and Acidobacteria (3%). Furthermore, the data at the species level classification showed that *Escherichia coli* is the most abundant species, then come *Candidatus solibacterusitatus* and *Achromobacter xylosoxidans*.

In another study, researchers were quite interested in life detection in the permafrost ice wedge of Axel Heiberg Island located in the Canadian high Arctic, stating that it is an analog to the polygonal permafrost terrain observed on Mars (Goordial et al., 2017). *In situ* field life detection was performed using cryo-iPlate for microbial culture by diffusing the *in situ* nutrients into semi solid media, in addition to the colorimetric assay to detect living species using Microbial Activity Microassay (MAM) plate. Similarly, the plates products and field samples were sequenced using both MinION rapid library prep kit and MinION low input library prep kit. Both methods resulted in similar functional and taxonomic profile of the microbial community in the ice wedge soil samples: bacteria were dominantly present in approximately 96%, though very low Archaea sequences (0.3-0.5%) were detected. With a minimal difference that the low input kit generated less viral origin sequences (0.2%) yet more eukaryotic reads (3%), compared to the rapid kit results (3% and 1% respectively). In total, 6,348 reads were generated by the low input kit with a mean length of 2,704 to 3,811 bp, while the rapid kit resulted in 9,530 reads with slightly shorter read length of 2,015 to 3,018 bp. Note that the Illumina MiSeq platform was also used for validation, and results

were truly consistent. At the level of phylum, Alphaproteobacteria, Actinobacteria, Acidobacteria, and Bacteroidetes were the most predominant in all three datasets; which is consistent again with previous molecular surveys at the same permafrost ice wedge soil site (Wilhelm et al., 2011) and with the wheat rhizosphere metagenome study in India mentioned above with slight proportional difference (Srivastava et al., 2020).

Besides, Mahoney and colleagues have studied the community structure, species variation, and potential functions of rhizosphere associated bacteria of different winter wheat (*Triticum aestivum*) cultivars in Washington, USA (Mahoney et al., 2017). Although DNA sequencing was performed by Illumina MiSeq, not MinION, the resulting data was still worth reviewing to evaluate the significance of our own generated data afterwards. The analyzed data of the rhizosphere composition revealed a total of 5,522,528 reads 350 bp in length, and consisted of 41% Proteobacteria, 17.4% Bacteroidetes, 16.7% Actinobacteria, 10.3% Acidobacteria, and 6% Gemmatimonadetes as the top 5 abundant phyla, with variable relative abundance of each phylum across their samples. Also, strong co-occurrence of community bacteria was detected between members of classes Alphaproteobacteria (genera Methylovirgula and Acidiphilium), Betaproteobacteria (genus Collimonas), Gammaproteobacteria (genus Serratia), Actinobacteria (genus Frankia), and Sphingobacteria (genus Mucilaginibacter).

Given all of these data references, the various advantages of third-generation nanopore sequencing over NGS technologies, the characteristics of MinION device, and particularly the fact that our 16S bacterial rRNA gene of interest is the size of 1.2 to 1.5 kb, using MinION for our soil metagenomic study is essential to unravel the bacterial diversity across different regions in Turkey, and attempt to define co-occurrence patterns of community bacterial genera or species that would highly serve the ultimate goal of the established project.

SCOPE OF THE THESIS

In our study, we aimed to investigate the microbiome of wheat rhizosphere in Turkey for bacterial distribution and variation through a metagenomic approach using third-generation sequencing platform. We have collaborated with Prof. Yelda Özden's team in Gebze Technical University for the TÜBİTAK 1001 project that aims to define wheat-specific microbial consortia, which we worked on in Sabanci University Nanotechnology Research and Application Center (SUNUM), then immobilize it to PolyHIPE polymer to be tested on different wheat cultivars. Our collaborators have accomplished soil sampling and DNA extraction. Yet, we also had dedicated trials to develop a soil DNA isolation protocol manually based on scientific literature. We proceeded with first step polymerase chain reaction (PCR), DNA purification using AMPure XP beads, second step barcoded PCR using ONT's adapter primers, and a second purification to eliminate any unbound primers and very short DNA fragments. Once the samples are prepared, we performed a Quant-IT assay to measure the molarity of each DNA sample and, according to the desired library concentration, we pooled a DNA library that served for setting the MinION flow cell. The sequencing would start right away, and at least 24 h later we were able to extract our dataset through the cloud Epi2ME labs platform to carry on with data analysis. The approach was repeated for 3 experiments, and we were able to obtain 3 remarkable datasets to analyze. We have identified the most dominant bacterial species across all the samples and have implemented a comparative analysis of the overall bacterial classification between our datasets, as well as between them and the experimental results of the soil microbiome sequencing that took place in other countries which we mentioned already. Nonetheless, to serve the ultimate goal of the project, we aimed to define particular microbial genera or species that we believed have significant co-occurrence, or selective, pattern of distribution across the different regions of cultivation in Turkey.

MATERIALS & METHODS

1. Soil sampling

To begin our metagenomic approach for soil DNA sequencing, soil sampling was carried out by our collaborators from Gebze Technical University (GTU) as follows. For soil acquisition, Diyarbakır, Konya, Ankara, Sivas, and Tekirdağ provinces in Turkey were selected, primarily due to the fact that they are major wheat production areas, they are geographically separated (figure 3), and have different soil structures, pH, Nitrogen, organic carbon, and phosphate levels that are predicted to provide microbial diversity. By consulting with provincial agricultural directorates, the fields suitable for the purchase of land management where the wheat cultivation program is high were determined and selected.

Figure 3. Geographical distribution of the five provinces in Turkey chosen for wheat-cultivated soil sampling for metagenomic sequencing of microbial consortia (en.wikipedia.org/wiki/Provinces_of_Turkey).

Soil collection and sampling were carried out according to (Simonin et al., 2020) for DNA isolation from rhizosphere soil. From each of the 5 provinces, 2 different soil samples were collected. In summary, at least 1 kg of soil (total 1 kg x 12 = 12 kg) from the land in each field was obtained at a depth of 0-15 cm from one of the 20 microregions to be displayed in the field to be sampled. Soil samples were analyzed for pH, 4 mm electrical conductivity, total organic carbon, total and usable nitrogen, and phosphate, with the assistance of Konya Gıda Agriculture University, Strategic Product Development, Application and Research Center (SARGEM). Also, the latitude and longitude of each soil sample's original location, type of agriculture (organic or traditional) and whether or not crop rotation is applied were learnt from the local farmer. In addition, 2 different soil samples, namely an unknown pathogen-infected potato dry soil sample P1, and a virus-infected tomato organic soil sample

A7H from a greenhouse in Antalya were included as examples of samples from non-wheat growing areas, only in experiment 3 of this study.

2. Optimization of soil DNA extraction protocols

We have experimented with soil DNA extraction in order to optimize a protocol that worked well. Primarily, it is vital to consider cell lysis efficiency and recovery when optimizing soil DNA extraction protocols specifically used to investigate the microbial consortium size or structure (Lever et al., 2015). Instead of the cationic surfactant CTAB that is usually used for plant DNA extraction, we have used phosphate buffer (PB) with sodium dodecyl sulfate (SDS) which is an anionic detergent that lyses the cell and denatures proteins without interfering with the DNA, and it is usually used for bacterial DNA extraction (Chatterjee et al., 2002). But prior to chemical lysis, we also used bead-beating of different sizes (3 mm and 5 mm) for primary mechanical lysis of the cells.

2.1 Soil DNA extraction using 1M PB, 0.5% SDS

This protocol was performed with a random sand soil sample in our laboratory. In brief, 200 mg of finely ground freeze-dried soil per sample were weighed into 2 ml tubes (We dried the soil with liquid nitrogen and finely ground them using mortar and pestle). The cells were mechanically lysed by bead beating using tungsten carbide beads in a swing mill at 25 Hz for 1 minute. For this, we used two different bead sizes: 2 samples with 2x 5 mm beads, and 2 samples with 5x 3 mm beads. Followed by chemical cell lysis, where 250 µl PB (1M PB with 0.5% SDS) was added, then vortexed for 10 seconds using a HS120209 vortexing unit. The samples were then incubated at room temperature for 10 minutes with shaking every minute for 5 seconds to facilitate the desorption of DNA. The samples were centrifuged at 7380 x *g* for 1 minute, and 90 µl of the supernatant was transferred to a new 2 ml tube (two phases were easily distinguished as a brownish pellet and an almost viscous supernatant). There, the supernatant was diluted (1:10) by adding 810 µl ddH_2O, then extracted with 900 µl phenol. Again, the samples were centrifuged at 7380 x *g* for 10 minutes, and 800 µl of the supernatant was pipetted into a new 2 ml tube. Next, the supernatant was extracted twice with chloroform:IAA in (24:1) ratio, which removes any residual phenols, proteins, lipids, and detergents by dissolution or accumulation at the aqueous interface. For DNA precipitation, 20% PEG-NaCl (polyethylene glycol 8000-5M NaCl) in (2:1) ratio was added, then the samples were centrifuged. The DNA pellets were almost invisible after centrifugation; we incubated them at 37 ^{0}C for 15 minutes, then centrifuged them again at maximum speed *g* for

15 minutes. Finally, the DNA pellets were washed once with ethanol, allowed to dry, then resuspended in 30 µl TE buffer for elution.

2.2 Optimizing soil DNA extraction using 120 mM PB, 5% SDS

Since the soil DNA extraction protocol above did not seem to be efficient judging by negative results of the agarose gel visualization of the samples, we worked on optimizing multiple key steps based on the literature. Starting with weighing 200 mg of finely ground freeze-dried soil into 2 ml tube per sample. The cells were mechanically lysed by bead beating using 5 tungsten carbide beads of 3 mm size only, as seen in the literature to work enough for the extraction, in a swing mill at 25 Hz for 1 minute. In this protocol, we optimized the PB solution by lowering the molarity down to 120 mM with a higher amount of SDS up to 5%. Of that prepared solution, 1 ml was added, and the samples were vortexed for 10 seconds. Then we increased the incubation period up to 1 hour at 65^0C with occasional stirring. The samples were centrifuged at 8,000 rpm for 10 minutes at 4^0C, and we transferred 900 µl of the supernatant of each to a new 2 ml tube. This time, the supernatant was extracted only once with chloroform:IAA (24:1) to avoid much loss of the DNA. Particularly for DNA precipitation, we tested a different precipitation method for each of the two samples. In the sample 1, half volume of 50% PEG and 1 volume of 5 M NaCl were added. Whereas the sample 2 was precipitated with 0.1 volume of 3 M sodium acetate and 2 volumes of ethanol. Then, both of the samples were incubated overnight at 4^0C for the maximum DNA yield possible. The next day, the samples were centrifuged at 12,000 rpm at 4^0C for 10 minutes to recover the pellets. Next, the pellets were washed twice with 200 µl of 75% ethanol for better DNA purity. Finally, the samples were resuspended in 30 µl TE buffer and visualized on nanodrop (Thermo Scientific Nanodrop 2000C Spectrophotometer). Apparently, the precipitation using 3 M sodium acetate and ethanol was not successful based on the nanodrop results of sample 2. However, a little bump could be reported in sample 2 that was worth running on agarose gel. From there, we have confirmed the efficiency of 50% PEG-5M NaCl to precipitate DNA, and therefore has been admitted.

3. Soil DNA extraction by ZymoBIOMICS

In our study, DNA was extracted from all the soil samples using the ZymoBIOMICS™ DNA Miniprep Kit (Zymo Research, D4300), according to the manufacturer's protocol. Briefly, for each sample, 250 mg of soil was weighed into 2 ml lysis tubes and 750 µl ZymoBIOMICS™ Lysis Solution was added. Samples were homogenized by bead-beating at maximum speed for 3 minutes to ensure complete lysis of the cells and access to the nucleic acid material. The

samples were centrifuged in a microcentrifuge at 10,000 x *g* for 1 minute. Then, 400 µl of the supernatant was transferred to the Zymo-Spin™ III-F Filter with a collection tube and centrifuged at 8,000 x *g* for 1 minute. The Zymo-Spin™ III-F Filter was discarded, and the filtrate was thoroughly mixed with 1,200 µl of ZymoBIOMICS™ DNA Binding Buffer. Next, only 800 µl of the mixture was transferred to a Zymo-Spin™ IICR Column in a collection tube and centrifuged at 10,000 x *g* for 1 minute. This step was repeated after discarding the flow through from the collection tube. After that, three DNA washes were applied to the Zymo-Spin™ IICR Column in a new Collection Tube, each followed by centrifugation at 10,000 x *g* for 1 minute then discarding the flow through, in the following order: first wash with 400 µl ZymoBIOMICS™ DNA Wash Buffer 1 (≤25% ethanol, ≤25% propan-2-ol), second wash with 700 µl ZymoBIOMICS™ DNA Wash Buffer 2 (≤25% ethanol, ≤25% propan-2-ol), and third wash with 200 µl ZymoBIOMICS™ DNA Wash Buffer 2. The column was then placed in a clean 1.5 ml microcentrifuge tube and 100 µl ZymoBIOMICS™ DNase/RNase Free Water was added directly to the column matrix. After a 1-minute incubation, the DNA got eluted by centrifugation at 10,000 x *g* for 1 minute again. Finally, to filter the DNA, a Zymo-Spin™ III-HRC Filter was prepared in a new collection tube and 600 µl ZymoBIOMICS™ HRC Prep Solution was added and centrifuged at 8,000 x *g* for 3 minutes. Once the filter was ready, the eluted DNA from the previous step was transferred to it with a clean 1.5 ml microcentrifuge tube, and then centrifuged exactly at 16,000 x *g* for 3 minutes. This would finally yield the eluted the DNA ready for the downstream experiments, precisely PCR.

4. Nanodrop quantification

Before carrying out with PCR experiments, the samples were quantified for DNA yield and purity using Thermo Scientific Nanodrop 2000C Spectrophotometer through 260 – 280 nm range. Nuclease-free water of 1.2 µl was first used to blank the system, and then 1.2 µl of each eluted DNA sample was measured.

5. First step Polymerase Chain Reaction (PCR)

The ONT's sequencing methodology has the advantage of not necessarily requiring an amplification step before sequencing. However, we aimed to increase the specificity of this approach by solely focusing on the bacterial 16S rRNA gene (figure 4) for our data.

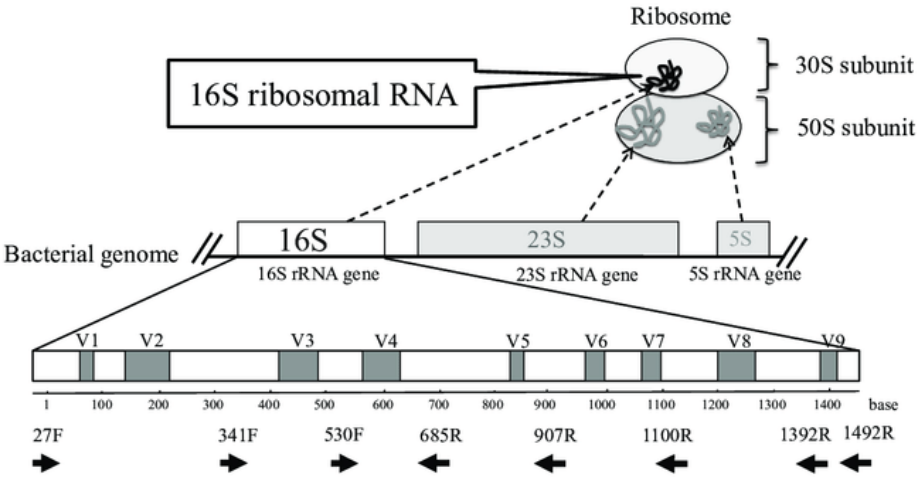

Figure 3. Structure of the bacterial 16S rRNA gene showing the binding sites of the 27F and 1492R primers used (Fukuda et al., 2016).

By amplifying this gene with a first step PCR, the prevalence of any other genome regions would reduce effectively, hence the increase in the bacterial 16S rRNA gene copies primarily targeted with the downstream PCR barcoding and sequencing. First step PCR was performed for all the samples using KAPA PCR reagents and the following primers for the bacterial 16S rRNA gene tailed with the ONT primer adapter sequences:

Ec27F-ONT1: TTTCTGTTGGTGCTGATATTGCAGAGTTTGATCCTGGCTCAGATTGA-3'

Ec1492R-ONT2: ACTTGCCTGTCGCTCTATCTTCCGATACGG**Y**TACCTTGTTACGACTT-3'

- The underlined bases allow binding to the ONT adapter primers
- The '**Y**' base means that the primer contains an equal mixture of pyrimidines (C or T) at this position, to allow binding to the widest possible variety of bacterial genome

5.1 1st PCR Reaction Conditions:

PCR reactions (total volume 20 µl per sample) were set up for the soil DNA samples as well as a negative control as follows, unless stated otherwise:

- 2 µl Kapa buffer (10X, 1.5 mM Mg at 1X)
- 0.4 µl $MgCl_2$ (25 mM)
- 0.6 µl dNTP (10 mM)
- 0.2 µl F primer – Ec27F-ont1 (10 µM)
- 0.2 µl R primer – Ec1492R-ont2 (10 µM)
- 0.1 µl *Taq* polymerase (5 U/µl, 500 U)
- 15.5 µl H_2O
- 1 µl DNA template

5.2 1st PCR Program JANA1:
Using *Eppendorf 5331 MasterCycler Gradient* thermal cycler*:*

- Initial denaturing, manual warm start, 95^0C, 5 min
- 35 cycles of: 95^0C for 30s, 60^0C for 30s, 72^0C for 90s (then select *Repeat/Hold at 2* for 34 cycles left
- Final extension, 72^0C, 5 min

6. Agarose gel electrophoresis

For visualizing the PCR products and proof of the successful amplification, 1% agarose gel electrophoresis was performed for all the experiments. To prepare the gel, 1 g of the agarose powder was dissolved into 100 ml of unused 0.5X TBE (or half the quantity for a short board gel, depending on the number of samples to be run) and stained with 1 µl Gel Red. Beside the 1 kb DNA ladder (Thermo Scientific GeneRuler, 0.5 µg/µl) as well as A5C and A15H as positive controls, the PCR products were mixed with 2x loading dye on (1:1) ratio, loaded, and run for 25 min, at 100 V. Then, the gel was visualized on the ImageLab software using UV tray of GelDoc EZ Bio-Rad product with the Stuart GelRed Program.

7. DNA purification

In order to remove any unbound primers or short fragments of DNA that did not succeed to amplify enough and thus will affect the quality of sequencing, the 1st step PCR products were cleaned up using the AMPure XP beads which have the ability to bind specifically long fragments, double stranded DNA. The purification was performed based on Oxford Nanopore's '4 primer PCR' protocol as follows.

Before initiating the experiment, the AMPure XP beads were allowed to come to room temperature and vortexed well. First, the PCR products were transferred to clean 1.5 ml Eppendorf tubes. For each reaction, 0.8x of resuspended AMPure XP beads was added, mixed by flicking the tube, and incubated on the rotator for 5 minutes at room temperature. Meanwhile, 500 µl fresh 70% ethanol per sample was prepared using Millipore or Nuclease-free water. After the incubation, the samples were spined down and pelleted on a magnet to pipette off the supernatant. While kept on the magnet, each sample was washed with 200 µl of the freshly prepared 70% ethanol carefully not to disturb the pellet. The ethanol was then removed by pipette and discarded. This washing step was repeated again to ensure the purification of the DNA material. Knowing that traces of ethanol might not be removed

completely, the samples were spined down briefly, placed back on the magnet, and any residual 70% ethanol was pipetted off. The samples were allowed to air dry briefly. Now to elute the DNA again for further applications, the samples were removed from the magnetic rack, and the pellets were resuspended in 10 µl of prepared 10 mM Tris-HCl pH 8.0, with 50 mM NaCl. Once incubated for 2 minutes at room temperature, the beads were pelleted back on the magnet until the eluate was clear and colorless. Finally, 10 µl of the DNA eluate was pipetted and retained into clean 1.5 ml Eppendorf tubes.

To verify this purification, the samples were visualized again on 1% agarose gel that would show only clear DNA bands in the expected size range, without any primer fragments left at the bottom of the gel. Only then, the samples would be ready to use for the next experiment directly or stored at -20^{0}C for later.

8. PCR barcoding

Also referred to as second-step PCR, PCR barcoding uses barcoded adapter primers to allow multiplexing of different samples. These adapter primers would attach to targeted regions of the genome to be analyzed through sequencing and basecalling. Based on the ONT's protocol, here we set up the experiment using the 2x LongAmp HotStart PCR mix (New England Biolabs), which is optimized for longer PCR products, and barcode primers from the PCR Barcoding Kit (Oxford Nanopore Technologies, SQK-PBK004 kit).

