KU LEUVEN

FACULTY OF BIOSCIENCE ENGINEERING

Deciphering the role of IS*6100* in microbial community adaptation to anthropogenic environmental stressors

Promoter:

Prof. Springael Dirk

Department Earth and Environmental Sciences

Division Soil and Water Management

Dissertation presented in fulfillment of the requirements for the degree of Master of Bioscience Engineering: Cellular and Genetic Engineering

Kristina YEFIMAK

June 2024

KU LEUVEN

FACULTY OF BIOSCIENCE ENGINEERING

Foreword

After years of studying theoretical concepts during my study, I really looked forward to bringing my knowledge into practice. I am so grateful that I could do this in a lab where I genuinely enjoyed my journey. I would like to explicitly thank my supervisor, Panji, whose guidance and expertise have been indispensable in shaping this work. Even though most communication was through WhatsApp we were able to stay in touch and bring this thesis to a good end. I do hope we can meet again with a free card to the Zoo in Leiden. I also wish to express my appreciation to my promotor, Professor Springael, for his genuine engagement with numerous update meetings and support throughout this year. I look with pleasure back on my moments in the lab, Yunus playing all his favorite sing alongs, Johanna asking me how her bestie is doing and if I am still motivated or some crazy other questions, Marte having 10000 moments of confusion with me and still being able to laugh our asses of together, Nina for being the best cantus praeses with me, Marie for bringing ‘delicious’ chocolate to our ‘VV’ and Toon, Quinten, Carsten, Lynn, Arthur, Kaat, Tim, Seppe, Joke for their spirit to go STUNTEN, but to also to create a fun ‘kiekenkot’ in our too long coffee breaks. I love ‘de harde kern’.

Especially, I want to thank my close friends and family for their invaluable support. Tippels for always struggling with me through bib sessions, especially Sarah for biking Gasthuisberg with me uphill and having a last snack night as motivation, Sara for listening to my enormous amount of voice messages, Anke for always opening up her dorm for a thesis session, Floris for being the best neighbor in my dorm and Jennie at home, Anouk for calling me when I needed a boost (not only for the thesis) and sending me motivational messages until the very last day and creating fun days to look forward to, Pollie for her thoughtfulness, my travel girls Lisa and Ine and Léontine to make home feel like nothing has changed, Lili to bring another smile on my face and comfort me with a warm hug or help me write some code, my mom for always standing by my side no matter what and helping out during all my crises and broski for keeping me up to date about what is cool. This past year has presented some of the hardest challenges I have encountered, and your presence has been a true blessing throughout.

Abstract

Anthropogenic activities across various land uses (e.g. roadsides, industrial areas, agricultural zones) have introduced xenobiotics, substances not natural to the ecosystem, in the environment, affecting soil physiochemical characteristics and soil bacterial communities. To adapt to this changing environment, bacteria have evolved to withstand these harsh conditions. Insertion sequences (IS) can form composite transposons that recruit and carry along passenger genes, including those involved in the degradation of organic pollutants. However, studies examining IS elements abundance and its cargo in bacterial communities within their natural context is limited. The purpose of this thesis is to improve our knowledge of the role of the IS element IS*6100* in genetic adaptation within soil bacterial communities. This research focused on two parts, the first part on Pakistani soil samples that determined how bacterial community structure and abundance of IS*6100* associates with land uses, spatial, and environmental factors. Land use and location exerted significant influence on alpha and beta diversity in soil ecosystems, notably with agricultural soil demonstrating higher diversity due to increased disturbance. While environmental factors such as Cu, total nitrogen, As, Cd, and total carbon contribute to soil bacterial community variability, their impact is relatively small compared to land use in shaping soil bacterial composition. Nevertheless, 86% unexplained variability suggests the presence of additional important factors that were not measured in our study. IS*6100* is widespread, with no differences in abundance across various geographic Pakistani samples with differing land use. While effects of environmental factors including community diversity was low, some toxic contaminants like DiPCB and TriPCB exhibited a negative correlation with IS*6100* abundance. The second part examined contaminated soils in Belgium, assessing IS*6100* abundance by qPCR. Moreover, a new long-range PCR (LR-PCR) method was applied to amplify the cargo of IS*6100* associated CTs for functional analysis by nanopore sequencing. IS*6100* was present across all samples, with significant variations in relative abundance between soils. Particularly, soil samples continuously exposed to pesticides exhibited elevated IS*6100* abundances compared to those with less clear contaminants. Regarding cargo genes identification, clear bands in the AGE for every sample suggested that the LR-PCR IS*6100* cargo genes were successfully and reproducibly amplified. The amplicons were successfully sequenced by nanopore sequencing for future functional analysis.

List of abbreviations and symbols

AGE	Agarose gel electrophoresis
As	Arsenic
ASV	Amplicon Sequence Variants
BSA	Bovine serum albumin
bp	Base pair
BPS	Bio purification system
Cd	Cadmium
Cr	Chrome
CT	Composite transposon
Cu	Copper
eDNA	Environmental deoxyribonucleic acid
EC	Electrical conductivity
HGT	Horizontal gene transfer
HMW	High-molecular-weight
ICE	Integrative conjugative elements
iPLSR	Iterative PLSR
IS	Insertion sequence
kb	Kilo base pairs
LR-PCR	Long-range PCR
MGE	Mobile genetic element
Pb	Lead
PBS	Phosphate buffered saline
PCB	Polychlorinated biphenyl
PCoA	Principle Coordinates Analysis
PFAS	Per- and polyfluoroalkyl substances
PLSR	Partial Least Squares Regression
qPCR	Quantitative PCR
RDA	Redundancy analysis
rRNA	Ribosomal ribonucleic acid
TE	Tris-EDTA
TIR	Terminal inverted repeats
tnpA	Gene encoding for a transposase enzyme in transposon
TPH	Total petroleum hydrocarbons
TSD	Target site duplication
Zn	Zinc

Table of Contents

1 Introduction

Anthropogenic activities such as industrial activities, traffic and transportation, and agricultural practises have led to the spread of inorganic and organic pollutants including xenobiotics in the environment for the past centuries (Nojiri *et al.*, 2004). These chemical pollutants pose significant hazards to both human and ecosystem health, raising critical environmental concerns (Madamwar *et al.*, 2021). Studies found that human activities and pollutants associated with industry, transportation, and agriculture can alter various soil physico-chemical properties, including pH and nutrient levels (Alengebawy *et al.*, 2021; Briffa *et al.*, 2020), which affect soil biodiversity structure and functionality (Nugent & Allison, 2022). For instance, the presence of heavy metals such as arsenic (As), cadmium (Cd), chrome (Cr), copper (Cu) and lead (Pb) could impact the metabolic activity, diversity, and abundance of soil microbiomes (Yang *et al.*, 2022). Therefore, there is an urgent need to enhance our understanding of anthropogenic pollutants across various land uses (e.g., roadsides, industrial areas, agricultural zones) and their influence on soil communities, including the soil microbiome that plays a pivotal role in biochemical cycling and the degradation of diverse pollutants (Nugent & Allison, 2022).

In response to the dynamic nature of their environment, microbes, particularly bacteria, have acquired the capacity to adjust to and endure these stressful conditions through adaptive evolution (Shi *et al.*, 2022). This selective pressure eradicates susceptible bacteria and accumulates adapted bacteria that can for example degrade or show resistance to pollutants (Nojiri *et al.*, 2004). This evolutionary process is driven by mutations and genetic exchanges, leading to the selection of advantageous traits that contribute to the species' fitness and survival (Shi *et al.*, 2022). Horizontal gene transfer (HGT) is a key mechanism driving the development of new catabolic pathways and the spread of catabolic/resistance genes in bacterial communities (Top & Springael, 2003). HGT is facilitated by Mobile genetic elements (MGEs), collectively making up the mobilome in a bacterial community. MGEs facilitate the transfer of DNA between bacterial cells or between different replicons within the genome (Siguier *et al.*, 2014). The simplest MGE is an insertion sequence (IS) that can recruit and carry along

passenger genes for instance by forming composite transposons (CTs). CTs consist of two ISs flanking cargo DNA, which can contain a variety of genes, for instance genes associated with metabolic pathways and enzymes essential for the breakdown of organic pollutants or resistance genes. When the CT is part of a conjugative or mobilizable plasmid it can be transferred among various bacterial strains and taxa, becoming a substantial factor in the spread of cargo and adaptive traits carried by the CTs, facilitating adaptation of the microbial community. IS*6100* is an IS element that has been associated in a CT structure, with genes related to antibiotic resistance as well as the catabolism of organic pollutants (Varani *et al.*, 2021). The connection between IS*6100* and a range of functional genes related to catabolism, detoxification, and resistance mechanisms suggests its involvement in HGT and bacterial adaptation, facilitating the ability of bacteria to survive in varied environmental conditions influenced by human activities (Mahillon & Chandler, 1998; Varani *et al.*, 2021). However, this knowledge is only based on studies on bacterial strains isolated for a particular adaptive characteristic like resistance to metals/antibiotics or xenobiotic catabolism. To acquire a better understanding of the role of IS*6100* and other IS elements or MGEs in microbial adaptation in real environmental settings, cultivation-independent and microbial community wide studies are required. Limited attention has been directed towards investigating the occurrence of the abundance of IS elements and the passengers/cargo genes they are associated with in CTs in bacterial communities within their natural environmental context.