8.1 2nd PCR Reaction Conditions:

In 0.2 ml thin-walled PCR tubes, we prepared:

- 23 µl Nuclease-free water
- 1 µl ONT barcode primer (BP01, BP02 etc. – one per sample)
- 1 µl XP cleaned, first step PCR product
- 25 µl LongAmp 2x PCR master mix

8.2 2nd PCR Program JANA2:

- Initial denaturing, 94^{o}C, 1 min
- 10 cycles of: 94^{o}C, 30s; 62^{o}C, 30s; 65^{o}C, 1 min 45 s.
- Final extension, 65^{o}C, 5 min.

This 10-cycle program named *JANA2* was used for the first set of soil DNA sample generating the first dataset. However, we tended to increase the number of cycles to 25, with

the program named *JANA225*, for the second and third datasets since we appreciated a higher quantity of PCR products for better quality and yield sequencing.

The 2nd step PCR products were visualized on 0.9% agarose gel with GelRed at 100 V for 25 minutes alongside 1 kb DNA ladder. All samples were expected to only have a single, bright PCR product in the expected size range. Next, the samples had to be cleaned up once again with the AMPure XP beads previously described. Note that the PCR products were either directly purified or stored at 4°C, but not frozen, since we planned to proceed with the sequencing protocol shortly within days.

9. Quantification and pooling

Once amplified and purified, the 2nd step PCR products were quantified then pooled into a barcoded DNA library for sequencing. Based on the Quant-It dsDNA Assay kit protocol, the working buffer solution was first prepared by diluting the fluorophore in the kit buffer solution at (1:200) ratio. For setting up the Quantus Fluorometer, we blanked then measured one standard (DNA reagent 8) at (1:200) ratio by diluting 10 µl of the standard reagent in 190 µl of working buffer solution. Then, the samples were measured at the same ratio by diluting 1 µl of each sample DNA in 199 µl working buffer solution. The concentrations were obtained in ng/µl then calculated in *f*mol/µl in order to pool the library in the desired ratio that is, based on the ONT's protocol, between 50 and 100 *f*mol/µl. According to the NEBioCalculator website, for 1.5 kb DNA fragments, 1 ng = 1.08 *f*mol thus each sample concentration was multiplied by 1.08 to get the *f*mol/µl measurements for all. After choosing a certain molarity per sample (e.g., 20 *f*mol per sample as in experiment 1), the pooled volume of each 2nd PCR product is calculated by dividing the desired number of *f*mol by the concentration of the sample in *f*mol/µl (V= 20 (*f*mol) / 52.92 (*f*mol/µl) = 0.37 (µl) of sample 1, etc.). Likewise, the volumes obtained are pipetted and pooled into a clean 1.5 ml Eppendorf tube and diluted in 10 µl of 10 mM Tris.HCl pH 8.0 with 50 mM NaCl. The library was kept at 4°C until we were ready to sequence.

10. MinION sequencing

On the sequencing day, the amplified 10 µl DNA library was first set by adding 1 µl RAP from the PCR Barcoding Kit (Oxford Nanopore Technologies, SQK-PBK004 kit), mixing gently by flicking the tube and spinning down briefly. The reaction was incubated for 5 minutes at room temperature then stored on ice until ready to load.

10.1 SpotON Flow Cell priming

Prior to loading, the SpotON Flow Cell embedded into the MinION Mk1B drive was necessarily primed using the Flow Cell Priming Kit (Oxford Nanopore Technologies, EXP-FLP002) and following ONT's protocol. Initially, the Sequencing Buffer (SQB), Loading Beads (LB), Flush Tether (FLT), and one tube of Flush Buffer (FB) were thawed at room temperature. Next, the Sequencing Buffer (SQB), Flush Tether (FLT), and Flush Buffer (FB) were thoroughly mixed by vortexing and spined down at room temperature.

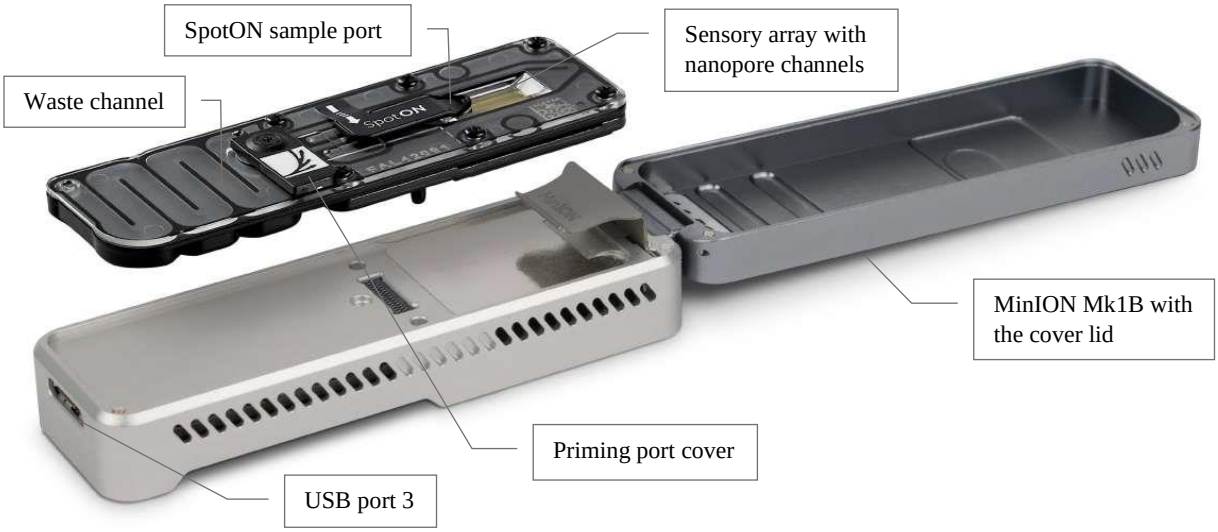

Figure 5. Diagram of annotated MinION Mk1B showing the sample port, priming port, waste channel, nanopore channels, USB port, and the cover lid (Oxford Nanopore Technologies).

Practically, MinION Mk1B was set up by opening its lid and sliding the SpotON Flow Cell under the clip (figure 5). The flow cell could be pressed down firmly to ensure correct thermal and electrical contact. The second step was to slide the priming port cover clockwise to open the priming port. There, we would check for small air bubbles within the priming channel. To get rid of any, we would set a P1000 pipette to 200 µl, insert the tip into the priming port, and draw back the wheel until we could see a small volume of buffer entering the pipette tip. There must be continuous buffer from the priming port across the sensor array. Now to prepare the flow cell priming mix, 30 µl of thawed and mixed Flush Tether (FLT) was added directly to the tube of thawed and mixed Flush Buffer (FB) and mixed by vortexing at room temperature. Of that priming mix, 800 µl was loaded into the flow cell via the priming port, avoiding the introduction of air bubbles, and left to prime. Five minutes later, the SpotON sample port cover was gently lifted to make the SpotON sample port accessible. A final volume of 200 µl of the priming mix was loaded again into the flow cell

via the priming port (not the SpotON sample port), avoiding the introduction of air bubbles as well.

10.2 Library loading

Meanwhile, the library was prepared in a new tube for loading as follows:

- 34 µl Sequencing Buffer (SQB)
- 25.5 µl Loading Beads (LB)
- 4.5 µl Nuclease-free water
- 11 µl amplified DNA library

Note that the Loading Beads (LB) tube contains a suspension of beads which settle very quickly. Therefore, it is vital that they are mixed immediately and only before use.

Just prior to loading, the prepared library was mixed gently by pipetting up and down not to leave the beads settled. The total of 75 µl library sample was added to the flow cell via the SpotON sample port in a dropwise fashion. It is crucial to ensure that each drop flows into the port before adding the next. Once finished, the SpotON sample port cover was gently replaced back making sure the bung enters the SpotON port. Likely, the priming port was closed, and the MinION Mk1B lid was finally replaced.

10.3 Starting the sequencing

The MinION Mk1B drive was plugged in to the computer via USB3 port. By already installing the MinKNOW software on the computer, the sequencing device control, data acquisition and real-time basecalling were carried out smoothly. The software allows for naming the sequencing experiment and dataset, selecting different options in the sequencing parameters, and performing a control check of the drive before starting the actual run. For a new flow cell, the control run would show up to 1200 active pores that decreases remarkably for a used one. When ensuring the drive is set up well, the experiment parameters can be defined, and the sequencing run can be started. The MinION drive is delicate and must be plugged in until the sequencing run is completely done, which may take up to 72 hours. However, in our study the sequencing run was completed after 24 hours only though real-time results were accessible while running, and these mainly include state time equivalent revealing the percentage and status of pores throughout sequencing, channels count, reads count, temperature, voltage, read length histogram, Qscore, and barcode hits for passed/failed bases.

11. Data analysis with Epi2ME

The Fastq sequence file format produced by MinKNOW was analyzed using ONT's own cloud-based platform, Epi2ME. Particularly, the Epi2ME Labs platform uses Python as a programming language and offers various bioinformatics tutorials designed for any sequencing approach to easily generate the dataset in the desired format. First, we were able view our data as a Fastq 16S easy-to-interpret report sheet on the epi2me.nanoporetech.com website by selecting the Fastq 16S Workflow which involves a number of parameters: min Qscore 7, length filter at 500 bp, e value = 0.01, min coverage of 30%, and min identity of 77%. The generated report provides a general overview of the dataset in terms of total classified read count, read count per barcode ID, cumulative reads per taxa, NCBI-based taxonomy tree, average sequence length, and more. Furthermore, from Epi2ME Labs Launcher installed on the computer and, through the "Analysis of Epi2ME 16S CSV Output" notebook, the raw dataset was uploaded and assessed through a series of bioinformatics steps to finally extract it in the desired BIOM format. The file was then downloaded to the desktop and further proceeded with Phinch2 (version 2.0.1) for complex data analysis (figure 6).

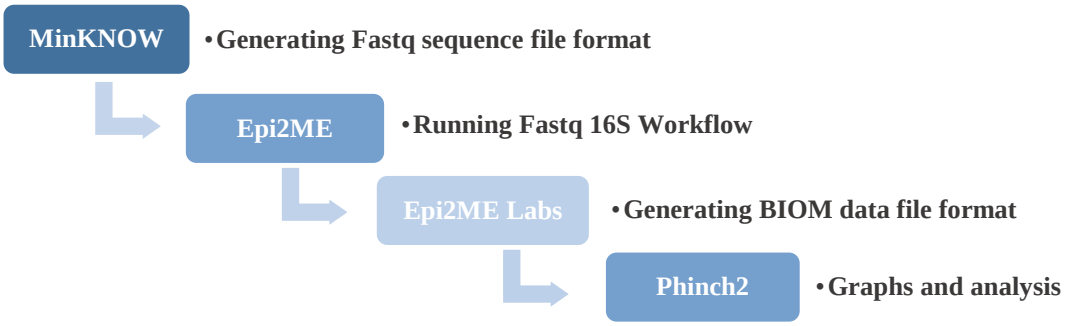

Figure 6. Schematic illustration of MinION data analysis pathway. MinKNOW is first used to obtain the Fastq sequence file format of the data, then it is run in Epi2ME cloud platform through the Fastq 16S Workflow. The resulting dataset is uploaded to Epi2ME Labs notebook, precisely "Analysis of Epi2ME 16S CSV Output" to generate the data in BIOM format file. The BIOM format file is finally uploaded into Phinch2 for complex data analysis and graph designing.

RESULTS

1. Soil DNA extraction on nanodrop

1.1 Experiment 1

For the first experiment, 12 soil samples were studied, 2 from each of the five different regions of Turkey (Diyarbakır 1 and 2, Konya 1 and 2, Ankara 1 and 2, Sivas 1 and 2, and Tekirdağ 1 and 2) as well as 2 extra samples Ankara1r and Tekirdağ2r, which were isolation repeats of the two samples for which DNA concentrations were < 25 ng/µl. The nanodrop measurements of the extracted DNA from these samples using ZymoBIOMICS™ DNA Miniprep Kit (Zymo Research, D4300) are evaluated below (figure 7).

08/12/2021 09:56

#	Sample ID	User name	Date and Time	Nucleic Acid	Unit	A260 (Abs)	A280 (Abs)	260/280	260/230	Sample Type	Factor
1	Diyar1	SUUSER	08/12/2021 09:41:43	94.5	ng/µl	1.891	1.007	1.88	1.40	DNA	50.00
2	Diyar2	SUUSER	08/12/2021 09:43:56	37.7	ng/µl	0.755	0.395	1.91	1.07	DNA	50.00
3	Konya1	SUUSER	08/12/2021 09:45:23	61.2	ng/µl	1.223	0.689	1.77	1.06	DNA	50.00
4	Konya2	SUUSER	08/12/2021 09:46:25	34.5	ng/µl	0.689	0.371	1.86	1.10	DNA	50.00
5	Tekirdag1	SUUSER	08/12/2021 09:47:25	35.9	ng/µl	0.719	0.394	1.82	1.22	DNA	50.00
6	Tekirdag2	SUUSER	08/12/2021 09:48:28	23.4	ng/µl	0.467	0.247	1.89	1.03	DNA	50.00
7	Sivas1	SUUSER	08/12/2021 09:49:44	61.2	ng/µl	1.225	0.681	1.80	1.36	DNA	50.00
8	Sivas2	SUUSER	08/12/2021 09:50:37	35.5	ng/µl	0.709	0.393	1.81	0.97	DNA	50.00
9	Ankara1	SUUSER	08/12/2021 09:51:38	16.1	ng/µl	0.322	0.166	1.94	0.95	DNA	50.00
10	Ankara2	SUUSER	08/12/2021 09:52:27	28.9	ng/µl	0.579	0.308	1.88	1.19	DNA	50.00
11	Ankara1r	SUUSER	08/12/2021 09:53:29	42.2	ng/µl	0.844	0.461	1.83	0.27	DNA	50.00
12	Tekirdag2r	SUUSER	08/12/2021 09:55:03	14.1	ng/µl	0.281	0.161	1.74	0.53	DNA	50.00

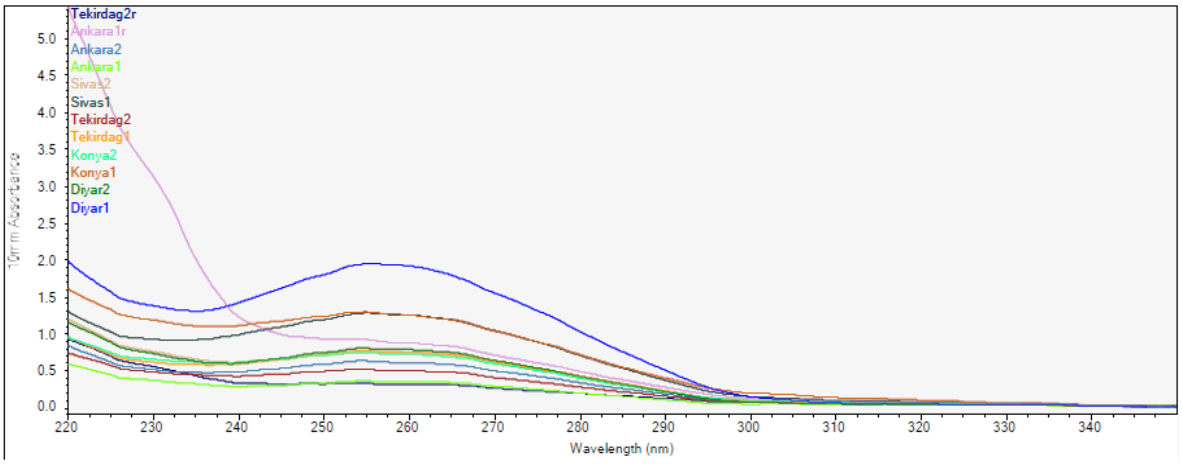

Figure 7. Absorbance spectra of extracted soil DNA from 12 wheat samples of experiment 1.

The table shows that the samples do contain isolated DNA and the yield is generally fine for all, note that Diyarbakır1 has the highest DNA concentration with 94.5 ng/µl while the lowest DNA yields are for Tekirdağ2r then Ankara1 with 14.1 and 16.1 ng/µl respectively. As for DNA purity, the ratio of absorbance at 260/280 nm reveals that the samples are ideally pure (±1.8), which indicates the absence of proteins, phenols, or other contaminants that absorb strongly at or near 280 nm. Regardless of their concentrations, Ankara1 sample shows the highest DNA purity record of 1.94, while Tekirdağ2r sample marks the lowest, still acceptable, DNA purity of 1.74. With the exception of Ankara1r sample showing a significant peak on the graph (pink) that implies major contamination by phenols with only a

0.27 ratio at 260/230 nm absorbance, therefore the moderate concentration of 42.2 ng/µl is not truly reliable. In that sense, better sequence results are to be expected from Ankara1 instead, given the better purity level of the sample despite the lower concentration. Nonetheless, we carried on with the samples given their good concentrations and purity on average.

1.2 Experiment 2

For the same batch of soil samples as in experiment 1 yet without any sample repeats, a second round of DNA extraction for a total of 10 samples only was performed by the same method and these are the results on Nanodrop (figure 8)

25/04/2022 14:47

#	Sample ID	User name	Date and Time	Nucleic Acid	Unit	A260 (Abs)	A280 (Abs)	260/280	260/230	Sample Type	Factor
1	Experiment2 Gebze DNA isolation - Diyar1 - 25.04.22	SUUSER	25/04/2022 14:24:30	39.9	ng/µl	0.798	0.422	1.89	0.98	DNA	50.00
2	Experiment2 Gebze DNA isolation - Diyar2 - 25.04.22	SUUSER	25/04/2022 14:25:53	31.6	ng/µl	0.632	0.334	1.89	0.82	DNA	50.00
3	Experiment2 Gebze DNA isolation - Konya1 - 25.04.22	SUUSER	25/04/2022 14:26:59	133.0	ng/µl	2.660	1.695	1.57	0.47	DNA	50.00
4	Experiment2 Gebze DNA isolation - Konya2 - 25.04.22	SUUSER	25/04/2022 14:27:55	36.9	ng/µl	0.738	0.416	1.77	0.72	DNA	50.00
5	Experiment2 Gebze DNA isolation - Ankara1 - 25.04.22	SUUSER	25/04/2022 14:28:54	26.7	ng/µl	0.535	0.302	1.77	0.67	DNA	50.00
6	Experiment2 Gebze DNA isolation - Ankara2 - 25.04.22	SUUSER	25/04/2022 14:29:42	19.4	ng/µl	0.388	0.222	1.75	0.60	DNA	50.00
7	Experiment2 Gebze DNA isolation - Sivas1 - 25.04.22	SUUSER	25/04/2022 14:31:34	43.6	ng/µl	0.871	0.487	1.79	0.88	DNA	50.00
8	Experiment2 Gebze DNA isolation - Sivas2 - 25.04.22	SUUSER	25/04/2022 14:32:32	30.3	ng/µl	0.606	0.350	1.73	0.77	DNA	50.00
9	Experiment2 Gebze DNA isolation - Tekirdag1 - 25.04.22	SUUSER	25/04/2022 14:33:39	30.9	ng/µl	0.618	0.352	1.76	0.88	DNA	50.00
10	Experiment2 Gebze DNA isolation - Tekirdag2 - 25.04.22	SUUSER	25/04/2022 14:34:26	247.0	ng/µl	4.939	4.276	1.16	0.77	DNA	50.00

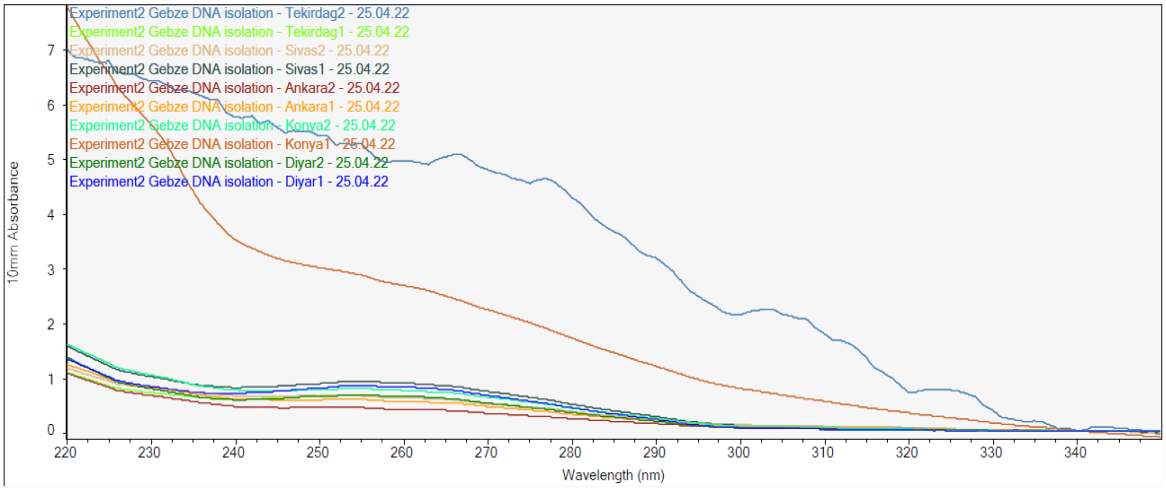

Figure 8. Absorbance spectra of extracted soil DNA from 10 wheat samples of experiment 2.