This thesis aims to enhance our understanding of IS*6100*'s role in genetic adaptation within complex soil microbial communities. The study consists of two parts. The first part focused on soils sampled in the North-Punjab province of Pakistan along a 400 km long road from Lahore in the south to Islamabad in the north, encompassing different geographical locations and land use types (agriculture, industry, roadside). High-throughput 16S rRNA gene amplicon sequencing was used to identify the composition of the bacterial community, and the relative abundances of IS*6100* were determined by real-time quantitative polymerase chain reaction (qPCR). Consequently relationships were sought between bacterial community composition, IS*6100* relative abundances, geographical location, land use type and physico-chemistry of the soils including the concentration of several contaminants. This to determine drivers of

community composition and IS*6100* prevalence. The second part focused on a number of contaminated soils in Belgium. IS*6100* relative abundances were determined across soils contaminated with different pollutants. In addition, a first attempt was made to recover and identify the cargo of IS*6100* associated CTs from those soils using a novel established long-range PCR (LR-PCR) approach combined with long-read nanopore sequencing. By exploring the complex dynamics of IS*6100* related CTs within polluted soils, it was aimed to obtain a deeper insight into the function of these elements in microbial community adaptability and pollutant degradation.

2 Literature study

2.1 The Role of HGT in Bacterial Community Adaptation

As the environment constantly changes, microbes have developed the ability to adapt to the challenges imposed by stressful events. This adaptive evolution is based on mutations and exchange of genetic information by HGT that allows a selection of certain beneficial treats to ensure survival of the species (Shi *et al.*, 2022). To ensure the presence of a transferred gene within the recipient over an extended period of time, the gene must confer a selective benefit (Soucy *et al.*, 2015). When the selection pressure is high for a specific treat, then the change of the bacterial genome in response to the altered environment is permanent (Shi *et al.*, 2022). An example of a mutation that resulted in adaptation is the exposure of *Mycobacterium tuberculosis* to the antibiotic rifampicin, which increased the mutation rate of *rpoB* gene, leading to rifampycin-resistant tuberculosis (Telenti *et al.*, 1993). Meanwhile, considering gene transfer as an adaptive mechanism, Barkay *et al.* (2003) showed that a bacterial community evolved to detoxify and resist mercury contamination by the spread of genetic elements that carry mercury-resistance genes. Microbial adaptation is further illustrated in the case of the uprise of industrial activity in the last centuries, which resulted into the environmental spread of xenobiotics (Nojiri *et al.*, 2004). Additionally, the growth of the human population has led to more crop production, consequently leading to an increase in the use of pesticides in agriculture (Wang *et al.*, 2022). Xenobiotics are chemicals that are not natural to the ecosphere (Leisinger, 1983). Nevertheless, bacteria have evolved to degrade these chemicals. At times, xenobiotics intentionally utilized and introduced into the environment (e.g., antibiotics, disinfectants, fungicides, and pesticides) were designed to exert toxic properties on specific target organisms. However, they also impose harmful and toxic effects on natural ecosystems and non-target organisms including humans and animals. This resulted in, with respect to microbial adaptation, that carbonous xenobiotics proliferated as a selective pressure, eliminating vulnerable organisms and leading to the build-up of adapted resistant microorganisms and even microbes that utilize the xenobiotic as a growth substrate (Madamwar *et al.*, 2021). Adaptation is especially expedited when these chemicals are released into the environment in substantial

quantities, as seen in extensive crude oil spills from pipelines, marine tankers, and oil wells (Nojiri *et al.*, 2004).

As mentioned, next to mutation, HGT is a driving force in microbial adaptation. Genetic transfer of genetic information can occur vertically and horizontally. Vertical gene transfer involves the transfer of genetic material from parent cells to offspring cells (Shi *et al.*, 2022). In contrast, HGT refers to the transfer of genes from one bacterium cell to another within the same generation and can even occur between cells belonging to different taxa (Keeling & Palmer, 2008). The occurrence of HGT is widespread as approximately 5-10% of a typical bacterial genome appears to originate from other species (Ochman *et al.*, 2000). The functions that bacteria transfer by HGT encompasses a wide variety of traits, such as catabolism of carbohydrates, resistance to metals like mercury, resistance to antibiotics, promotion of virulence, and catabolism/degradation of (xenobiotic) organic pollutants (Böhm *et al.*, 2015; Hehemann *et al.*, 2010; Hemme *et al.*, 2016; Huddleston, 2014; Ikuma & Gunsch, 2012; Summers *et al.*, 1993). For example, the transfer of chlorocatechol catabolic genes to a toluene-degrading organism within a chlorobenzene contaminated aquifer led to the establishment of a novel metabolic pathway for chlorobenzene degradation (Van der Meer *et al.*, 1991). Xenobiotic biodegradation might be initially facilitated by collaboration of multiple bacterial strains, each harbouring distinct genes responsible for various steps in the entire pathway. The assembly of these genes within a single organism, may evolve gradually through HGT between diverse making strains part of the cooperative consortium (Top & Springael, 2003).

2.2 MGEs and HGT Mechanisms

MGEs play a key role in driving HGT, by transferring genes between varied positions within the genome or transferring genes among distinct cells/taxa. MGEs consist of DNA segments that encode enzymes and other proteins responsible for facilitating transfer of DNA and can be divided in 2 categories: intercellular and intracellular MGEs. Intercellular mobility facilitates transfer between two bacterial cells involving MGEs like plasmids, bacteriophages and integrative conjugative elements (ICE) that transfer DNA between two living cells but also includes transformation in which naked DNA for

instance derived from lysed cells, is taken up by living cells (Frost *et al.*, 2005; Nazarian & Boedicker, 2018). In contrast, intracellular mobility relates to the translocation of DNA between different locations of the genome in the same cell (Frost *et al.*, 2005). Intracellular MGEs include transposons, ISs, integrons and introns which can recruit DNA. Moreover, in instances where these intracellular MGEs are found on intercellular MGEs, like an IS on a plasmid, these intercellular MGEs have the capability to seize and mobilize DNA across different cells and genomes (Vandecraen *et al.*, 2017). The collective set of MGEs withing a community together make up the mobilome (Siguier *et al.*, 2014).

2.2.1 HGT mechanisms

Intercellular HGT uses four main mechanisms as illustrated in Fig. 1: transformation, transduction, conjugation and vesiduction. **Transformation** is the process of bacteria taking up naked DNA from their environment, in plasmid form or as DNA fragments. Only a limited number of species have the competence to naturally take up DNA, other taxa need to be made competent (Nazarian & Boedicker, 2018), for example by increasing the membrane permeability to DNA (Chen & Dubnau, 2004). The induction of competence can be done in the lab through chemical treatment or electroporation (Hasegawa *et al.*, 2018). **Transduction** involves bacteriophages. Phages are independently replicating bacterial viruses (Frost *et al.*, 2005). Both the donor cell and recipient cell of the transferred DNA must be susceptible for infection by the same phage (Nazarian & Boedicker, 2018). In the lysogenic cycle, the phage DNA integrates into the bacterial chromosome and composes the prophage. The prophage DNA is inherited vertically during host replication and remains in the hosts' chromosome, while the bacterial host can use the genes that the prophage carries. The prophage can eventually initiate a lytic phase, leading to the demise of their host cells and release of the virulent particles (De Sousa *et al.*, 2023). Lytic and lysogenic phages can both engage in HGT of the host DNA through the process of transduction. There are 2 forms of transduction: the first is called generalized transduction involving the inclusion of a random fragment of host DNA during cell lysis. The second is specialized transduction, where a prophage imprecisely removes itself from the host genome and takes up part of the flanking host DNA, resulting in encapsulation of neighbouring genes (Soucy *et al.*, 2015). Phage-facilitated HGT enables genetic exchange to span diverse bacterial

lineages and due to the transportation of DNA within the viral capsids the DNA is protected from environmental factors that can degrade the DNA (De Sousa *et al.*, 2023). **Conjugation** transfers genes from donor to recipient by means of cell to cell contact (Nazarian & Boedicker, 2018). It requires conjugative plasmids or ICEs to transfer the DNA from one cell to another. While ICEs elements become part of the host's genome to maintain the transferred DNA in the new host, plasmids differ in that they replicate and exist independently, not integrating into the host's genetic material. The transfer is mediated by forming a pilus serving as the molecular bridge that attracts the cells to each other, and through which the plasmids or ICEs move into the recipient cell. In contrast to conjugative plasmids and ICEs, non-conjugative and non-mobilizable plasmids lack the capability to independently perform HGT. However, non-conjugative mobilizable plasmids can make use of the transfer functions of conjugative transfer to become exchanged (Frost *et al.*, 2005). **Vesiduction** is a recently described process and involves small extracellular vesicles, which are membrane vesicles including exosomes, encompassing DNA for intercellular transfer. Previously, it was assumed that only lipids, proteins and RNA could be transferred by these vesicles. The exact mechanisms underlying vesiduction remains to be studied (Ghanam *et al.*, 2022).