Interestingly, we have got lower DNA concentrations for the samples on average compared to those of experiment 1, as well as less pure DNA measurements at 260/280 nm. Particularly, the samples of issue were Konya1 and Tekirdağ2. For instance, Tekirdağ2 shows the very highest DNA concentration (247 ng/µl) yet the lowest purity ratio of 1.16 at 260/280 nm and 0.77 at 260/230 nm, which only signifies for the strong contamination by proteins or phenols during the extraction process. Similarly, very low purity (1.57 at 260/280 nm, 0.47 at 260/230 nm) is observed in Konya1 which DNA yield is relatively high (133 ng/µl). Generally, the samples' measurements were not as good as expected, but we were interested to see whether or not they would amplify through 1st step PCR and thus carried on with them.

1.3 Experiment 3

For this last sequencing experiment, we used the same DNA extractions as Experiment 2 and in addition extracted the DNA with ZymoBIOMICS™ DNA Miniprep Kit (Zymo Research, D4300) from 2 soil samples that have been already stored at -20°C in our lab for previous research interests, namely P1 and A7H. In fact, P1 is a dry potato soil sample collected from a pathogen-infected field in Karaman province, whereas A7H is an organic tomato soil sample originally from a pathogen-infected greenhouse in Antalya. Our objective was to compare the microbiome diversity in soils different from wheat crops, as well as to examine the effectiveness of the ZymoBIOMICS™ DNA Miniprep Kit to extract DNA from a variety of soil complexity (dry, organic, etc...).

At last, the purified barcoded PCR products of these samples were quantified alongside those of the set of 10 samples studied in experiment 2, and together pooled for a desired library ratio for the third sequencing run (12 samples in total). Regardless, the nanodrop results of soil DNA extraction from P1 and A7H are right below (figure 9).

28/04/2022 12:59

#	Sample ID	User name	Date and Time	Nucleic Acid	Unit	A260 (Abs)	A280 (Abs)	260/280	260/230	Sample Type	Factor
1	P1 DNA isolation - Zymo kit - 28.04.22	SUUSER	28/04/2022 12:55:53	208.3	ng/µl	4.165	2.249	1.85	1.66	DNA	50.00
2	A7H DNA isolation - Zymo kit - 28.04.22	SUUSER	28/04/2022 12:58:11	157.3	ng/µl	3.146	1.681	1.87	1.66	DNA	50.00

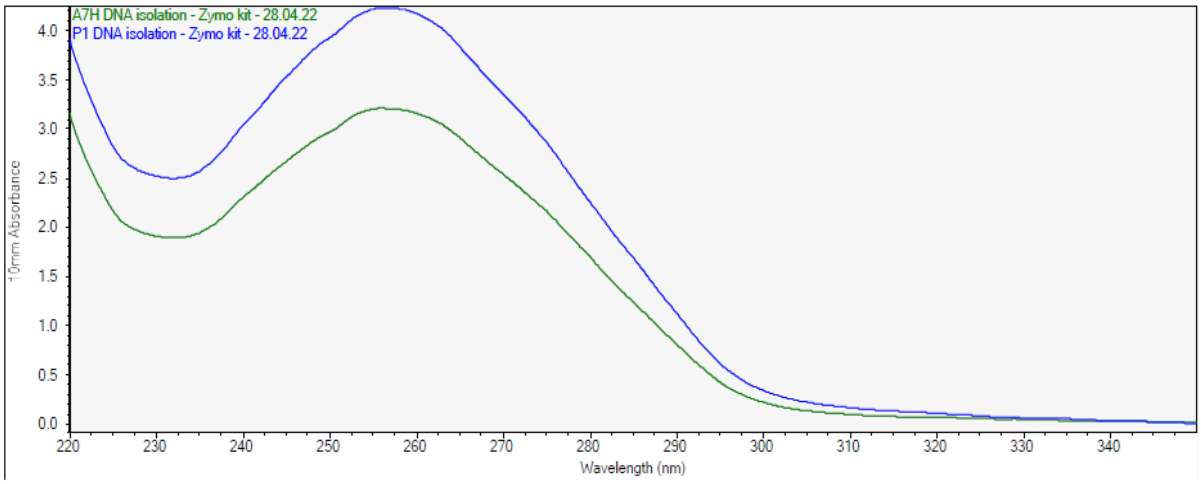

Figure 9. Absorbance spectra of extracted soil DNA from pathogen-infected samples of experiment 3.

Our soil DNA extraction for these samples is the most successful of all the other extractions. Either sample's measurement gives a great DNA concentration, with ideal purity slightly above 1.8 at 260/280 nm ratio and > 1.5 for the 260/230 nm ratio.

2. First step PCR on agarose gel electrophoresis

2.1 Experiment 1

After the DNA extraction, the samples of experiment 1 were prepared for 1st step PCR. Unexpectedly, the results came out negative for all samples. This could be explained by the presence of PCR inhibitors like humic acids and other compounds that inhibit *Taq* polymerase and are co-isolated with the DNA, hence interrupting the amplification process. One way to counteract that was two-fold serial dilution of each extracted stock sample into (1:10), (1:20), (1:40), and (1:80) ratios using Millipore or nuclease-free water. Following each dilution step, the samples would be vortexed well then briefly centrifuged to collect the samples in the bottom of the tubes. The diluted samples were run for PCR then on 1% agarose gel, and the best dilution ratio for each sample was determined judging by the clarity and thickness of the bands.

Starting with the samples Diyarbakır1 and Konya1 to test the serial dilution strategy, each has been diluted into 4 ratios as mentioned above. The resulting PCR products on gel are shown here (figure 10). Only the dilutions Diyarbakır1 (1:80) and Konya1 (1:10) showed positive yet faint, thin DNA bands, whereas no amplification is noted for the other sample dilutions.

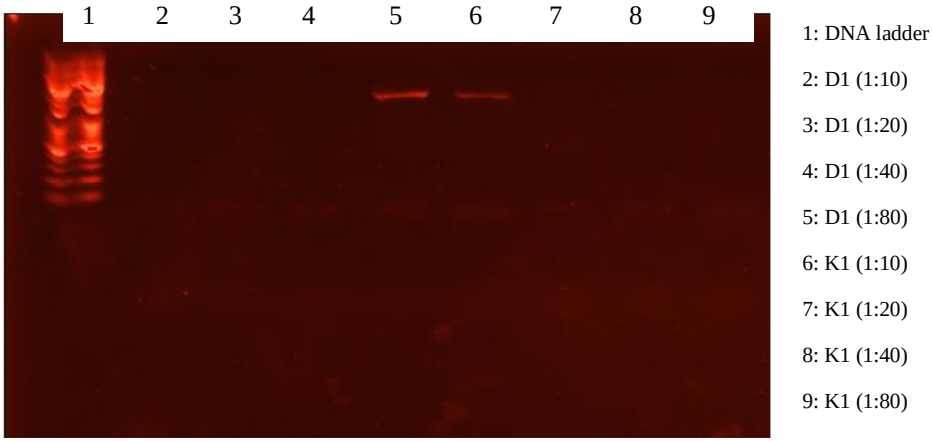

Figure 4. Agarose gel image of diluted D1 and K1 1st PCR products of experiment 1.

The second set of samples to be diluted and prepared for PCR was: Diyarbakır2, Konya2, Tekirdağ1, Tekirdağ2, and Sivas1. Similarly, none of the Diyarbakır2 and Konya2 diluted products showed any DNA bands on the gel (figure 11). However, we have obtained positive results in each of Tekirdağ1 (1:40) and (1:80), Tekirdağ2 (1:10), (1:20) and (1:40), and all of Sivas1 dilutes. The DNA bands are generally thin and faint, though barely seen in Tekirdağ1.

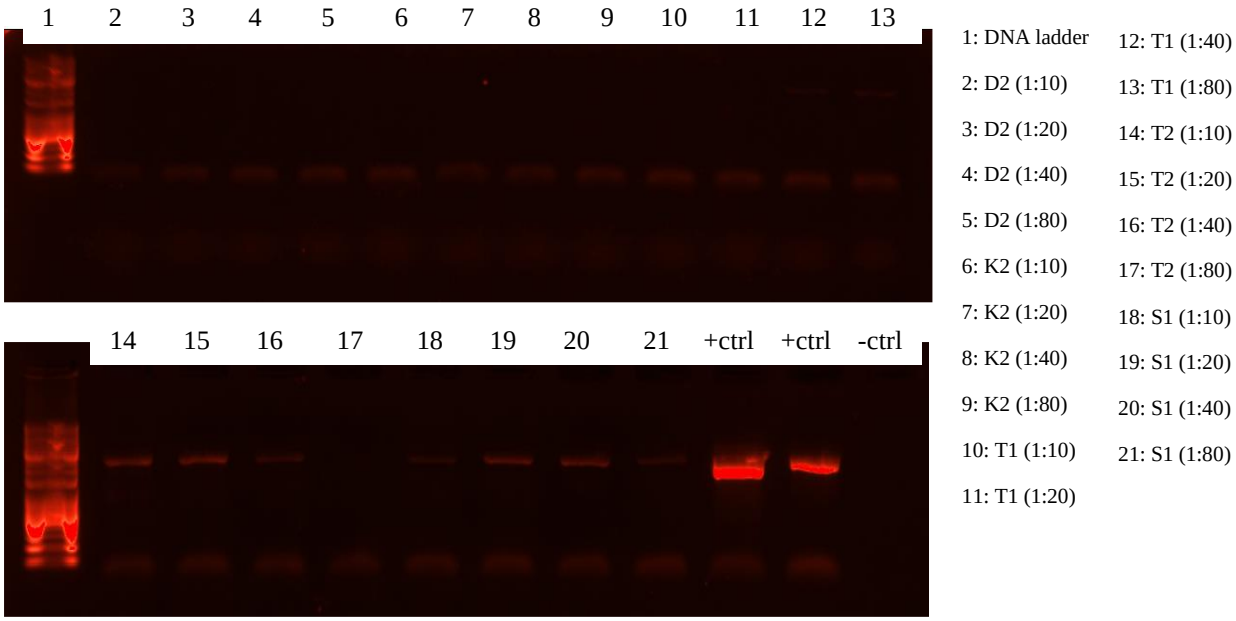

Figure 5. Agarose gel images of the second set of diluted 1st PCR products of experiment 1, using A15H and A5C as positive controls, and a DNA template-free sample as a negative control.

Since the dilution of Diyarbakır2 and Konya2 samples could not succeed in PCR amplification, a second round of two-fold serial dilution was suggested in higher ratios as (1:160), (1:320), and (1:640). In fact, we picked and began the dilution with only Diyarbakır2 (1:80) and Konya2 (1:80) samples given that the more diluted the sample is, the less PCR inhibitors would be. The agarose gel confirms positive PCR products for all the diluted samples, with slightly thicker DNA bands in Diyarbakır2 (figure 12).

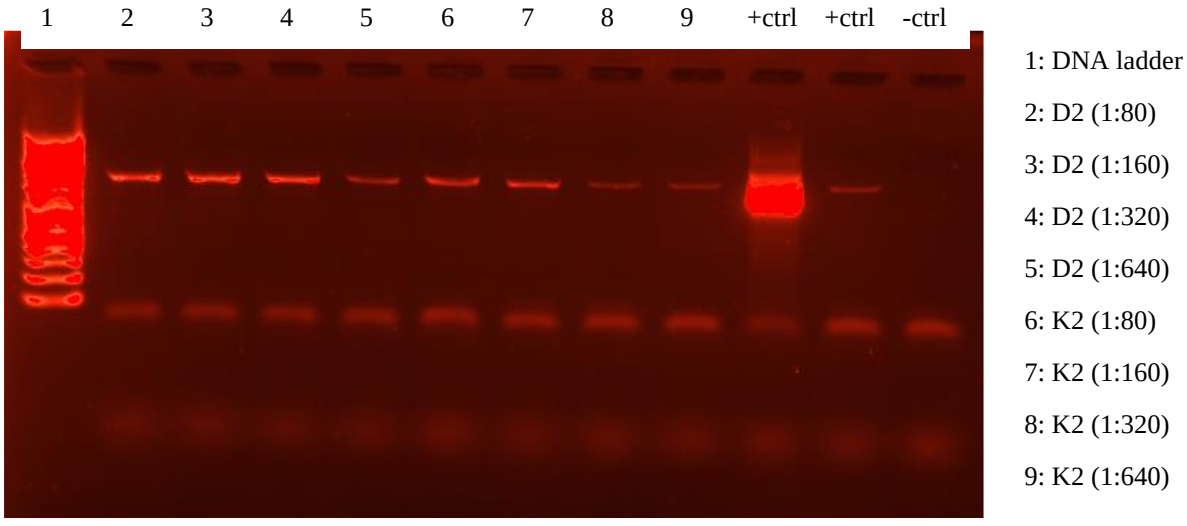

Figure 6. Agarose gel image of second-diluted D2 and K2 1st PCR products of experiment 1.

The following gel image corresponds to the last set of PCR products: Sivas2, Ankara1, Ankara2, Ankara1r, and Tekirdağ2r (figure 13). The DNA bands indicate positive PCR

products in all of Sivas2, Ankara1, and Ankara2 samples. Despite the bad quality image and faint bands, the sample repeats Ankara1r and Tekirdağ2r also showed positive results.

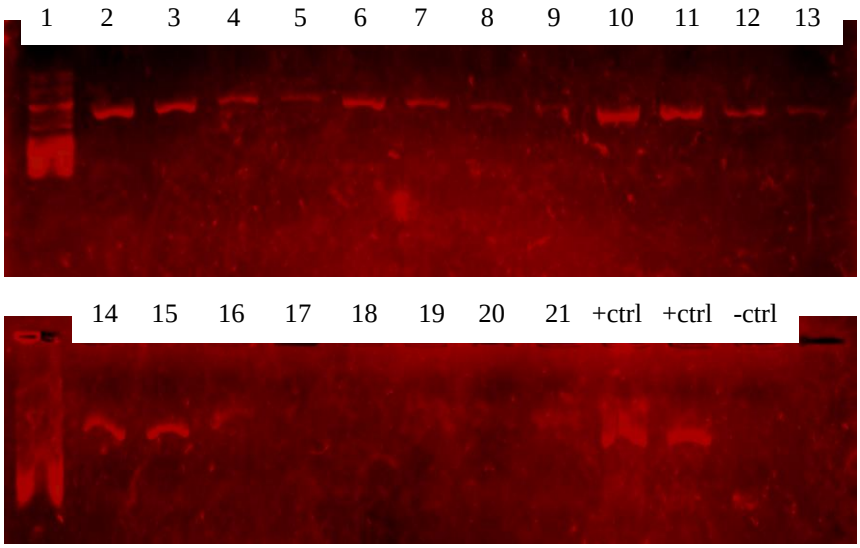

Figure 7. Agarose gel images of the last set of diluted 1st PCR products of experiment 1.

2.2 Experiment 2

After nanodrop measurements, we prepared (1:10) diluted working samples from the fresh stocks that we received from GTU. The samples were set up for 1st step PCR and the results are illustrated in the gel image (figure 14). Unlike the exhaustive results in the experiment 1, this time positive clear PCR products were confirmed for all the samples with both Sivas2 and Tekirdağ2 showing less defined DNA bands. Thus, implying considerably less amount of PCR inhibitors and purer samples that there was no need for further serial dilution of the samples.

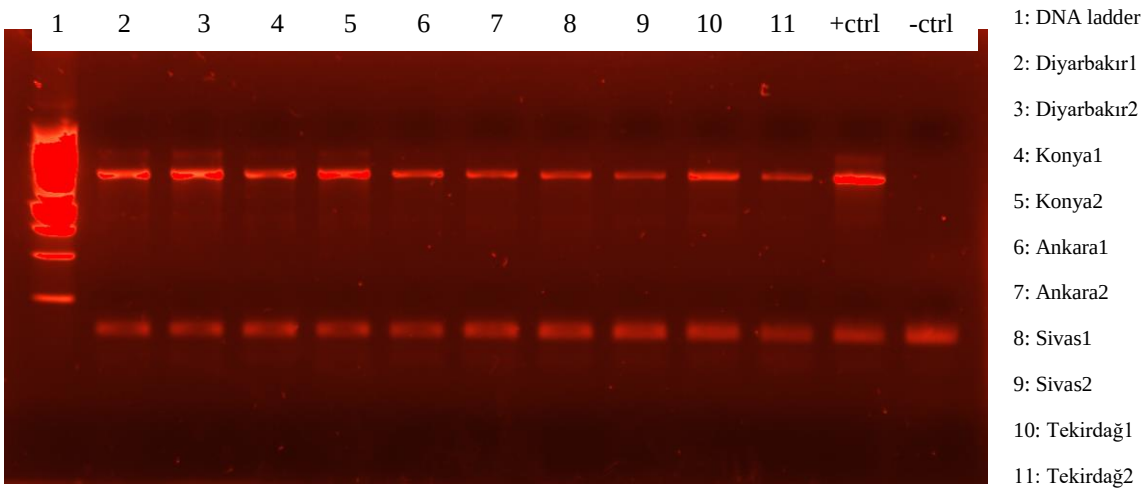

Figure 8. Agarose gel image of (1:10) diluted 1st PCR products of experiment 2.

2.3 Experiment 3

Initially, the isolated samples of P1 and A7H were set directly for 1st step PCR without any dilution step, yet no bands have been seen on the gel. To uncover the reason behind that, 1 µl of each of the DNA sample was run on agarose gel. Interestingly, the gel image (not shown) marked large smears for both samples at the same top level of the ladder, which suggests no major degradation but rather PCR inhibition. To counteract the PCR inhibitors obstacle that was most assumed, both samples were diluted similarly into (1:10), (1:20), (1:40), and (1:80) ratios. As shown below, the samples were positive in all diluted ratios (figure 15). For P1, only the (1:80) sample was the least successful as it appeared faint, whereas the (1:10) sample of A7H was the clearest and most reliable.

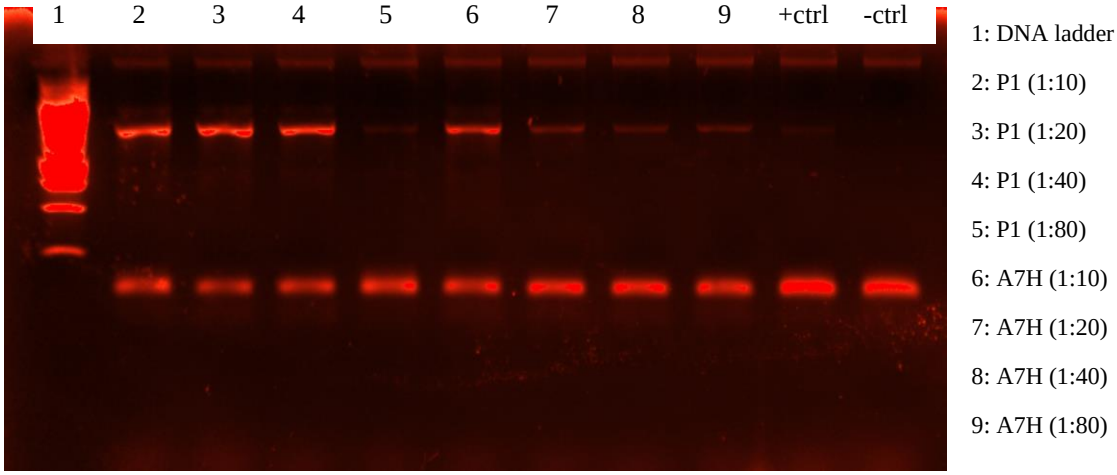

Figure 15. Agarose gel image of diluted P1 and A7H 1st PCR products.