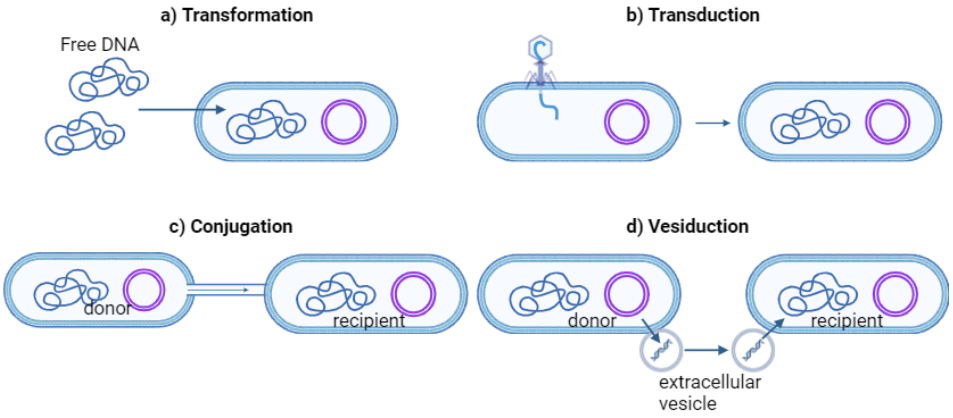


Fig. 1. The HGT process of transformation, transduction, conjugation and vesiduction. a) Transformation refers to the bacterial process of absorbing genetic material directly from the surroundings, where the DNA is not enclosed within a cell or membrane. b) Transduction is a process, in which bacteriophages, which are self-replicating viruses that infect bacteria, perform the transfer. c) Conjugation involves the direct transfer of genes from a donor to a recipient cell through physical contact between cells d) Vesiduction makes use of extracellular vesicles for DNA transfer, although the exact mechanism remains to be studied (created with Biorender.com).

Intracellular mobility occurs mainly through the mechanisms of homologous recombination and site-specific recombination (Fig. 2) (Siefert, 2009). The transfer of genetic material between DNA molecules with similar or identical sequences is facilitated by homologous recombination, while site-specific recombination involves the exchange of DNA sequences at specific recognition sites within DNA molecules (Trzilova & Tamayo, 2021). **Homologous recombination** is initiated by double stranded breaks in one pair of two aligned DNA segments with homologous sequences leading to ssDNA overhangs. The overhangs insert themselves into an intact homologous DNA strand and complementary base pairs are formed between the single-stranded overhangs and the intact DNA strand. This results in the formation of a Holliday junction, an intermediate where the exchange of genetic material between the two DNA strands takes place and eventually leads to resolution to finalize the recombination (Sung & Klein, 2006). **Site-specific recombination** uses two mechanisms, i.e., conservative and transpositional site-specific recombination. Conservative site-specific recombination involves shorter homologous DNA sequences compared to homologous recombination and a specialized recombinase enzyme that breaks and rejoins the DNA strands at the recombination sites (Parks & Peters, 2018). Transpositional site-specific recombination does not need large regions of homology (Siguier *et al.*, 2014). The DNA 'jumps' between replicons, i.e., the DNA sequence moves from one location to the other within the genome of an organism. The DNA can even insert into phage DNA or plasmids, making it prone to intercellular transfer (Frost *et al.*, 2005).

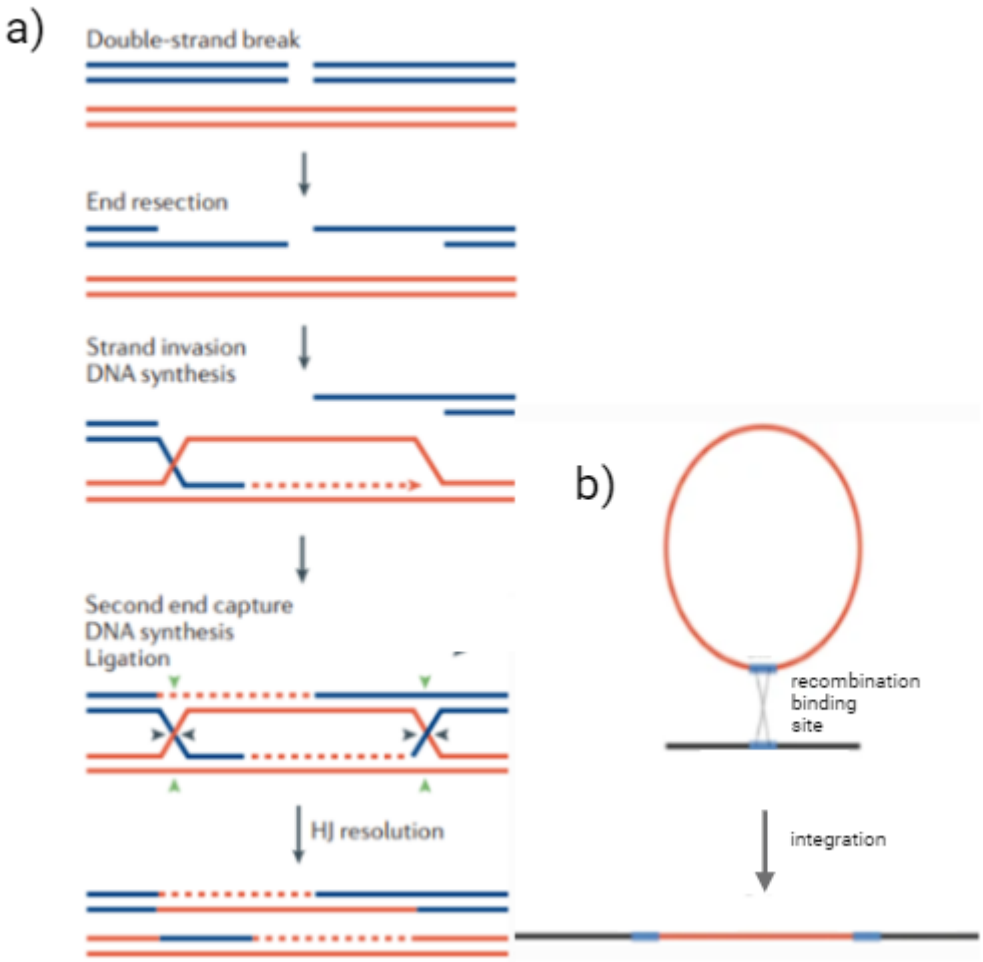


Fig. 2. The mechanism of recombination showing homologous recombination and conservative site-specific recombination. a) Homologous recombination where both strands are represented in a different colour. It starts with the formation of double stranded breaks which results in a holiday junction to exchange DNA and eventually resolution. b) Conservative site-specific recombination where the recombinase enzyme breaks the donor DNA strand to transfer it to the target molecule (adjusted from Parks & Peters (2018) with Biorender.com).

2.2.2 Transposable elements

Transposable elements are DNA sequences that relocate in the genome by site-specific transposition. Transposable elements occupy a large percentage of a species' genome as they have a repetitive nature and are hypothesised to have played a role in adjusting the genome during evolution by causing mutations during insertion. The recombination-induced interference in the transposition process leads to structural changes like excisions, insertions, translocations and duplications (Pimpinelli *et al.*, 1995). Due to deletions genomic DNA gets lost, but in case of duplications, the new

duplicated raw material can be used for evolutionary changes (Muñoz-López & García-Pérez, 2010). Furthermore, when the transposable element relocates, it may disrupt the coding region of a gene, which results in a loss of function (Kawakami *et al.*, 2017). Hence, both the host organism and transposable elements have devised strategies to reduce the harmful effects of transposition. For instance, certain transposable elements tend to integrate into genomic regions of heterochromatin, regions typically harbouring non-essential genetic elements and so have a minimal adverse impact. Additionally, these elements are less likely to be removed by natural selection, because they're inserted into parts of the genome that do not undergo significant evolutionary changes (Muñoz-López & García-Pérez, 2010; Pimpinelli *et al.*, 1995). As shown by Sinzelle *et al.* (2009) some transposable elements became domesticated in the genome and are transformed in a fully operational host gene. In terms of structure, these domesticated genes have no repetitive nature and do not exhibit typical molecular features associated with transposition such as terminal inverted repeats (TIR). This happened for example to the PIF/ Harbinger transposons identified in vertebrates (Sinzelle *et al.*, 2009).

Prokaryotes have different kinds of transposable elements, including DNA transposons and ISs, but do not have retrotransposons like eukaryotes. Transposable elements are present in both autonomous and non-autonomous forms. Autonomous copies contain all the required enzymes for self-mobility, while non-autonomous copies lack any coding ability, relying entirely on the enzymatic mechanisms of autonomous counterparts for mobility (Brown & Evans, 1991; Sinzelle *et al.*, 2009).