3. First step PCR product purification

The DNA purification results of experiment 2 were not imaged on agarose gel as we ensured its effectiveness and the following PCR barcoding step succeeded anyways. For experiment 1 however, we determined the most successful dilution ratio of each sample and worked them out to be cleaned up with the beads. The list goes as follows: Diyarbakır1(1:80), Diyarbakır2 (1:160), Konya1 (1:10), Konya2 (1:10), Ankara1 (1:10), Ankara2 (1:10), Sivas1 (1:40), Sivas2 (1:10), Tekirdağ1 (1:80), Tekirdağ2 (1:10), Ankara1r (1:20), and Tekirdağ2r (1:20).

Here is the gel image that indicates the successful purification of all the PCR products except for Konya2 (1:10). Generally, the DNA bands looked very thin and sometimes faint, especially Sivas1, Sivas2, and Tekirdağ1 samples (figure 16).

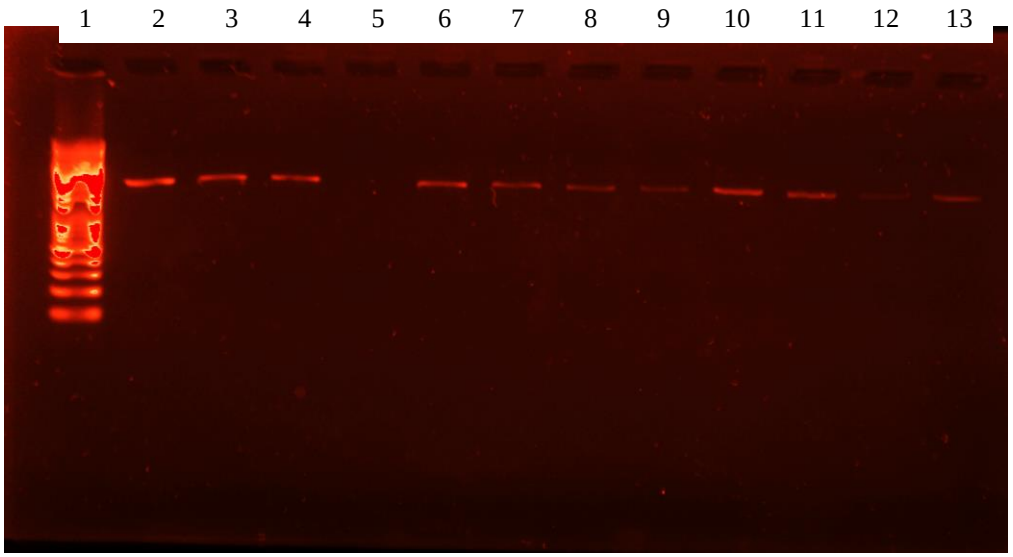

Figure 16. Purification of the best diluted samples 1st PCR products of experiment 1 on gel.

Eventually, the purification of Konya2 sample was repeated yet with the (1:160) and both (1:320) and (1:640) combined together into one diluted sample to increase the volume so the beads would work better. The results were positive for both loaded samples (figure 17), and Konya2 (1:160) was chosen for downstream PCR barcoding since it appeared slightly better than the combined Konya2 (1:320/1:640).

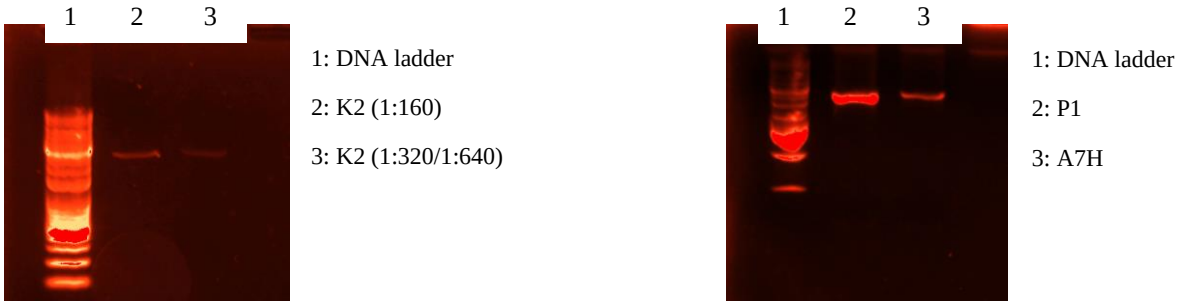

Figure 17. Purified second-diluted K2 of experiment 1, and P1 and A7H of experiment 3, 1st PCR products on gel.

For the purification of P1 and A7H 1st step PCR products of experiment 3, we suggested to pool the three of (1:10), (1:20) and (1:40) diluted samples into one single tube for each. That way we could obtain a higher yield of pure PCR products, and the 0.8x volume of AMPure XP beads used was scaled accordingly. It was obvious to have a nicer DNA band for P1 (figure 15), since A7H PCR results have already come out faint on the gel.

4. PCR barcoding on agarose gel electrophoresis

4.1 Experiment 1

The PCR barcoding for the experiment 1 samples was not very satisfactory since the DNA bands visualized on the agarose gel were somewhat thin and faint, especially for the Sivas and Tekirdağ samples, while the rest of the samples were better recognized (figure 18). Still, the samples were quantified and prepared for sequencing.

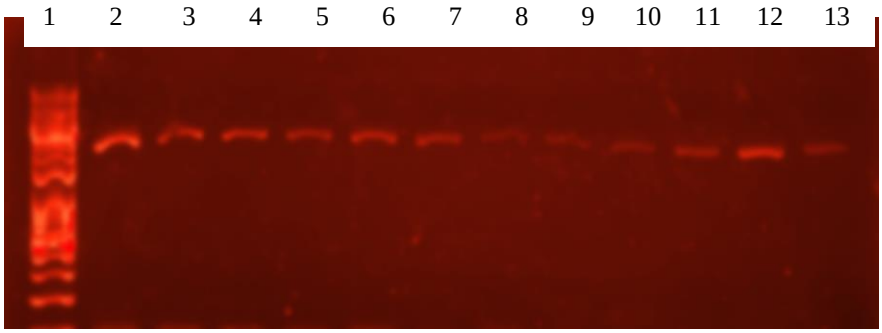

Figure 18. PCR barcoding products of 12 samples of experiment 1 on gel.

4.2 Experiment 2

The purified 1st step PCR products were set up for a 25-cycle PCR barcoding by coupling each sample to a barcoded primer randomly. On a 25-cycle PCR program, the results are as follows: thick, very bright DNA bands were obtained for all the samples except for Diyarbakır1 which seemed quite less (figure 19).

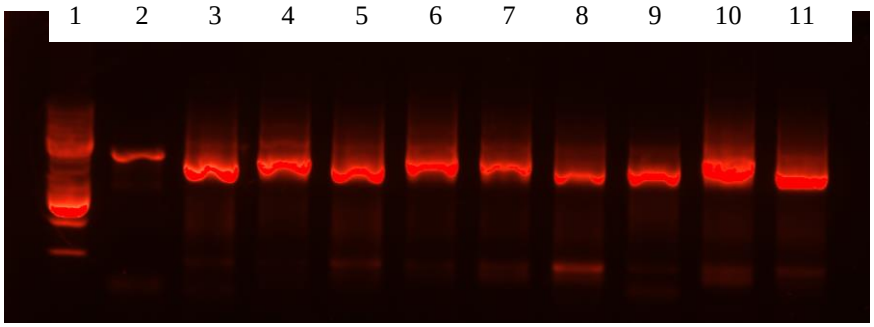

Figure 19. PCR barcoding products of 10 samples of experiment 2 on gel.

4.3 Experiment 3

Likely, P1 and A7H were run on a 25-cycle PCR barcoding program. Successfully, both samples gave positive DNA bands although A7H seemed to amplify a bit less with a thinner band marked on the gel (figure 20).

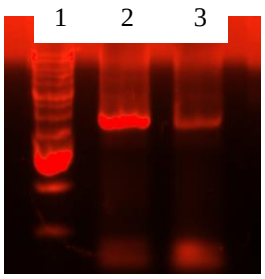

Figure 20. PCR barcoding products of P1 and A7H on gel.

5. Quantification and pooling

5.1 Experiment 1

#	Sample ID	Barcode	Quant-It conc. (ng/µl)	Molar conc. (*f*mol/µl)	Pooled volume for library (µl)
1	Diyarbakır1	BP01	21	22.7	0.53
2	Diyarbakır2	BP02	8.8	9.50	1.26
3	Konya1	BP03	8.8	9.50	1.26
4	Konya2	BP04	5,8	6.26	1.92
5	Ankara1	BP05	8.9	9.61	1.25
6	Ankara2	BP06	5.9	6.37	1.88
7	Sivas1	BP07	3.22	3.48	3.45
8	Sivas2	BP08	2.29	2.47	4.86
9	Tekirdağ1	BP09	4.57	4.94	2.43
10	Tekirdağ2	BP10	4.62	4.99	2.40
11	Ankara1r	BP11	13	14.04	0.85
12	Tekirdağ2r	BP12	4.60	4.97	2.41

Table 1. DNA quantification and pooling of 12 samples of experiment 1 using Quant-It HS assay.

The DNA concentrations obtained from the Quant-It HS assay for the first sequencing run were low compared to the other two experiments that are to be listed next (table 1). Thus, in order to combine an equimolar amount of each barcoded PCR product into aDNA library, 12 *f*mol of each sample was pooled according to the calculations already explained in the methods' section. The resulting total volume of the library was 24.5 µl containing a total of 144 fmol DNA and therefore when 10 µl of this library was used for sequencing, it contained (144 x 10/24.5) =58.7 *f*mol of barcoded DNA, within the desired ratio as in the protocol.

5.2 Experiment 2

#	Sample ID	Barcode	Quant-It conc. (ng/µl)	Molar conc. (*f*mol/µl)	Pooled volume for library (µl)
1	Diyarbakır1	BP03	49	52.92	0.38
2	Diyarbakır2	BP06	91	98.28	0.20
3	Konya1	BP01	81	87.48	0.23
4	Konya2	BP10	73	78.84	0.25
5	Ankara1	BP08	81	87.48	0.23
6	Ankara2	BP02	74	79.92	0.25
7	Sivas1	BP09	33	35.64	0.56
8	Sivas2	BP07	78	84.24	0.24
9	Tekirdağ1	BP05	81	87.48	0.23

10	Tekirdağ2	BP04	91	98.28	0.20

Table 2. DNA quantification and pooling of 10 samples of experiment 2 using Quant-It HS assay.

In experiment 2, we obtained remarkably better DNA concentrations for all the samples (table 2) ranging above 70 ng/µl, except for Sivas1 which recorded the lowest amount of 33 ng/µl only and 49 ng/µl of DNA in Diyarbakır1. On the other hand, both Diyarbakır2 and Tekirdağ2 contained the highest DNA concentration (91 ng/µl). Therefore, we increased the quantity of each barcoded PCR product in the pooled DNA library up to 20 *f*mol. The final total volume of the library was 2.77 µl containing 73.5 *f*mol, that is 26.53 *f*mol/µl.

5.3 Experiment 3

#	Sample ID	Barcode	Quant-It conc. (ng/µl)	Molar conc. (*f*mol/µl)	Pooled volume for library (µl)
1	Diyarbakır1	BP03	49	52.92	0.92
2	Diyarbakır2	BP06	91	98.28	0.50
3	Konya1	BP01	81	87.48	0.56
4	Konya2	BP10	73	78.84	0.62
5	Ankara1	BP08	81	87.48	0.56
6	Ankara2	BP02	74	79.92	0.61
7	Sivas1	BP09	33	35.64	1.37
8	Sivas2	BP07	78	84.24	0.58
9	Tekirdağ1	BP05	81	87.48	0.56
10	Tekirdağ2	BP04	91	98.28	0.50
11	P1	BP11	57	61.56	0.79
12	A7H	BP12	28	30.24	1.60

Table 3. DNA quantification and pooling of 10 wheat soil samples along with P1 and A7H of experiment 3 using Quant-It HS assay.

For the last sequencing run, the pathogen-infected soil samples P1 and A7H were quantified then pooled together with the 10 wheat soil samples into one adjusted DNA library within the desired ratio (table 3). In fact, P1 has got a relatively good DNA concentration (57 ng/µl) thanks to the idea of pooling the 3 different dilutes into one sample as previously mentioned, whereas A7H has marked the very lowest amount out of the bunch (28 ng/µl) that was expected from the thin PCR product band on the gel (figure 12). This time, we intended to increase the molarity of the DNA library the most, judging by the most and least concentrated samples. For instance, taking 0.5 µl from the most concentrated sample D2 or T2 (98.28 *f*mol/µl) makes 49.14 *f*mol per sample. That is, 49.14 *f*mol of the least concentrated sample A7H (30.24 *f*mol/µl) equals 1.6 µl. Given that we obtained more than only 1.6 µl of A7H,

we agreed on 49.14 *f*mol molarity of the samples. The added volumes were calculated per each. The pooled library had 9.17 µl total volume, 589.68 *f*mol in total, that is 64.30 *f*mol/µl. The increased molarity of the library after suspension in 10 µl mM Tris.HCl pH 8.0 with 50mM NaCl buffer was adjusted right before sequencing by further diluting it with the buffer in (1:5) ratio, in order to bring it down to the recommended range.

6. Data analysis

#	Reads analyzed	Reads classified	Reads unclassified	Avg. sequence length
Exp1	372,904	267,814	5,050	1,348 bases
Exp2	1,235,401	666,412	38,491	856 bases
Exp3	2,948,000	1,282,403	47,712	723 bases

Table 4. Epi2ME outputs of the 3 experiments with regard to read classification and average sequence length.

The overall data outputs generated by Epi2ME workflow reports across all of the three experiments are summarized above (table 4). Generally, the table indicates that our sequencing experiment has become more successful upon practice. The largest generated dataset corresponds to the experiment 3 with a value of 2,948,000 total reads analyzed (that is roughly 8 times the number of total reads analyzed in experiment 1).

The reads were classified according to the QC filters (Qscore >7, read length >500 bp), so 96% of them have been classified and 4% were unclassified (figure 21 C). Though, the average sequence length in this dataset has interestingly dropped to 723 bases only compared to the expected average sequence length as reported in experiment 1 (about 1.35 kb) (table 4). Similarly, the number of total reads analyzed in experiment 2 accounts for 3 times that of in experiment 1 (1,235,401 total reads analyzed), 95% of which are classified reads, again with a decreased average sequence length at 856 bases (table 4, figure 21 B). While only 2% of the total reads were unclassified in experiment 1, that is relatively not surprising given that only 372,904 total reads could be generated and analyzed (table 4, figure 21 A).

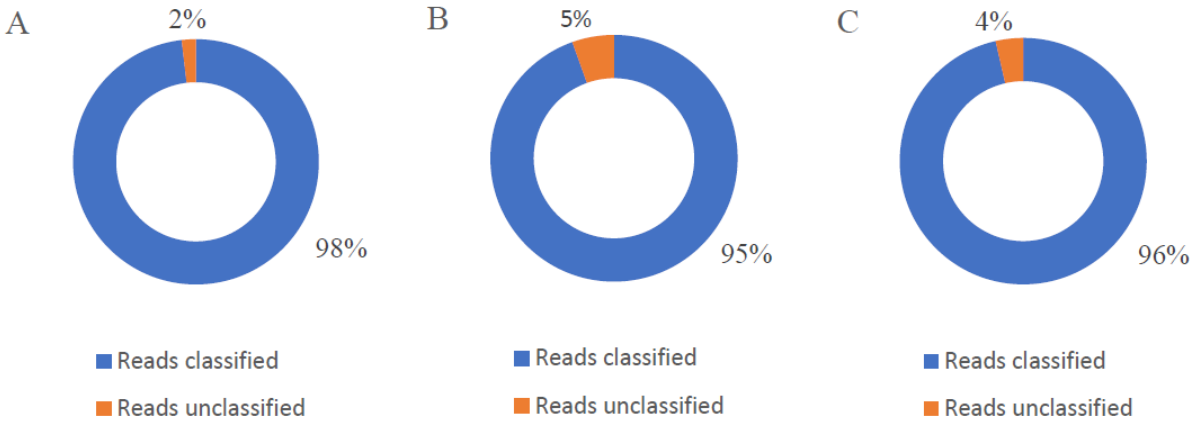

Figure 21. Comparative illustration of the total read classification by percentage across the three experiments. A: Experiment 1; B: Experiment 2; C: Experiment 3.

The total number of classified reads per each sample was checked as well for all 3 experiments (table 5). In the experiment 1, the highest number of classified reads corresponds to Diyarbakır 2 (39,143 reads), while the lowest number of classified reads corresponds to Tekirdağ 1 (13,492 reads). In the experiment 2, Sivas 1 indicates the highest number of classified reads with 99,465 reads, and interestingly Tekirdağ 1 marks the lowest output wıth 23,013 classıfied reads only. In the experiment 3, all the samples demonstrate a high output of classified reads that is the highest in P1 (181,066 reads). Except for Diyarbakır 1 which shows 37,716 reads only.

Sample ID	**Experiment 1**	**Experiment 2**	**Experiment 3**
Diyarbakır 1	19,591	70,269	37,716
Diyarbakır 2	39,143	71,427	180,862
Konya 1	23,114	49,215	77,712
Konya 2	19,242	70,915	121,383
Ankara 1	18,159	52,512	119,303
Ankara 2	16,204	80,497	78,776
Sivas 1	16,363	99,465	76,401
Sivas 2	19,117	57,745	54,884
Tekirdağ 1	13,492	23,013	132,996
Tekirdağ 2	25,061	92,343	100,715
Ankara 1r	25,757	Not tested	Not tested
Tekirdağ 2r	27,689	Not tested	Not tested
P1	Not tested	Not tested	181,066

A7H	Not tested	Not tested	110,489

Table 5. Total number of classified reads per soil sample for all 3 experiments.

6.1 Experiment 1

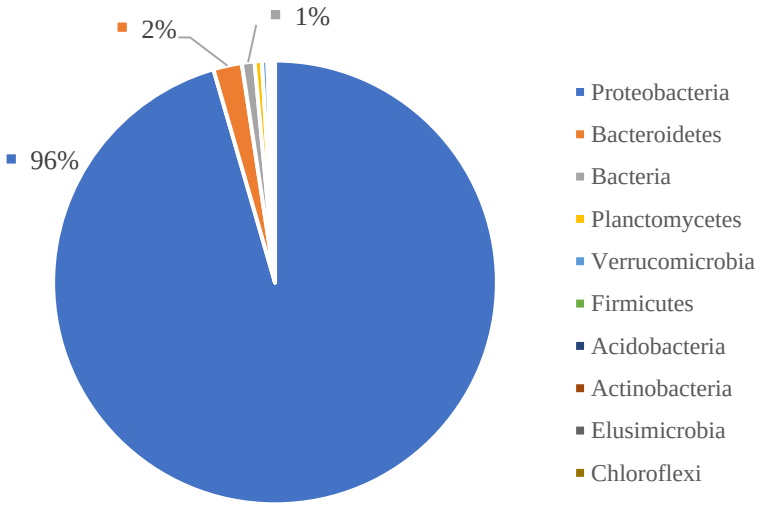

Figure 22. Classification of top 10 phyla derived from 16S rRNA bacterial clone sequences of experiment 1 in all samples. The percentage corresponds to the total reads classified per phylum.

The generated dataset of experiment 1 had the size of 503 Mbases total yield only. Primarily, Phinch2 has helped with the several data visualization and analysis. In the first place, the top 10 phyla derived from the 16S rRNA bacterial clone sequences out of all the samples were determined (figure 22). Proteobacteria constitute 96% of the total reads. Second come Bacteroidetes with only 2% classification, then 1% Bacteria, and the rest of the classified phyla accounts for no more than 1% in total. These include Planctomycetes, Verrucomicrobia, Firmicutes, Acidobacteria, Actinobacteria, Elusimicrobia, and Chloroflexi.

At the genus level, the top 10 taxa dominantly ranked by total sequence reads were also visualized on the bar chart above (figure 23). Eventually, Betaproteobacteria ranks first in the dominant taxon list, and other common names are included as well, such as *Massilia*, *Ramlibacter*, *Pseudomonas, Lysobacter*, and *Variovorax*. Though, the distribution of these genus varies in read counts per sample as the bar chart demonstrates. For instance, Betaproteobacteria roughly accounts for only 5% of the sequence reads in Diyarbakır1 and Diyarbakır2 samples where *Ramlibacter*, *Massilia*, Oxalobacteraceae, and Comamonadaceae, which constitute 10% or more each. Similarly in Tekirdağ2 and its sample repeat Tekirdağ2r, *Pseudomonas* ranks the most dominant bacterial genus with almost 11% of the total read counts in each sample, and the appearance of *Lysobacter* as the third most dominant genus in both of these samples compared to the other samples is also marked.

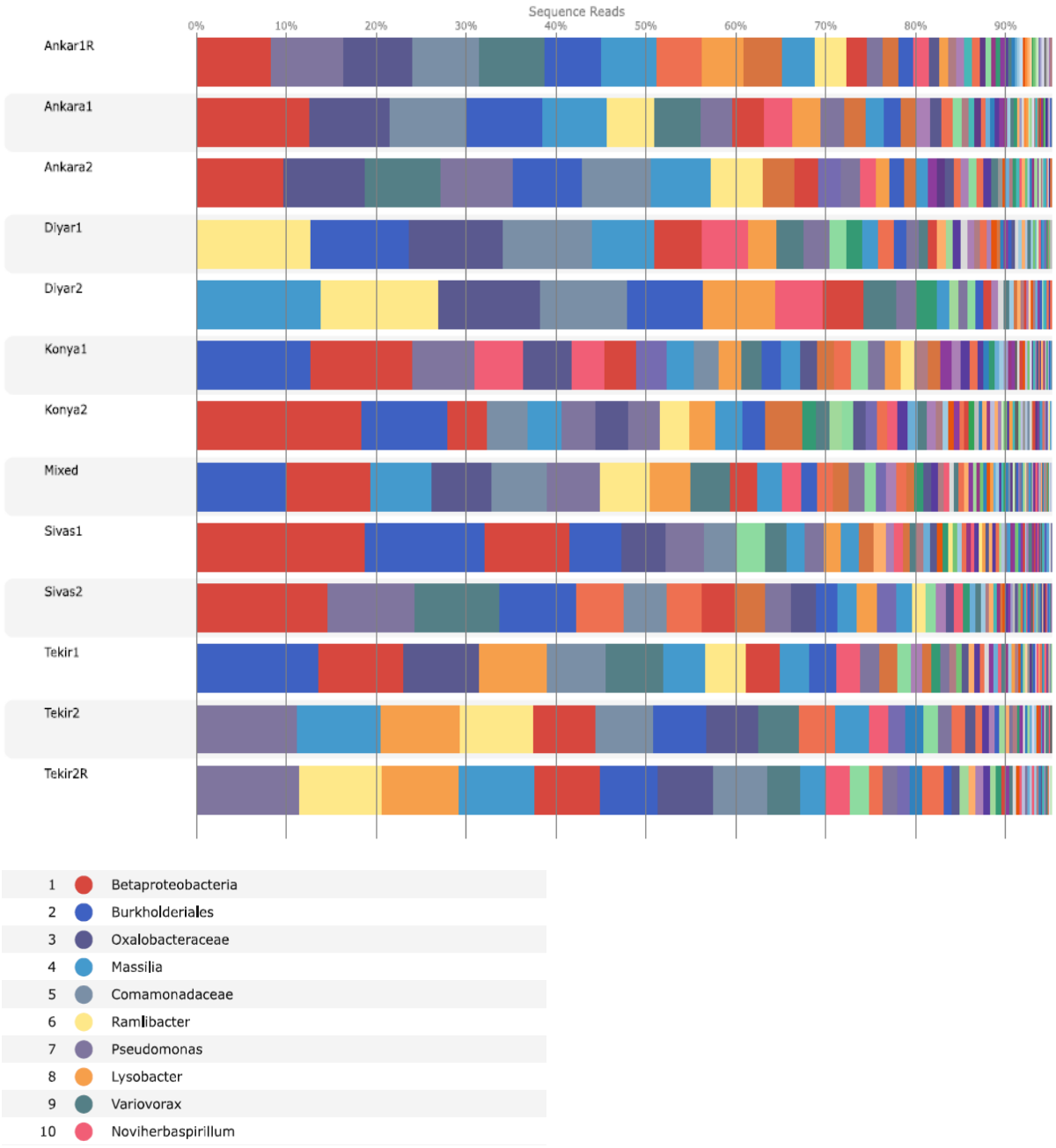

Figure 23. Bacterial genus distribution in wheat soil samples of experiment 1 with the 10 most dominant genus by total sequence reads, generated with Phinch2 (bioRxiv 009944; https://doi.org/10.1101/009944).

At the level of species, the 10 most abundant species overall have been determined (figure 24) as follows: *Ramlibacter monticola* (11,382 total reads), *Variovorax paradoxus* (6,669 total reads), *Lysobacter terricola* (5,469 total reads), *Pseudomonas orientalis* (5,334 total reads), *Ramlibacter ginsenosidimutans* (3,358 total reads), *Variovorax ginsengisoli* (2,777 total reads), *Massilia putida* (2,687 total reads), *Massilia atriviolacea* (2,630 total reads), *Nitrosospira multiformis* (2,553 total reads), and *Pseudomonas kilonensis* (2,549 total reads).

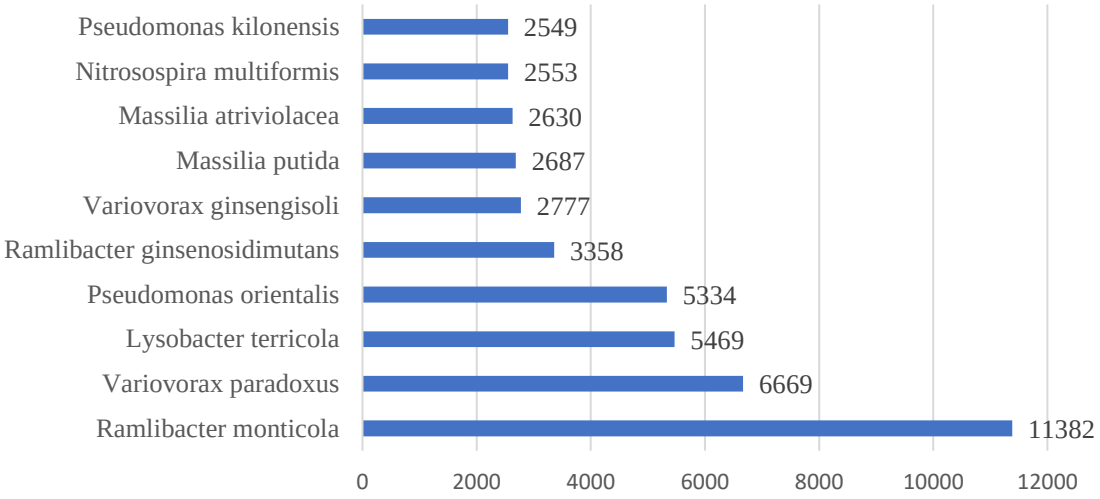

Figure 24. Classification of the 10 most abundant bacterial species in experiment 1 by total read count.

Distribution of these most dominant bacterial species of experiment 1 are illustrated in the graph below (figure 25). As expected from the bacterial genus distribution graph (figure 23), Sivas1 is a poor sample in terms of bacterial variation relatively to the total number of classified reads across the 3 experiments (see table 5), as such the most abundant species *Ramlibacter monticola* does not appear (1.83 relative abundance) (figure 23). While the largest abundance for *Ramlibacter monticola* is marked in Diyarbakır2 (90.10 relative abundance).

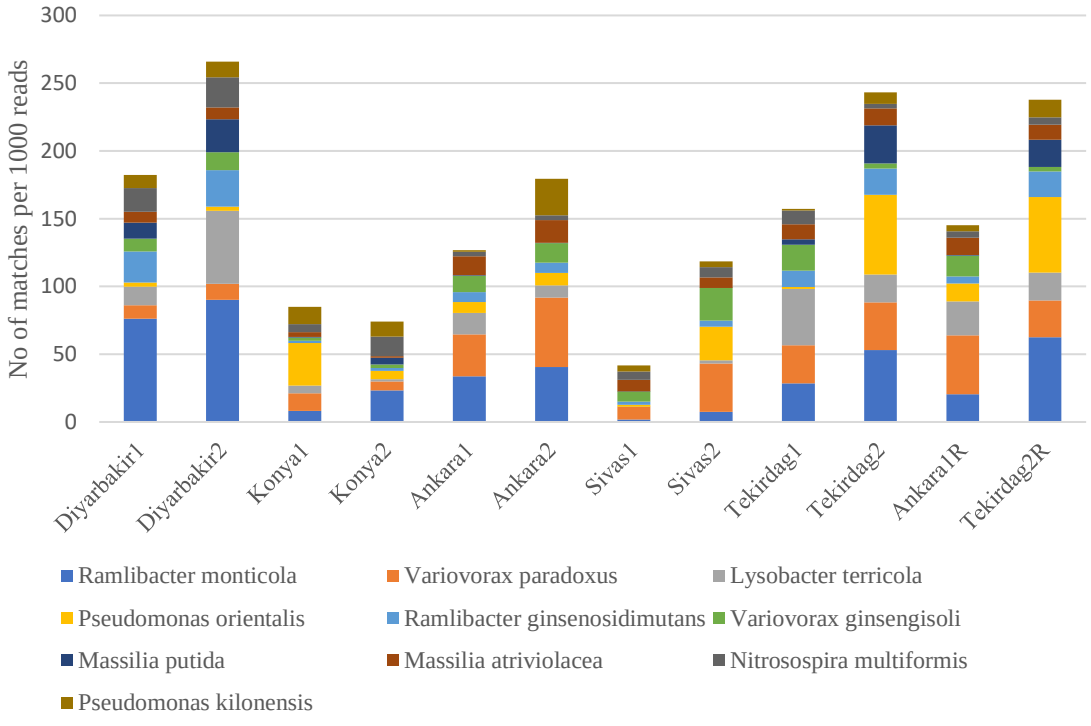

Figure 25. Distribution of the 10 most abundant bacterial species, experiment 1, across the 12 wheat soil samples by relative abundance.

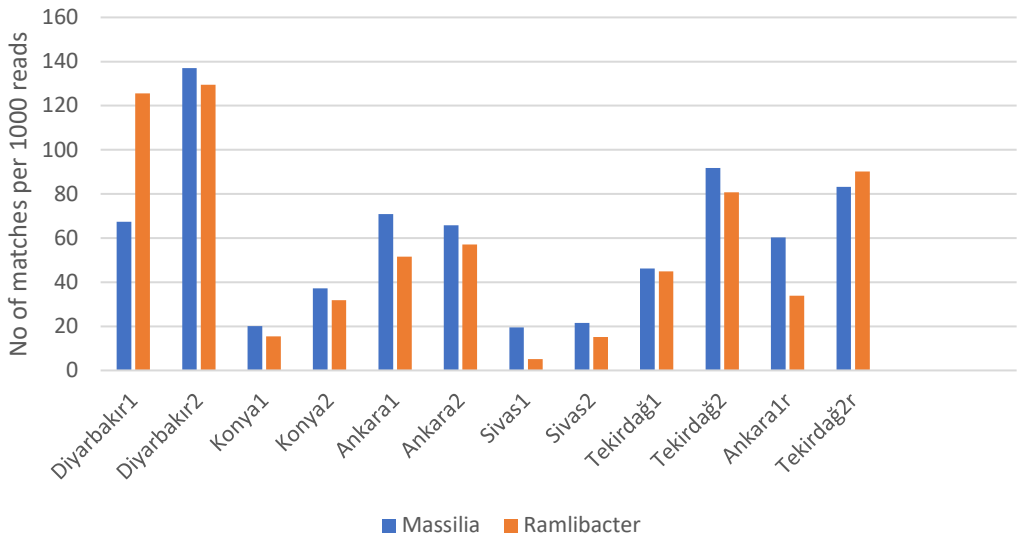

Figure 26. Relative abundance of community bacterial genus *Massilia* and *Ramlibacter*, experiment 1, in wheat soil samples.

Another significant thing is that some bacterial taxa tend to form community genus or species. Specifically, *Massilia* and *Ramlibacter* bacterial genus have shown to obtain relatively similar pattern of appearance in all samples. As demonstrated in (figure 26), the relative abundance of *Ramlibacter* increases with the relative abundanceof *Massilia* in almost all the samples simultaneously and decreases again with that of *Massilia*. Except for the Diyarbakır1 and Ankara1r samples where the relative abundance for each genus did not keep that constant pattern of distribution.

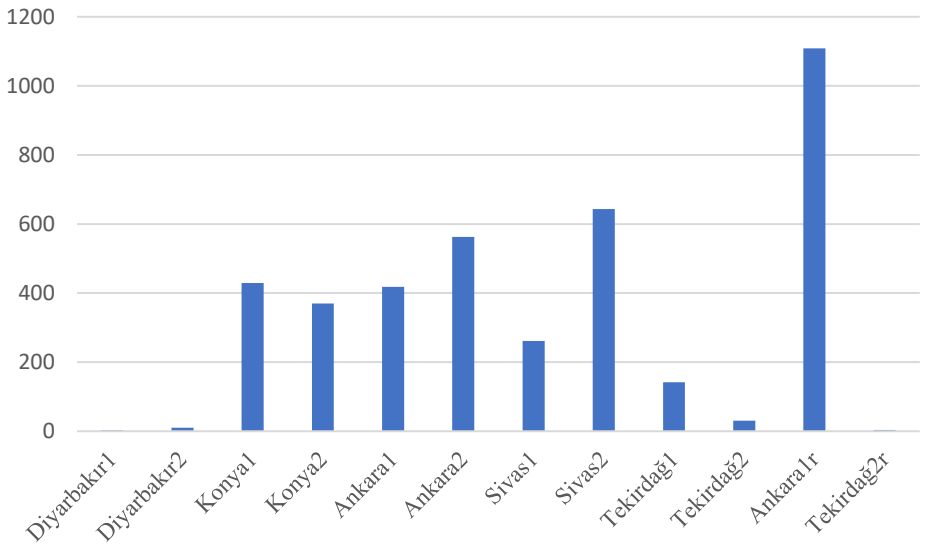

Figure 27. Selective distribution and abundance of bacterial genera *Pseudoxanthomonas*, experiment 1, across the wheat soil samples by read count.

Other bacterial genera seem to have selective pattern of distribution across samples of different and distinct regions (figure 27). *Pseudoxanthomonas* was considered to have most relative abundance in Ankara1r and Sivas2 samples (1,109 and 643 reads, respectively). As well as in Ankara 2 with 563 reads, and less variable in the other samples. Whereas almost no occurrence of *Pseudoxanthomonas* was detected in Diyarbakır1 that got barely 3 reads only, Tekirdağ2r with 4 reads only, Diyarbakır2 with 10 reads, and Tekirdağ2 with 31 reads.

On the other hand, we were keen to look at the sample repeats and compare their genus distribution, to define whether or not their consistency is relative. Starting with Ankara1 and Ankara1r (figure 28), it is surprising that the distribution of the most dominant bacterial genus looks quite variable comparatively in both samples. For example, the read count of *Variovorax*, *Lysobacter*, *Pseudoxanthomonas*, and *Xanthomonadaceae* in Ankara1r is 2 times that of Ankara1. Also, *Pseudomonas* is present 3 times more in Ankara1r than in Ankara1, while *Stenotrophomonas* prevails almost 5 times more in the sample repeat than in Ankara1. That is by considering that the total read count we obtained in Ankara1r is relatively higher (25,757 total reads) than that of Ankara1 (18,159 total reads).

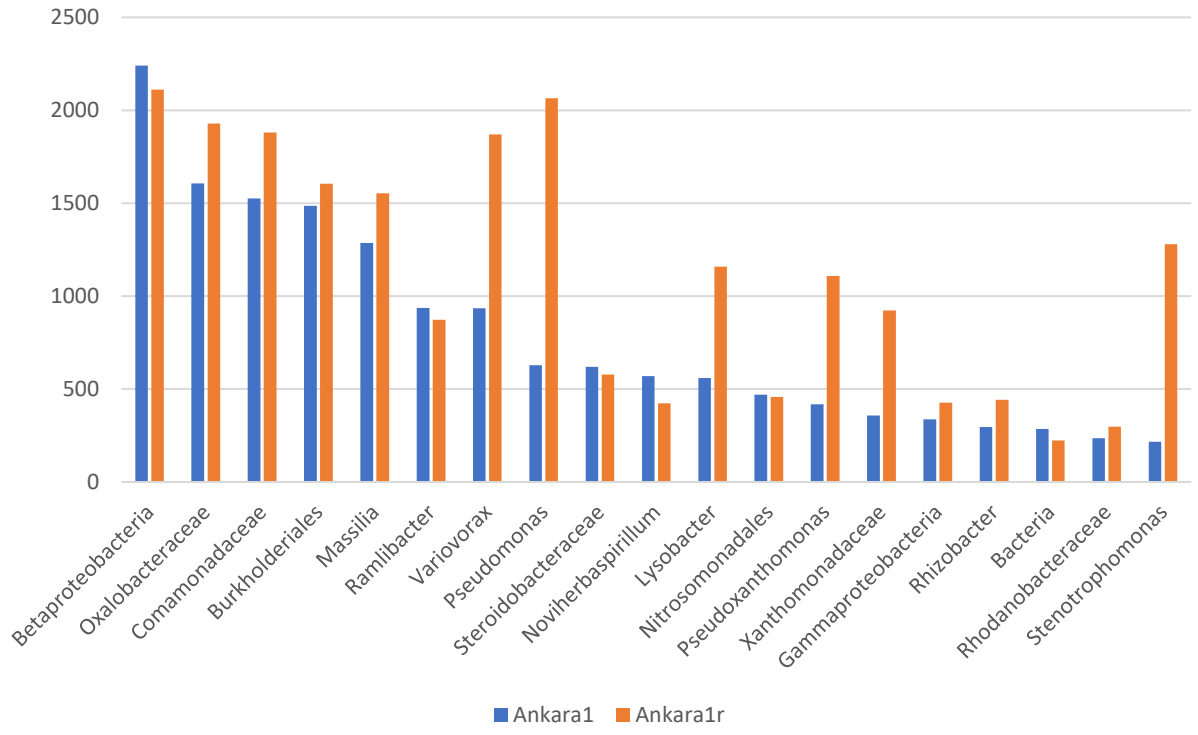

Figure 28. Comparative distribution of the most dominant genus of experiment 1 in sample repeats Ankara1 and Ankara1r, by read count.

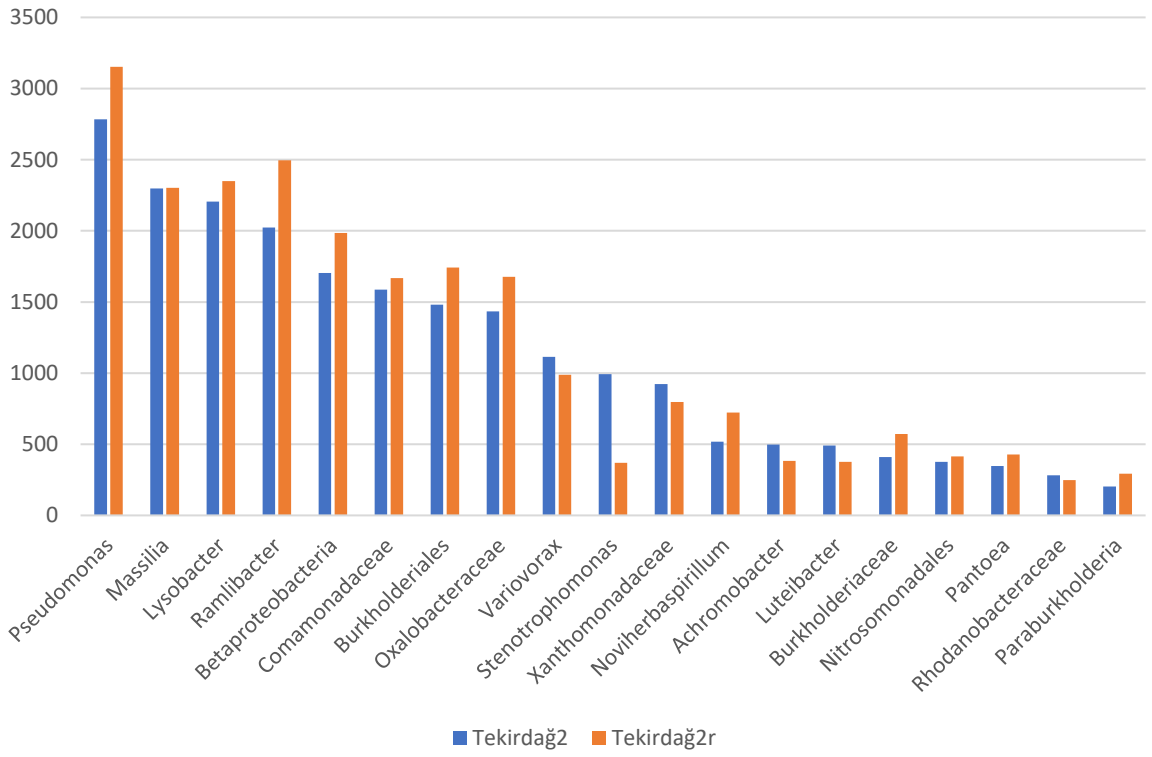

Figure 29. Comparative distribution of the most common genus of experiment 1 in sample repeats Tekirdağ2 and Tekirdağ2r, by read count.