DNA transposons are characterized by the occurrence of short TIRs at their ends (Alamro *et al.*, 2021), of which the length is specific for every transposon family and by a gene encoding the transposase enzyme, located between the two TIRs and necessary for translocation of the transposon (Mcdonald, 1993; Muñoz-López & García-Pérez, 2010). Furthermore, compared to ISs, DNA transposons often carry additional functional genes, contributing to their genetic complexity and potential impact on genomic evolution (Benler *et al.*, 2021). The TIRs play an essential role as recognition sites for the transposase cleaving the donor DNA and facilitating strand transfer (Vandecraen *et al.*, 2017). DNA transposons incorporate themselves in locations on the chromosome or plasmid through non-homologous recombination

(Frost *et al.*, 2005) using either a replicative or non-replicative mechanism for transposition (Muñoz-López & García-Pérez, 2010). The replicative class of transposons consists of 9 superfamilies, while the non-replicative includes 2 families. This distinction arises from the fact that non-replicative transposons generate double-stranded breaks as part of their transposition mechanism, while the replicative mechanism makes single-stranded breaks (Sinzelle *et al.*, 2009). An **autonomous non-replicative DNA transposon** has a cut-and-paste mechanism, where the transposon shifts from one location to another, and as it moves, it creates an empty space (Muñoz-López & García-Pérez, 2010; Sinzelle *et al.*, 2009). These non-replicative DNA transposons comprise a transposase-coding region flanked on both ends by TIRs (Fig. 3) and are bordered by target site duplications (TSD), which are also specific to each transposon family. The non-replicative mechanism starts with the transcription of the transposase gene and binding to its recognition sequence at each end of the transposon. The transposase removes the transposon by cleaving DNA at the TIR ends after forming a synaptic complex. The transposase identifies a target site and inserts the transposon into the target DNA. The cellular DNA repair machinery seals the excision site of the donor strand and forms a transposon scar of varying lengths, unique for each transposon family (Sinzelle *et al.*, 2009).

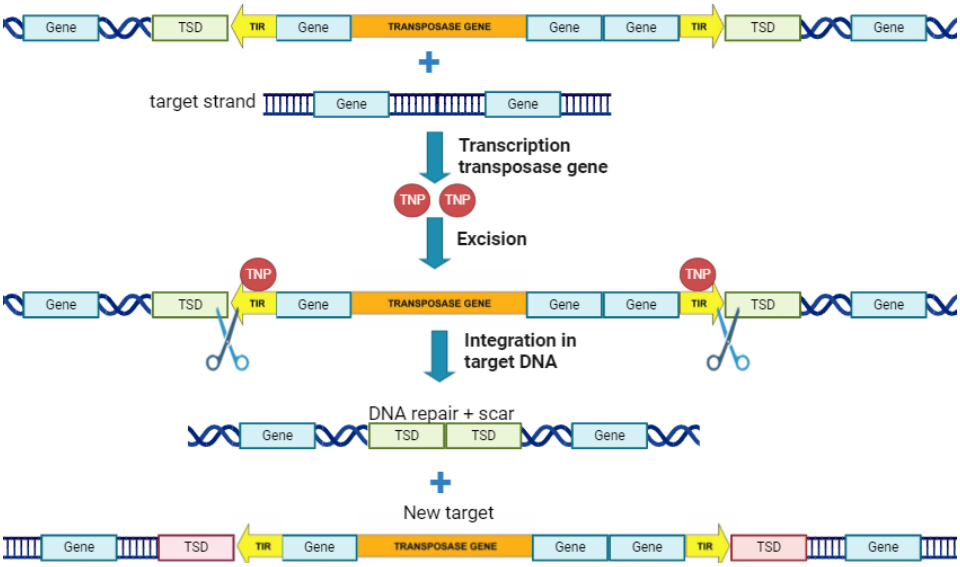


Fig. 3. The mechanism of non-replicative DNA transposition. In the process of transposition, the transposon relocates from one position to another, leaving a space behind as it moves. The transposon gene is coloured orange with the TIRs in yellow and functional genes in blue. The transposase (in red) cuts the transposon out of the donor DNA molecule (TSD in green) and inserts it into the target DNA molecule (TSD in pink) (created with Biorender.com).

Autonomous replicative DNA transposons use a paste-and-copy mechanism as displayed in Fig. 4. Overall, the transposon duplicates itself from the donor DNA and deposits a copy at the target DNA forming a cointegrate intermediate. Firstly, the transposase enzyme makes single stranded cuts in the donor strand at the level of the transposon and the target site DNA, which results in joining of the single stranded ends. Then, the complementary strand of the transposon is copied by the replicative machinery of the host, resulting in the formation of the cointegrate. Lastly, the resolvase enzyme resolves the donor DNA molecule from the target molecule at the level of the resolution sites making double stranded breaks, resulting in both the donor and target DNA having a copy of the transposon (Nicolas *et al.*, 2015).