That is not the case with the sample repeats Tekirdağ2 and Tekirdağ2r, as the read count histogram indicates a consistant pattern of genus distribution at the first sight (figure 29). Indeed, the read count for most of the listed predominant bacterial genus in Tekirdağ2r tends to be equal (like *Massilia*, Comamonadaceae, and Nitrosomonadales), slightly higher (like *Pseudomonas*, *Lysobacter*, *Ramlibacter*, Betaproteobacteria, Burkholderiales, Oxalobacteraceae, Noviherbaspirillum, Burkholderiaceae, Pantoea, and Paraburkholderia) or slightly lower (like *Variovorax*, *Xanthomonadaceae*, *Achromobacter*, and *Luteibacter*) than that in Tekirdağ2r. With the exception of Stenotrophomonas which appears 2 times more in Tekirdağ2 than in the sample repeat. This overall consistency pattern is expected when considering the total read count in Tekirdağ2 of 25,061 reads that is quite close to that of Tekirdağ2r with 27,689 total reads.

6.2 Experiment 2

Here the generated dataset was quite larger than the first (1.1 Gbases total yield). As observed in Experiment 1, 98% of the total reads classified in all samples correspond to Proteobacteria, and there is only 1% Bacteroidetes and 1% Planctomycetes. While the rest of the 10 most abundant phyla in experiment 2 are as follows: Verrucomicrobia, Firmicutes, Actinobacteria, Acidobacteria, Elusimicrobia, Chloroflexi, and Gemmatimonadetes (figure 30).

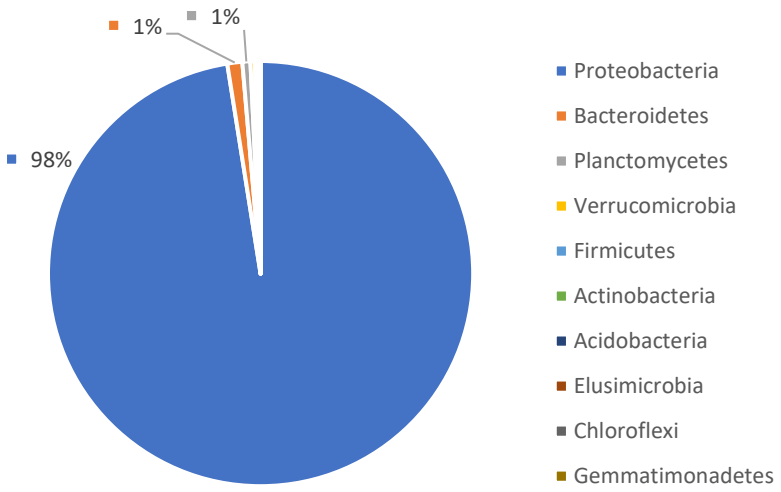

Figure 30. Classification of top 10 phyla derived from 16S rRNA bacterial clone sequences of experiment 2 in all samples. The percentage corresponds to the total reads classified per phylum.

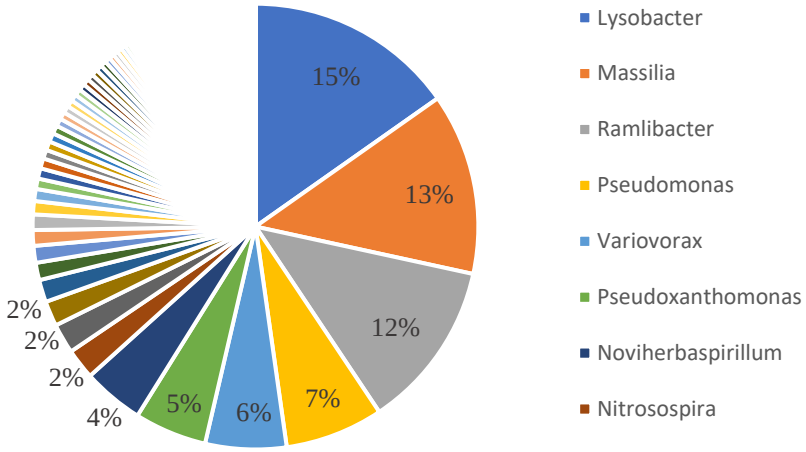

Figure 31. Classification of the 10 most abundant genera of experiment 2. The percentage represents the total read count out of all 10 samples.

These phyla are then classified into the 10 most abundant genera of the experiment 2 by percentage of the total sequence reads in all the samples (figure 31). The ranking list goes as follows: 15% *Lysobacter*, 13% *Massilia*, 12% *Ramlibacter*, 7% *Pseudomonas*, 6% *Variovorax*, 5% *Pseudoxanthomonas*, 4% *Noviherbaspirillum*, 2% *Nitrosospira*, 2% *Arenimonas*, and 2% *Achromobacter*. Considering that most of these already derive from the Proteobacteria phylum or genus that is shown in the genus distribution of the other two experiments in the color graphs; hence no significant difference is seen in the genus classification results.

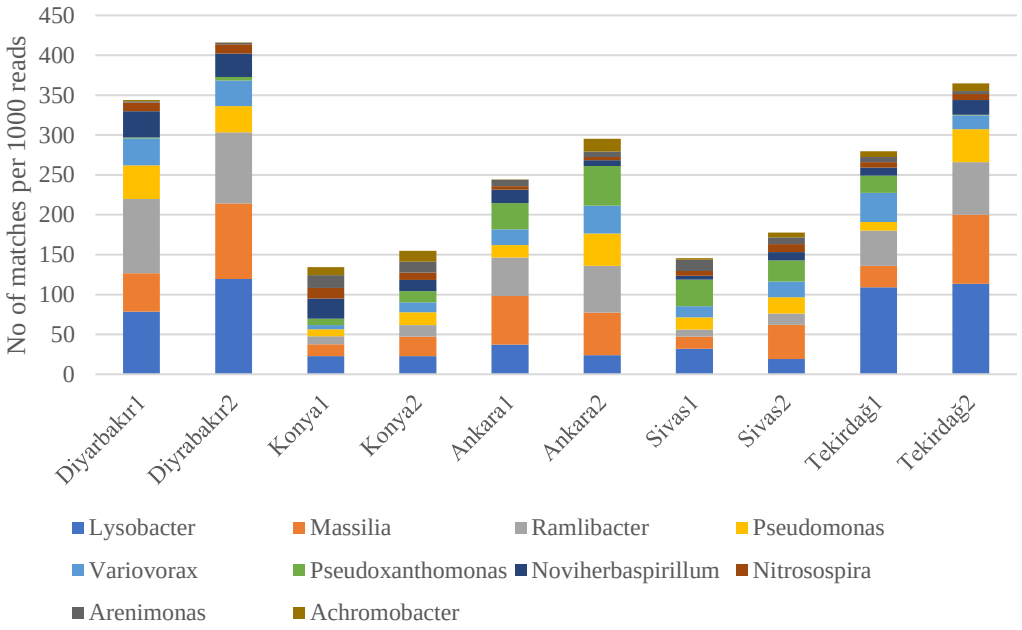

Figure 32. Distribution of the 10 most dominant bacterial genera, experiment 2, across the wheat soil samples by relative abundance.

Distribution of these abundant bacterial genera is demonstrated (figure 32). Shortly, it is obvious that the three of *Lysobacter*, *Massilia*, and *Ramlibacter* genera do share a co-occurrence pattern of distribution in almost all the samples, except for Sivas1 and Sivas2, referring to the observation of community bacteria. Besides, *Pseudoxanthomonas* seem to be barely present in Diyarbakır1, Diyarbakır2, and Tekirdağ2, relatively to their increased total classified reads. Rather, it is most abundant in Ankara2 (49.7 relative abundance) and Sivas1 (33.5 relative abundance), relatively in Ankara1 (32.9 relative abundance) and Sivas2 (26.7 relative abundance) as well. Something that also suggested a selective abundance pattern of *Pseudoxanthomonas* with regards to each sample (figure 33).

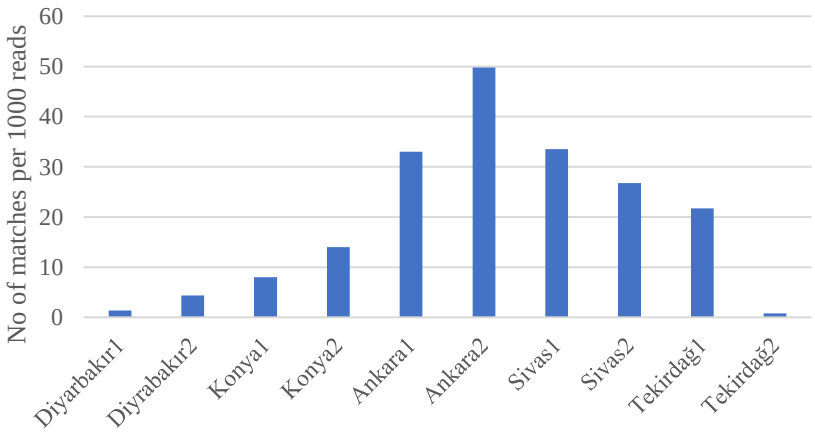

Figure 33. Selective abundance of the genus *Pseudoxanthomonas*, experiment 2, across the wheat soil samples by relative abundance.

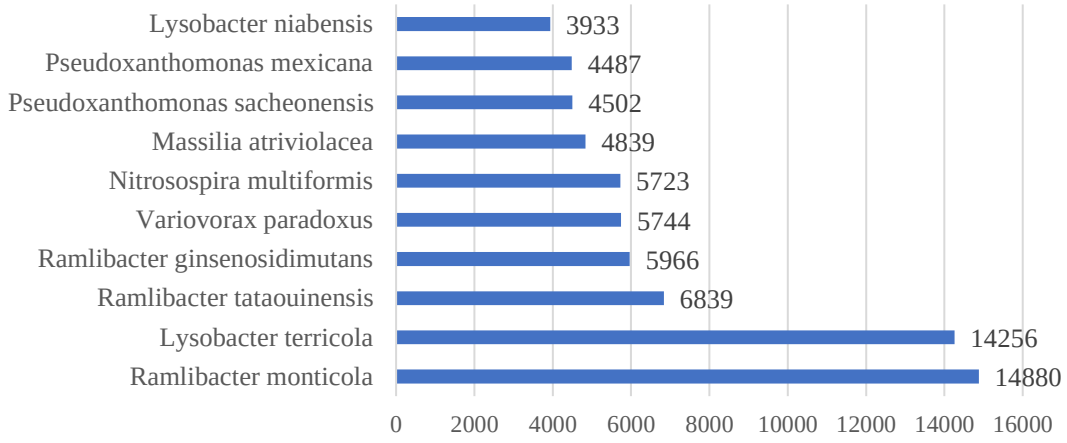

Figure 34. Classification of the 10 most abundant bacterial species in experiment 2 by total read count.

Though, there is quite a variation in the classification by species in experiment 2 compared to the previous results in experiment 1 (figure 34). For example, *Lysobacter terricola* takes over as the second most abundant specie (14,256 total reads), after *Ramlibacter monticola* (14,880 total reads), followed by two Ramlibacter species, namely *Ramlibacter tataouinensis* (6,839 total reads) and *Ramlibacter ginsenosidimutans* (5,966 total reads). *Variovorax paradoxus* appears next (9,171 total reads), then *Nitrosospira multiformis* (5,723 total reads), *Massilia atriviolaceae* (4,839 total reads), *Pseudoxanthomonas sacheonensis* (4,502 total reads), *Pseudoxanthomonas mexicana* (4,487 total reads), and *Lysobacter niabensis* (3,933 total reads).

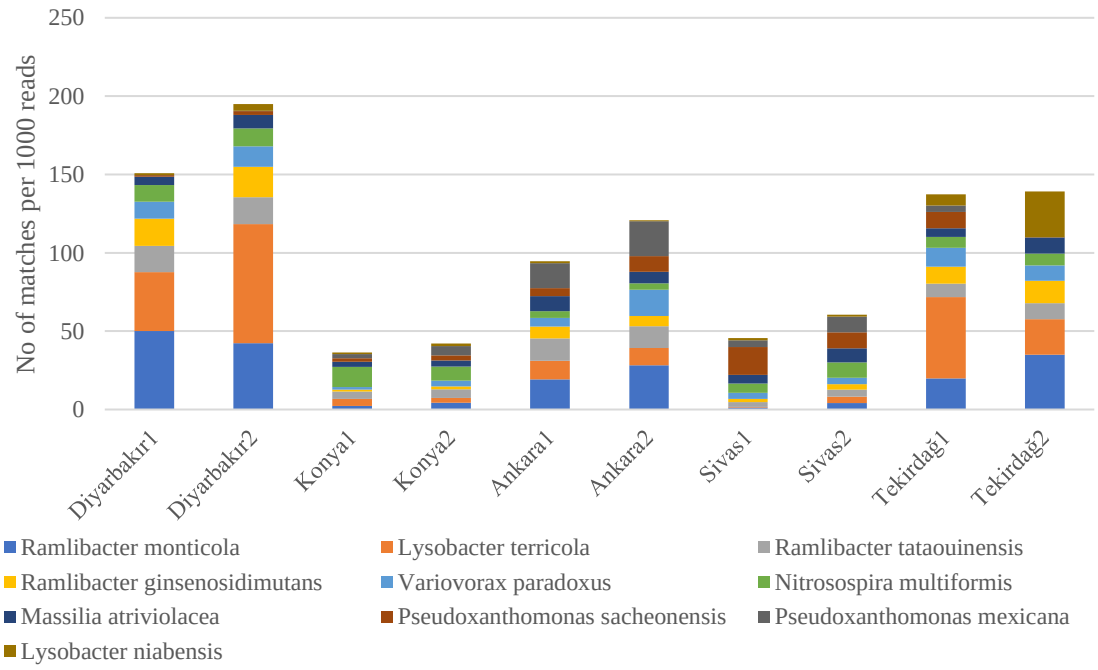

Figure 35. Distribution of the 10 most dominant bacterial species, experiment 2, across the wheat soil samples by relative abundance.

By analyzing the bacterial species distribution here, there seem a very distinct occurrence pattern for each across the 10 wheat soil samples (figure 35). For instance, *Ramlibacter monticola* is abundant at relatively constant read count in Diyarbakır1 (3,525 reads), Diyarbakır2 (3,029 reads), and Tekirdağ2 (3,225 reads). A little bit less relative abundance in Ankara2 (2,272 reads) and Ankara1 (1,006 reads). Then its abundance declines dramatically in Tekirdağ1 (456 reads), Konya2 (310 reads), Sivas2 (241 reads), Sivas1 (128 reads), and least abundant in Konya1 (114 reads). *Lysobacter terricola* also is most abundant in Diyarbakır2 (5,422 reads), then variably occurs less throughout the samples, and interestingly drops down to 68 reads only in Sivas1. Likely, *Pseudoxanthomonas sacheonensis* is found to be well abundant in Sivas1 uniquely with 1,764 reads. As well as *Lysobacter niabensis* which only has a remarkable read count in Tekirdağ2 (2,687 reads).

6.3 Experiment 3

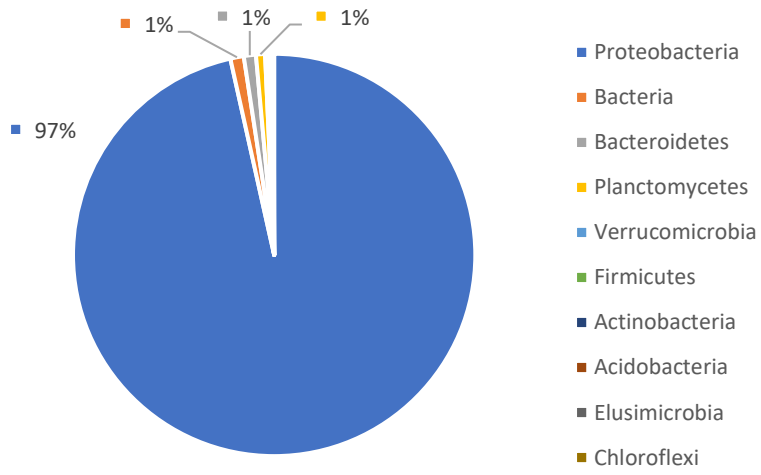

Figure 36. Distribution of top 10 phyla derived from 16S rRNA bacterial clone sequences of experiment 3 in all samples. The percentage corresponds to the total reads classified per phylum.

Although the experiment 3 has generated the largest dataset with 2.1 Gbases total yield, it did not differ much from the previous datasets in terms of phyla as illustrated in the pie chart (figure 36). Proteobacteria constitues 97% of the entire dataset classification by phylum, with 1% of Bacteria, Bacteroidetes, and Planctomycetes each, then the rest of sequence reads corresponds to a list of the commonly seen bacterial phyla: Verrucomicrobia, Firmicutes, Actinobacteria, Acidobacteria, Elusimicrobia, and Chloroflexi.

At the genus level, there is quite a bit difference in the genus distribution across all the samples (figure 37). Despite the huge abundance of Betaproteobacteria with about 157,999 total sequence reads, it is affected by the decreased appearance in both Tekirdağ1 and

Tekirdağ2 samples. Therefore, Burkholderiales ranks first on the list with 167,478 total reads. Lysobacter comes fifth with an increased total reads of 75,558, then 64,120 for Xanthomonadaceae, 55,133 for Ramlibacter, and 53,779 total reads for Massilia.

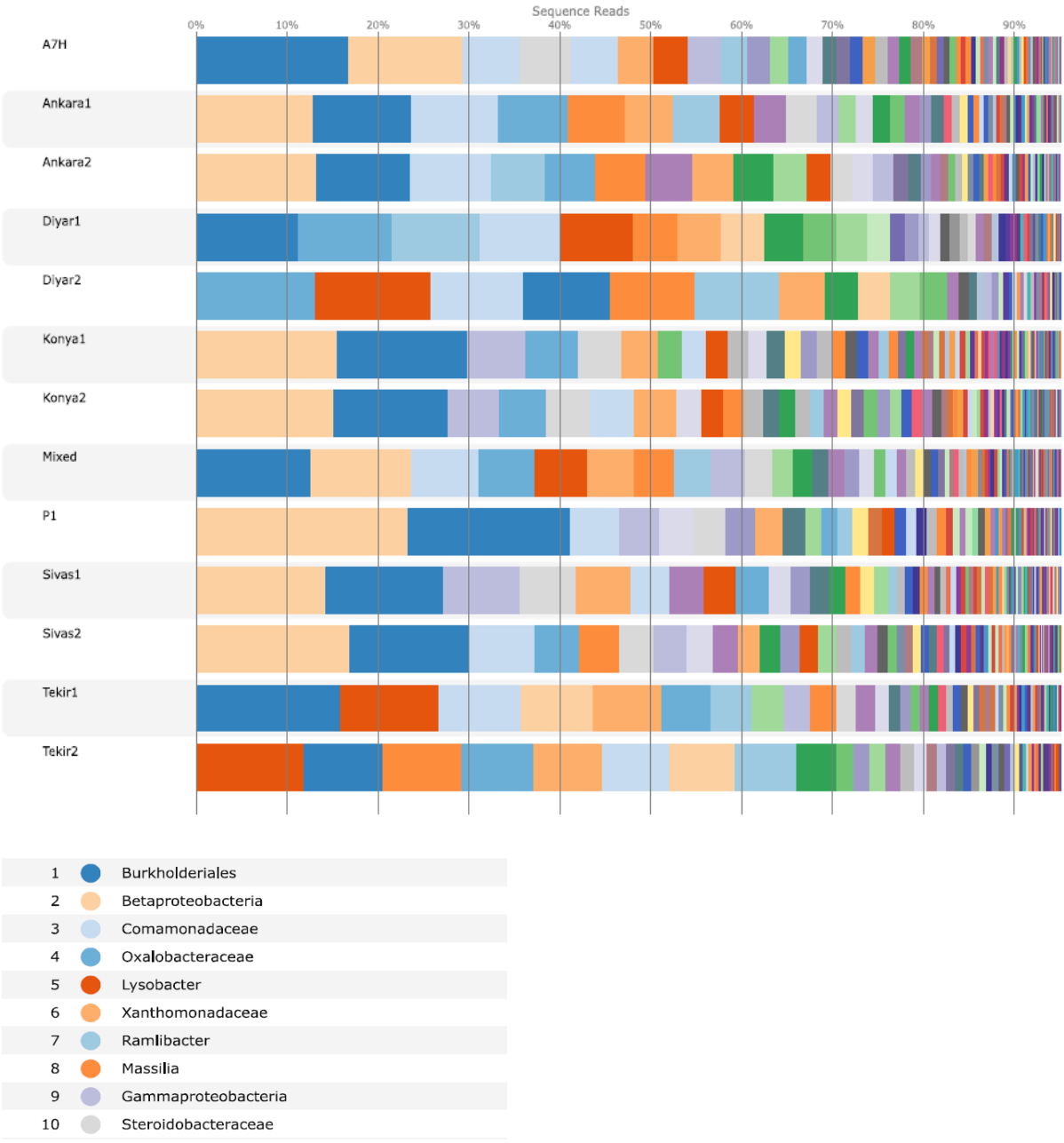

Figure 37. Bacterial genus distribution in different soil samples of experiment 3 with the 10 most dominant genus by total sequence reads, generated with Phinch2 (bioRxiv 009944; https://doi.org/10.1101/009944).

Below is the graph showing the 10 most abundant bacterial species by total read count for experiment 3 (figure 38). Similar to the classification by species of experiment 2, except that *Lysobacter terricola* takes over *Ramlibacter* species to be the most abundant with 30,087 total reads. Then followed by the three of Ramlibacter species: *monticola* (23,639 total reads), *tataouinensis* (12,174 total reads), and *ginsenosidimutans* (11,795 total reads).

Nitrosospira multiformis and *Variovorax paradoxus* come next (10,345 and 9,171 total reads, respectively). One species of *Pseudoxanthomonas mexicana* (8,293 total reads), and the last three species are as indicated on the graph.

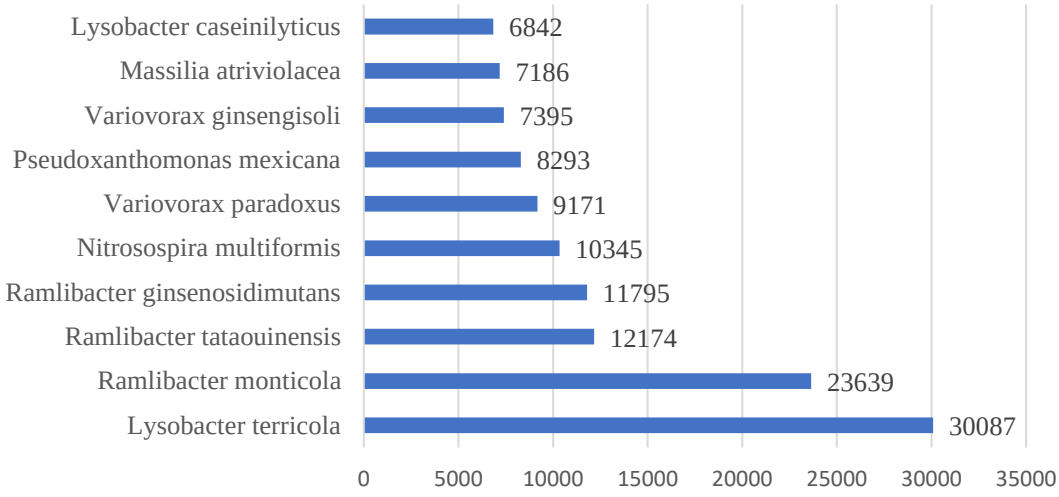

Figure 38. The 10 most abundant bacterial species in experiment 3 by total read count.

Speaking of *Ramlibacter* and *Massilia*, it is one more time demonstrated that these are grown in community as illustrated in the graph (figure 39). They both share relative abundance with pretty close total read counts (55,133 and 53,779 respectively), as well as in the different soil samples individually. Best seen in Diyarbakır2, the read count for *Ramlibacter* is 16,448 that is only a few sequences reads less than for *Massilia* with 16,575 reads. This very minimal variation between the two genus is also observed in the other soil samples.

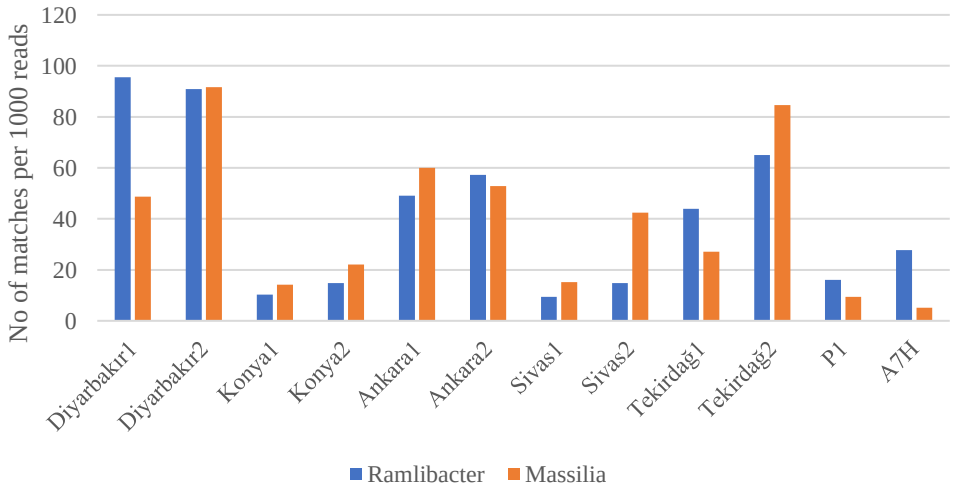

Figure 39. Relative abundance of community bacterial genus *Massilia* and *Ramlibacter*, experiment 3, in different types of soil samples.

Except for A7H which is almost 5 times higher in *Ramlibacter* sequences (3062 reads) than in *Massilia* (573 reads), as well as Sivas2, where *Massilia* takes over with 2328 reads and *Ramlibacter* appears with 813 reads only. While in Diyarbakır1, Ramlibacter overrides almost with double the relative abundance of Massilia (3,601 and 1,835 reads respectively).

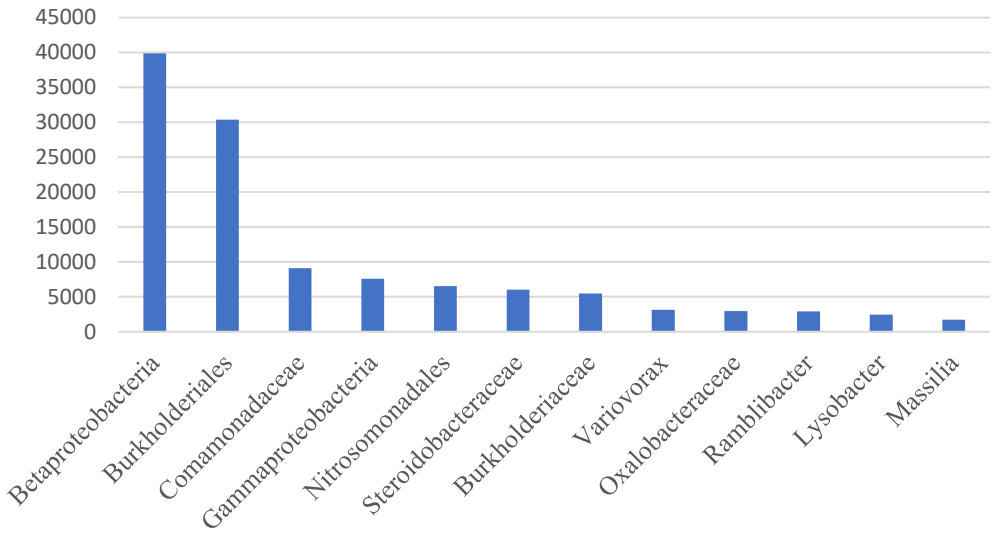

Figure 40. Demonstration of minimal bacterial genus variety in P1 sample based on the read count of the predominant genus classification.

Several other observations were pointed out too. Unlike the wheat soil samples which comprise a nice mixture of different genus distribution, the dry soil pathogen infected P1 sample does not look the same. In (figure 40) marking the read counts of the first few predominant bacterial genus, we only observe a remarkable abundance of Betaproteobacteria (39,865 sequence reads) that is the highest out of all the samples in all of the three experiments. Followed by a close rate of abundance in Burkholderiales with 30,385 sequence reads. Then the read count dramatically drops to 9,124 reads of Comamonadaceae, until it reaches only 1,700 reads of *Massilia*.

Another thing that got our attention, is the selective abundance or distribution of *Pseudoxanthomonas* genus across the different soil samples (figure 41). *Pseudoxanthomonas* ranks amongst the 20 most dominant bacterial genus by total read count, yet in some soil samples way more than others. For instance, both Ankara1 and Ankara2 samples are the richest in *Pseudoxanthomonas* with 4,018 (18.9%) and 3,927 (18.5%) reads respectively. This occurrence decreases to more than the half in Konya2 (1,765 reads; 8.30%) and Sivas2 (1,452 reads; 6.82%). Whereas there is barely any occurrence in Diyarbakır1 (49 reads, that is roughly 0.23%) and Tekirdağ2 (74 reads, that is roughly 0.34%), considering that the total read count for *Pseudoxanthomonas* is 21,267 reads.

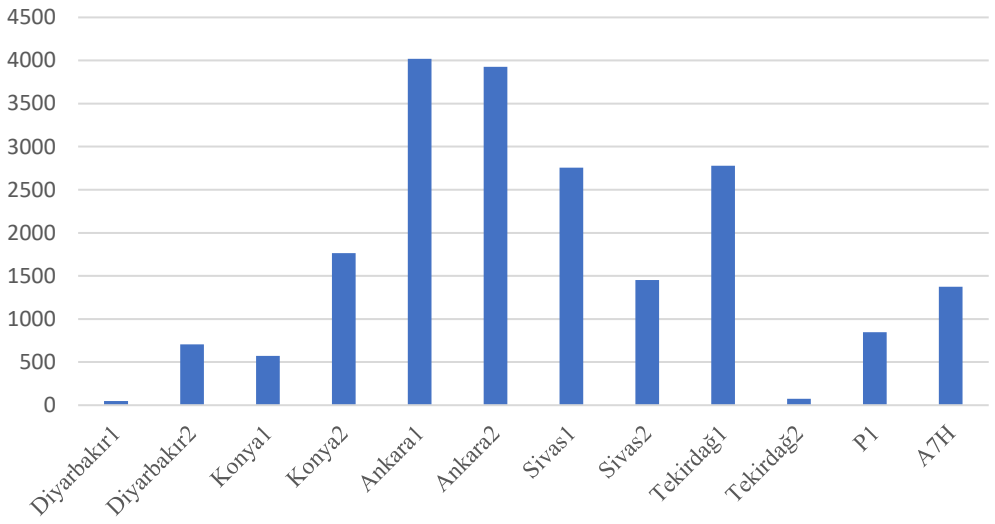

Figure 41. Selective abundance variation of *Pseudoxanthomonas* genus, experiment 3, in different soil samples by read count.

DISCUSSION

1. PCR inhibition

1.1 PCR inhibition can be eliminated by serial dilution of stock samples

PCR is an experimental method used to prove the taxonomic or functional abundance of certain organisms or organism groups, like bacteria. Technically, it requires the extraction of nucleic acids from the sample of study beforehand and the use of specific primers tailored to the gene/product to be amplified. However, there has been many concerns about the PCR inhibitors that may be co-extracted along with the nucleic acids during the extraction process. Those primarily include humic acids that limit the activity of the *Taq* polymerase enzyme, thus inhibit the PCR reaction. This effect can be overcome by several strategies, including pre-PCR treatments and PCR modifications. For the modifications of the PCR reaction, alterations of the PCR recipe can be applied by adding performance-enhancing additives, such as bovine serum albumin (BSA), and selectively using thermostable DNA polymerases (Kreader A., 1996). On the other hand, pre-PCR treatments comprise purification and dilution of the DNA extracts before setting the PCR reaction (LaMontagne et al., 2002).

In our study, we adopted the serial dilution strategy for our samples provided that it is simpler and cost effective. In fact, dilution of the DNA extract is able to reduce the concentration of the co-extracted PCR inhibitors in soil until they no longer inhibit the reaction. This might be challenging for the DNA yield as template concentration is also reduced, except that PCR reactions produce a high quantity of the amplified sequence through exponential amplification enough to carry on with the downstream applications. Researchers have been concerned about determining the proper dilution range or ratio for soil samples. For instance, no dilution or minor dilution like two-fold was shown to result in complete or partial PCR inhibition (Wang et al., 2017). Instead, for each soil type there was a corresponding dilution range at which the PCR inhibition was eliminated (within 4- to 400- fold dilution), and 40- to 60- fold dilution range was optimal based on their DNA extraction protocol. Yet, findings indicated that the dilution of soil DNA extracts aiming to reduce the co-extracted inhibitory components simultaneously reduces the target gene concentration (Bustin et al., 2009; McKee et al., 2015). In other words, dilution of genes with low concentrations causes copies number to be undetected (Jane et al., 2015). Thus, we decided to proceed with the two-fold dilution strategy on our samples and consider possible modifications if needed.

To ensure that serial dilution was ever needed, we have tested the first-step PCR reaction at the very first with the wheat soil DNA extracts directly, and expectedly, results were always negative. The two-fold serial dilution strategy was definitely successful on our samples,

however with variable results. For the experiment 2, only (1:10) working samples of the fresh stocks were prepared for PCR and the agarose gel was positive for all of them, especially for Diyarbakır 1 and 2 samples which bands looked most clear and thick. It was only a bit minimal for Sivas2 and Tekirdağ2 samples which gave thin DNA bands. So, there was no need for full serial dilution with higher ratios.

In the experiment 3, 4 diluted samples were prepared for each of P1 and A7H, and the PCR results were positive for all. Except that the most significant detection appeared in the (1:10) ratio in P1 where the DNA band gradually decrease in thickness until it became very minimal in (1:80) ratio. Likely in A7H, only the (1:10) ratio sample was clearly detectable, whereas the other 3 ratios looked most thin and faint. We conclude that, for these pathogen-infected soil samples, excessive dilution over (1:10) ratio causes reduction of the DNA yield simultaneously with the PCR inhibitors.

The major challenge we faced was in the experiment 1. For the sake of organizing the loaded samples in order to fit on the agarose gel, we have divided the samples into 3 sets in which each stock sample is diluted into 4 ratios as we discussed. The first set of Diyarbakır1 and Konya1 samples was interesting as the PCR reaction worked for Diyarbakır1 only in (1:80); suggesting that the DNA within was affected by a large quantity of humic acids. While for Konya1, only the (1:10) ratio was positive suggesting that the sample is too fragile for over dilution.

Interestingly, the second set of samples was challenging as well. None of the Diyarbakır2 and Konya2 samples were positive, indicating again a high rate of PCR inhibition and the need for further serial dilution. For Tekirdağ1, we could hardly detect extremely faint DNA bands in (1:40) and (1:80) ratios by adjusting the contrast of the gel image. Similarly in Tekirdağ2 and Sivas1, the bands were not satisfying yet the dilution seemed somehow optimal in (1:10) and (1:20) for Tekirdağ2, and in (1:20) and (1:40) in Sivas1. So, we performed a second round of two-fold serial dilution for Diyarbakır2 and Konya2 with higher dilution ratios, starting from the (1:80) samples to (1:160), (1:320), and (1:640). All of the samples showed positive bands with little difference in thickness. The dilution ratio (1:160) was considered optimal for Diyarbakır2 for purification, where the higher dilutions became less successively. Likely, Konya2 (1:160) looked optimal, and we pooled both (1:320) and (1:640) samples together for purification as well for testing although it could not get as remarkable as the (1:160) ratio sample.

For the last set of samples, the results were much better regardless of the bad gel image. PCR inhibition bas eliminated in almost all of the samples with positive DNA bands in Sivas2, Ankara1, and Ankara2 in all the dilution ratios; with (1:10) being the best result then gradually reduced. The sample repeats Anakar1r and Tekirdağ2r were positive as well. In Ankara1r, the (1:10) sample looked the most successful and the (1:80) sample got barely seen. However in Tekirdağ2r, the bad-quality gel image did not allow us to judge clearly on the dilution although the PCR seemed to have worked. Actually, we have repeated the visualization of the samples on agarose gel yet we did not show it in the results because it also appeared dirty for the other samples. Nevertheless, judging by that, (1:20) ratio was chosen in Tekirdağ2r for purification. Overall, Tekirdağ2r was a hard sample to deal with compared to the other samples.

All of these outcomes and the interesting variability to determine the appropriate dilution factor for each sample can be explained by the complexity of the soil itself. Indeed, Turkey is a large country with many vast greenfields and agricultural areas, where each is characterized by its unique geographical features, planting mode or strategy, use of fertilizers, crop types, the very unstable weather, and more. These factors are of great significance to constitute the unique identity and composition of the soil underneath. Hence, sequencing one soil sample of a determined region or area differs from another in many aspects, including DNA extraction, serial dilution, PCR reactions, product quantification, and the generated data. In addition, by comparing some of the wheat soil samples across the experiments 1 and 2, the optimal dilution factor for the same sample was variable in each. Particularly, Diyarbakır2 and Konya2 which could not amplify successfully with slight dilution in experiment 1 (with (1:10) dilution), were absolutely fine with it in experiment 2 and the PCR inhibition was eliminated. Same applies to Diyarbakır1, which best amplified in experiment 2 with minimal (1:10) dilution but could only succeed in higher dilution ratio (1:80) in experiment 1. Basically, we may relate this to the extraction process beforehand that much less humic acids and other PCR inhibitors were co-extracted with the soil DNA in the experiment 2 compared to the experiment 1, as well as in the experiment 3 even though the same extraction kit has been used. This suggests that the extraction method is not highly reproducible; which all in all may have affected the quantification of the PCR barcoding products by resulting in much higher DNA yeild for all the samples in experiment 2 judging by (table 1) and (table 2), and in experiment 3 (table 3).

2. Soil microbial diversity in Turkey revealed with metagenomic sequencing

2.1 Proteobacteria predominates different types of soils in Turkey

Our generated data analysis has revealed that Proteobacteria is the absolute predominant bacterial phylum in different soil types in Turkey. Not only in wheat cultivated soil samples from 5 different regions, but also the pathogen infected soil samples P1 and A7H. Recalling that P1 is a dry soil potato sample from an unknown pathogen-infected field, while A7H is an organic soil tomato sample that was collected from Antalya. Hence, those totally different parameters of soil types including the region, field of collection area, agricultural mode, and infection/health, probably would have had a great impact on the bacterial composition or variation in the samples. Yet, it was obvious through all of the sequencing experiments that 96-97% of the bacteria is classified as Proteobacteria, and the few percentages left would comprise all of the other bacterial phyla including Bacteroidetes, Planctomycetes, Firmicutes, Verrucomicrobia, Actinobacteria, Acidobacteria, and more.

That is not the case in the other soil DNA sequencing studies that we mentioned early in the introduction. For instance, in the wheat rhizosphere DNA sequencing study recently accomplished in India using MinION (Srivastava et al., 2020), the metagenome dataset analysis has generated only 51,909 reads analyzed out of 10 wheat soil samples randomly collected; that is 1/7 of the smallest dataset we obtained in the experiment 1 (372,904 total reads analyzed). Their dataset analysis uncovered that the predominant Proteobacteria phylum accounts for 68% only, followed by Firmicutes (13%), Bacteroidetes (3%), Actinobacteria (3%), and Acidobacteria (3%).

Also, the MinION sequencing of Canadian ice-wedge soil microbial communities has revealed that the predominant bacterial phyla were, first Alphaproteobacteria, followed by Acidobacteria, Actinobacteria, and Bacteroidetes. Although not precisely indicating the percentages of each, the graphs demonstrated a less composition in Alphaproteobacteria strictly, instead more rich and variable in the other phylum compared to our results (Goordial et al., 2017). Bearing in mind that this ice-wedge soil comprises mainly a permafrost layer that is completely frozen, topped with an active layer that is constantly thawing during summer and completely frozen with the permafrost during winter and spring.

In addition, the metagenomic analysis of microbial community and their function in Cadmium (Cd)-contaminated soil samples were investigated in China (Feng et al., 2018). Though the sequencing was performed using an Illumina cBot sequencer with 2 soil samples only, the researchers were able to summarize 77 taxa in total. Out of which, Proteobacteria,

Gemmatimonadetes, Thaumarchaeota, and Acidobacteria were determined as the predominant phyla accounting for over 75% of the total population. In detail, the most abundant phylum was Proteobacteria again, but comprising of no more than 38.56% in one of the samples, and 57.85% in the other.

In fact, Proteobacteria are an essential component of a healthy soil providing basic functions in the biogeochemical cycle. It is a major phylum of Gram-negative bacteria. It can basically embrace Betaproteobacteria which was most revealed in our results, alongside Alphaproteobacteria, Gammaproteobacteria, Deltaproteobacteria, Epsilonproteobacteria, as well as some other child taxa like *Ramlibacter monticola*, *Lysobacter terricola*, Burkholderiales, and more. These microorganisms play a significant role in soil nitrogen fixation. Regarding their importance as iron-oxidizing microorganisms, Proteobacteria can be subdivided into four main physiological categories: acidophilic, aerobic iron-oxidizers; neutrophilic, aerobic iron-oxidizers; neutrophilic, anaerobic iron-oxidizers (nitrate-dependent); and anaerobic photosynthetic iron oxidizers (Hedrich et al., 2011). On a side note, Alphaproteobacteria are oligotrophs, organisms that are capable of living in low-nutrient medium or environment such as deep oceanic sediments or deep undersurface soil. They comprise a number of pathogenic species like *Agrobacterium* and *Brucella*, and other essential nitrogen-fixing bacterial species like *Rhizobium* and *Methylocystis*. Whereas our predominantly found Betaproteobacteria are eutrophs, symbolling for microorganisms that require a copious amount of organic nutrients to survive in a certain environment. Likely, they also include several pathogenic (e.g., *Neisseria*, *N. meningitides* causing human diseases…) and healthy (e.g., *Leptothrix* as aquatic iron- and manganese- oxidizer, *Thiobacillus* as an aerobic acidophilic iron- and sulfur- oxidizer…) bacterial species (https://opentextbc.ca/microbiologyopenstax/chapter/proteobacteria/). Basically, our data did not reveal any of those severely pathogenic species, with very large abundance in Betaproteobacteria, that might affiliate to the overall healthy, good-functioning soil composition in the selected regions of Turkey.

In fact, the given observations are slightly surprising compared to the studies' results in other countries that we discussed. Something to take a careful look at to consider any experimental factors that might have biased our output results in this study, provided that it is the first of its kind in Turkey. For instance, the sequencing datasets observations suggest that the same DNA extraction method we used for all the experiments, in collaboration with the GTU team, is not highly reproducible since it could not give constant outputs from the same soil samples

across the experiments 1 and 2 as we mentioned before. Nevertheless, we must consider that the laboratory equipment used in the experimental setups are also a factor that could have affected the quality of the DNA extraction process. One way that could be suggested to rule out the possibility of the biased results towards Proteobacteria, is by repeating the sequencing experiment using another NGS technology, like Illumina MiSeq, similarly to the experimental approach that has been adopted in other studies (Goordial et al., 2017). This would allow us to better evaluate the reproducibility of the technology we used, and the accuracy of our data analysis results.

2.2 Ramlibacter and Massilia in microbial community

Ramlibacter and *Massilia* are both genera derived from the phylum of Proteobacteria, class of Betaproteobacteria, order of Burkholderiales. Yet, they apart at family taxa where *Ramlibacter* derive from Comamonadaceae, and *Massilia* from Oxalobacteraceae. Our findings have demonstrated a defined pattern of co-occurrence of these two genera throughout the entire study. As listed in the results section, *Ramlibacter* and *Massilia* tend to be abundant collectively in certain soil samples like Diyarbakır2, Ankara1, and Tekirdağ2 (figures 26 and 39), then the decline in abundance for one genus in a sample consequently affects that of the other like in Konya1, Sivas1, Sivas2, and A7H. This consistent distribution pattern can be explained by the fact of growing as community bacteria that greatly support the microorganisms living within collectively. In one of the studies mentioned earlier in the introduction, networks of co-occurrence of wheat rhizosphere in Washington, USA were established through complex analysis (Mahoney et al., 2017). However, strong networks of different community genera were demonstrated like members of classes Alphaproteobacteria (genera *Methylovirgula* and *Acidiphilium*), Betaproteobacteria (genus *Collimonas*), Gammaproteobacteria (genus *Serratia*), Actinobacteria (genus *Frankia*), and Sphingobacteria (genus *Mucilaginibacter*). That is not surprising provided that they obtained quite a different, more variable, classification of phyla as we described before. Therefore, it is essential to undergo more complex analysis in our sample study to better define the correlation between *Ramlibacter* and *Massilia*, target specific species line that contribute to this network, and link to other potential networks of community bacteria.

2.3 Pseudoxanthomonas acquires selective distribution pattern

As for some bacterial genera or species that contribute to the rhizosphere in a community fashion of co-occurrence, regardless of their abundance levels, other observations have been elaborated throughout the results of this study. To our attention, *Pseudoxanthomonas* genus

have proven to appear selectively in soil samples of different regions. Particularly, in Ankara and Sivas samples, the sequence read counts for *Pseudoxanthomonas* were quite remarkable in all the three experiments. While in other sample regions, like Diyarbakır and Tekirdağ, there would only be extremely low or none at all occurrences detected, in all three experiments again.

As an overview, *Pseudoxanthomonas* is affiliated to the class Gammaproteobacteria, family Xanthomonadaceae, phylum Proteobacteria. Members of *Pseudoxanthomonas* genus are described as non-spore forming rods, Gram-negative bacteria. Some classified species, like *P. kaohsiungensis* and *P. gei* were isolated from an oil-polluted site and plant stem respectively (Chang et al., 2005; Zhang et al., 2014). It has been reported that these members important ecological contributors as they are capable of reducing both nitrite and nitrate, and degrading various hydrocarbons (like benzene, toluene, ethyl-benzene, and xylene) (Nayak et al., 2011; (Xu et al., 2014). Recently, a novel species was identified, *Pseudoxanthomonas* arseniciresistens sp. nov., closest neighbors with *P. mexicana*, *P. japonensis*, *P. putridarboris*, and *P. indica*, as potential Arsenic (As)-reducing bacteria in the subsurface aquifer environments (Mohapatra et al., 2018).

In fact, our datasets have also revealed species of *Pseudoxanthomonas*; 17 species in experiment 1, and 21 species in each of experiment 2 and 3 differently. Amongst the lists are actually species that have been studied as mentioned lately: *P. mexicana*, *P. gei*, *P. japonensis*, *P. indica*, *P. kaohsiungensis*, and many others. This might be an interesting indicator that the sample soils in which *Pseudoxanthomonas* was relatively most abundant are in fact high in plant soil toxins or contaminators… And therefore, it would serve the project aim to target them with dedicated strategies of soil detoxification and recycling for better agricultural potentials.

CONCLUSION AND FUTURE WORK

As a summary, our study has contributed to the current and future scientific findings on soil fertility and plant health in Turkey. Indeed, one of the main objectives of the TÜBİTAK 1001 has been met, that is to define and characterize specific complex microbial diversity and consortia in wheat-cultivated soils in Turkey. We have established novel metagenomic analysis of soil DNA extracts using third generation nanopore sequencing technology with ONT's MinION Mk1B device, which proudly has never been adopted before in Turkey for soil microbiome research. In particular, our findings reveal a major abundance (96 to 98%) of Proteobacteria in all the 5 biogeographically different regions in Turkey, that is remarkably higher than in other countries like India, China, Canada, and USA. There seem to be constant relative abundance of the top 10 dominant bacterial genera and species in Turkey, like *Ramlibacter*, *Massilia*, *Variovorax*, *Lysobacer*, *Pseudomonas*, *Pseudoxanthomonas*, and more. Of these, we have identified *Ramlibacter* and *Massilia* as community bacterial genera based on their co-occurrence pattern across all 5 regions, for which we aim to construct more complex and relative analysis. In addition, a very unique observation has been marked for *pseudoxanthomonas* genus which has proven to acquire selective abundance pattern across the 5 regions, particularly in Ankara and Sivas. This bacterial genus has got 17 to 21 different species sequenced in all soil samples with alterations relative abundance, some of which have already been shown to have pivotal roles in soil detoxification and nutrient recycling. For effective contribution to the ultimate aims of the project, we suggest that pseudoxanthomonas must be fairly studied for the mechanism of counteracting heavy metals activity in soil, and the recycling of essential nutrients to be embedded in the soil fertility and plant growth and health strategies by the corresponding authorities and scientific communities.

REFERENCES

Astier, Y., Braha, O., & Bayley, H. (2006). Toward single molecule DNA sequencing: Direct identification of ribonucleoside and deoxyribonucleoside 5′-monophosphates by using an engineered protein nanopore equipped with a molecular adapter. *Journal of the American Chemical Society*, *128*(5), 1705–1710. https://doi.org/10.1021/ja057123+

Bassler, B. L., & Losick, R. (2006). Bacterially Speaking. *Cell*, *125*(2), 237–246. https://doi.org/10.1016/j.cell.2006.04.001

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., & Smith, G. P. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*. https://doi.org/10.1038/nature07517

Besser, J., Carleton, H. A., Gerner-Smidt, P., Lindsey, R. L., & Trees, E. (2018). Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clinical Microbiology and Infection*, *24*(4), 335–341. https://doi.org/10.1016/j.cmi.2017.10.013

Bustin, S. A., Benes, V., Garson, J. A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M. W., Shipley, G. L., Vandesompele, J., & Wittwer, C. T. (2009). The MIQE guidelines: Minimum information for publication of quantitative real-time PCR experiments. *Clinical Chemistry*, *55*(4), 611–622. https://doi.org/10.1373/clinchem.2008.112797

Chang, J. S., Chou, C. L., Lin, G. H., Sheu, S. Y., & Chen, W. M. (2005). Pseudoxanthomonas kaohsiungensis, sp. nov., a novel bacterium isolated from oil-polluted site produces extracellular surface activity. *Systematic and Applied Microbiology*, *28*(2), 137–144. https://doi.org/10.1016/j.syapm.2004.11.003

Chatterjee, A., Moulik, S. P., Majhi, P. R., & Sanyal, S. K. (2002). Studies on surfactant-biopolymer interaction. I. Microcalorimetric investigation on the interaction of cetyltrimethylammonium bromide (CTAB) and sodium dodecylsulfate (SDS) with gelatin (Gn), lysozyme (Lz) and deoxyribonucleic acid (DNA). *Biophysical Chemistry*, *98*(3), 313–327. https://doi.org/10.1016/S0301-4622(02)00107-2

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., … Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, *323*, 133–138.

https://doi.org/10.1126/science.1162986

Fatima, F., Pathak, N., & Rastogi Verma, S. (2014). An Improved Method for Soil DNA Extraction to Study the Microbial Assortment within Rhizospheric Region. *Molecular Biology International*, *2014*, 1–6. https://doi.org/10.1155/2014/518960

Feng, G., Xie, T., Wang, X., Bai, J., Tang, L., Zhao, H., Wei, W., Wang, M., & Zhao, Y. (2018). Metagenomic analysis of microbial community and function involved in cd-contaminated soil. *BMC Microbiology*, *18*(1), 1–13. https://doi.org/10.1186/s12866-018-1152-5

Fukuda, K., Ogawa, M., Taniguchi, H., & Saito, M. (2016). Molecular approaches to studying microbial communities: Targeting the 16S ribosomal RNA gene. *Journal of UOEH*, *38*(3), 223–232. https://doi.org/10.7888/juoeh.38.223

Goordial, J., Altshuler, I., Hindson, K., Chan-Yam, K., Marcolefas, E., & Whyte, L. G. (2017). In situ field sequencing and life detection in remote (79°26'N) Canadian high arctic permafrost ice wedge microbial communities. *Frontiers in Microbiology*, *8*(DEC), 1–14. https://doi.org/10.3389/fmicb.2017.02594

Handelsman, J., & Tiedje, J. (2007). The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet. *Washington, DC, USA: THE NATIONAL ACADEMIES PRESS*. https://doi.org/10.17226/11902

Handelsman, Jo, Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chemistry and Biology*, *5*(10). https://doi.org/10.1016/S1074-5521(98)90108-9

Hedrich, S., Schlömann, M., & Barrie Johnson, D. (2011). The iron-oxidizing proteobacteria. *Microbiology*, *157*(6), 1551–1564. https://doi.org/10.1099/mic.0.045344-0

Jane, S. F., Wilcox, T. M., McKelvey, K. S., Young, M. K., Schwartz, M. K., Lowe, W. H., Letcher, B. H., & Whiteley, A. R. (2015). Distance, flow and PCR inhibition: eDNA dynamics in two headwater streams. *Molecular Ecology Resources*, *15*(1), 216–227. https://doi.org/10.1111/1755-0998.12285

Kreader A., C. (1996). Relief of amplification inhibition in PCR with bovine serum albumin or T4 Gene 32 Protein. *Applied and Environmental Microbiology*, *62*(3), 1102–1106.

LaMontagne, M. G., Michel Jr, F. C., Holden, P. A., & Reddy, C. A. (2002). Evaluation of extraction and purification methods for obtaining PCR-amplifiable DNA from compost for microbial community analysis. *Journal of Microbiological Methods*, *49*, 255–264. https://doi.org/10.1016/s0167-7012(01)00377-3

Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G., & Webb, W. W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, *299*, 682–686. https://doi.org/10.1126/science.1079700

Lever, M. A., Torti, A., Eickenbusch, P., Michaud, A. B., Šantl-Temkiv, T., & Jørgensen, B. B. (2015). A modular method for the extraction of DNA and RNA, and the separation of DNA pools from diverse environmental sample types. *Frontiers in Microbiology*, *6*(MAY). https://doi.org/10.3389/fmicb.2015.00476

Lugtenberg, B., & Kamilova, F. (2009). Plant-Growth-Promoting Rhizobacteria. *Annual Review of Microbiology*, *63*, 541–556. https://doi.org/10.1146/annurev.micro.62.081307.162918

Mahoney, A. K., Yin, C., & Hulbert, S. H. (2017). Community structure, species variation, and potential functions of rhizosphere-associated bacteria of different winter wheat (Triticum aestivum) cultivars. *Frontiers in Plant Science*, *8*(February), 1–14. https://doi.org/10.3389/fpls.2017.00132

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–380. https://doi.org/10.1038/nature03959

McCarty, N. S., & Ledesma-Amaro, R. (2019). Synthetic Biology Tools to Engineer Microbial Communities for Biotechnology. *Trends in Biotechnology*, *37*(2), 181–197. https://doi.org/10.1016/j.tibtech.2018.11.002

McKee, A. M., Spear, S. F., & Pierson, T. W. (2015). The effect of dilution and the use of a post-extraction nucleic acid purification column on the accuracy, precision, and inhibition of environmental DNA samples. *Biological Conservation*, *183*, 70–76.

Mendes, L. W., Tsai, S. M., Navarrete, A. A., de Hollander, M., van Veen, J. A., & Kuramae, E. E. (2015). Soil-Borne Microbiome: Linking Diversity to Function. *Microbial*

Ecology, *70*(1), 255–265. https://doi.org/10.1007/s00248-014-0559-2

Mohapatra, B., Sar, P., Kazy, S. K., Maiti, M. K., & Satyanarayana, T. (2018). Taxonomy and physiology of pseudoxanthomonas arseniciresistens sp. nov., an arsenate and nitrate-reducing novel gammaproteobacterium from arsenic contaminated groundwater, India. *PLoS ONE*, *13*(3), 1–18. https://doi.org/10.1371/journal.pone.0193718

Nayak, A. S., Sanjeev Kumar, S., Santosh Kumar, M., Anjaneya, O., & Karegoudar, T. B. (2011). A catabolic pathway for the degradation of chrysene by Pseudoxanthomonas sp. PNK-04. *FEMS Microbiology Letters*, *320*(2), 128–134. https://doi.org/10.1111/j.1574-6968.2011.02301.x

Nora, L. C., Westmann, C. A., Martins-Santana, L., Alves, L. de F., Monteiro, L. M. O., Guazzaroni, M. E., & Silva-Rocha, R. (2019). The art of vector engineering: towards the construction of next-generation genetic tools. *Microbial Biotechnology*, *12*(1), 125–147. https://doi.org/10.1111/1751-7915.13318

Pareek, C. S., Smoczynski, R., & Tretyn, A. (2011). Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, *52*(4), 413–435. https://doi.org/10.1007/s13353-011-0057-x

Priya, P., Aneesh, B., & Harikrishnan, K. (2021). Genomics as a potential tool to unravel the rhizosphere microbiome interactions on plant health. *Journal of Microbiological Methods*, *185*(December 2020), 106215. https://doi.org/10.1016/j.mimet.2021.106215

Rusk, N. (2009). *Cheap third-generation sequencing*. *6*(4), 244–245.

Sadigov, R. (2022). Rapid Growth of the World Population and Its Socioeconomic Results. *Scientific World Journal*, *2022*(1930). https://doi.org/10.1155/2022/8110229

Sebastián-Domingo, J. J., & Sánchez-Sánchez, C. (2018). From the intestinal flora to the microbiome. *Revista Espanola de Enfermedades Digestivas*, *110*(1), 51–56. https://doi.org/10.17235/reed.2017.4947/2017

Şeker, M. G., Şah, I., Kırdö, E., Ekinci, H., Çiftçi, Y. Ö., & Akkaya, Ö. (2017). A Hidden Plant Growth Promoting Bacterium Isolated from In Vitro Cultures of Fraser Photinia (Photinia × fraseri). *International Journal of Agriculture & Biology*, *July 2018*, 1511–1519. https://doi.org/10.17957/IJAB/15.0455

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., &

Waterston, R. H. (2017). DNA sequencing at 40: Past, present and future. In *Nature* (Vol. 550, Issue 7676). https://doi.org/10.1038/nature24286

Simonin, M., Dasilva, C., Terzi, V., Ngonkeu, E. L. M., DIouf, Di., Kane, A., Béna, G., & Moulin, L. (2020). Influence of plant genotype and soil on the wheat rhizosphere microbiome: Evidences for a core microbiome across eight African and European soils. *FEMS Microbiology Ecology*, *96*(6), 1–18. https://doi.org/10.1093/femsec/fiaa067

Srivastava, R., Srivastava, A. K., Ramteke, P. W., Gupta, V. K., & Srivastava, A. K. (2020). Metagenome dataset of wheat rhizosphere from Ghazipur region of Eastern Uttar Pradesh. *Data in Brief*, *28*, 105094. https://doi.org/10.1016/j.dib.2019.105094

Stenuit, B., & Agathos, S. N. (2015). Deciphering microbial community robustness through synthetic ecology and molecular systems synecology. *Current Opinion in Biotechnology*, *33*(1), 305–317. https://doi.org/10.1016/j.copbio.2015.03.012

Stoddart, D., Heron, A. J., Mikhailova, E., Maglia, G., & Bayley, H. (2009). Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(19), 7702–7707. https://doi.org/10.1073/pnas.0901054106

Wang, H., Qi, J., Xiao, D., Wang, Z., & Tian, K. (2017). A re-evaluation of dilution for eliminating PCR inhibition in soil DNA samples. *Soil Biology and Biochemistry*, *106*, 109–118. https://doi.org/10.1016/j.soilbio.2016.12.011

Wilhelm, R. C., Niederberger, T. D., Greer, C., & Whyte, L. G. (2011). Microbial diversity of active layer and permafrost in an acidic wetland from the Canadian high arctic. *Canadian Journal of Microbiology*, *57*(4), 303–315. https://doi.org/10.1139/w11-004

Xu, M., Zhang, Q., Xia, C., Zhong, Y., Sun, G., Guo, J., Yuan, T., Zhou, J., & He, Z. (2014). Elevated nitrate enriches microbial functional genes for potential bioremediation of complexly contaminated sediments. *ISME Journal*, *8*(9), 1932–1944. https://doi.org/10.1038/ismej.2014.42

Zhang, L., Wei, L., Zhu, L., Li, C., Wang, Y., & Shen, X. (2014). Pseudoxanthomonas gei sp. nov., a novel endophytic bacterium isolated from the stem of Geum aleppicum. *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, *105*(4), 653–661. https://doi.org/10.1007/s10482-014-0119-2