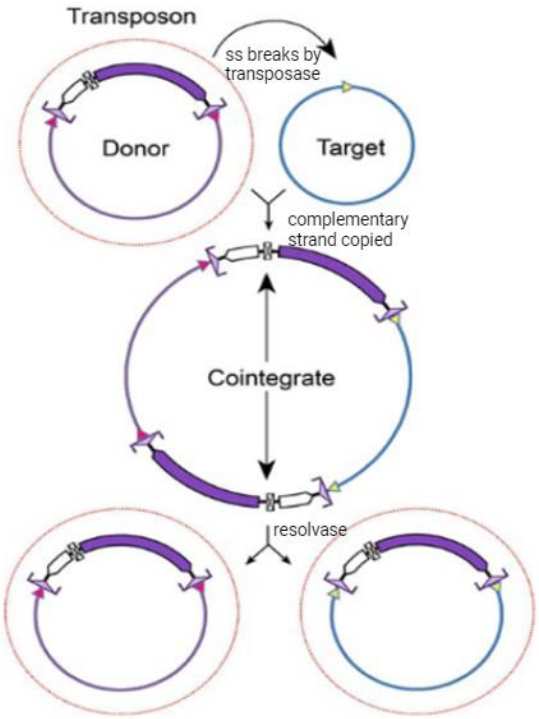


Fig. 4. Autonomous replicative mechanism of transposition. The transposon (in purple) replicates itself from the donor DNA and places a duplicate in the target DNA (in blue), creating a cointegrate intermediate. Then the resolvase resolves the donor from the target molecule at the copies of the transposon resolution sites indicated in yellow boxed crosses. The resolvase enzyme is indicated by the double arrow. The transposon's TIRs are indicated by bracketed triangles, while the TSDs that are produced upon insertion into the target are displayed as the small triangles (Nicolas *et al.*, 2015).

2.3 IS Elements

DNA transposons' most simple form is an IS (Vandecraen *et al.*, 2017). ISs can be found in prokaryotes like archaea and bacteria, and are not found in eukaryotes, although some eukaryotic DNA transposons share similarities with prokaryotic ISs (Kojima & Bao, 2023; Siguier *et al.*, 2015). ISs are compact genetic elements, often smaller than 3 kilobases between 0.7-2.5 kb, that autonomously move within the genome. They typically have short TIRs on either end and contain one or two open reading frames that code for proteins necessary for their mobility like transposase enzymes (Fig. 5) (Mahillon & Chandler, 1998; Siguier *et al.*, 2014). ISs are frequently accompanied by adjacent sequences known as TSDs, which are generated during transposition but are not part of the ISs. The length of these TSDs typically ranges from 2 to 14 base pairs (bp) with sequences that are specific for a particular IS element or IS family. These TSDs can serve as a genetic indicator or signature in the genome. Their presence or absence can help to identify which specific transposon was previously inserted into a DNA sequence when the transposon is no longer present (Siguier *et al.*, 2014, 2015).

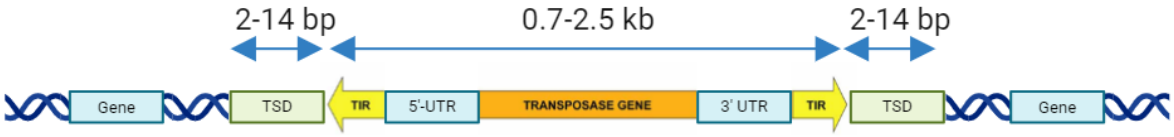


Fig. 5. The composition of an IS. The open reading frame encoding for the transposase is flanked by TIRs. During transposition TSDs are generated in the DNA flanking the IS (created with Biorender.com).

The primary distinguishing characteristic between IS elements and transposons is the absence of passenger genes in the IS. On the other hand, ISs can form a structure termed as CT as illustrated in Fig. 6. These elements share a genetic connection with IS elements by carrying extra (functional) passenger/cargo genes unrelated to the transposition flanked at both sites with a copy of the same IS element. The CT composed of the two IS copies flanking the cargo DNA, can relocate as a complete entity. In addition, the individual IS elements can move independently from the entire unit (Tansirichaiya *et al.*, 2016). Moreover, the CTs, when located on a conjugative or mobilizable plasmid, become prone to intercellular transfer between two cells within a bacterial community. The mobility obtained by conjugative plasmids emerges as a key

driver in the spread of CTs, significantly influencing the evolutionary paths and ability to adapt bacterial populations (Tansirichaiya *et al.*, 2016). An inverse correlation exists between the size of the intervening sequence and the transposition frequency of CTs, wherein a larger intervening sequence is associated with a lower transposition frequency (Nojiri *et al.*, 2004). As key players in shaping the genetics of bacterial communities, CTs represent the interplay between molecular mechanisms and microbial evolution (Bennett, 2008; Nojiri *et al.*, 2004).

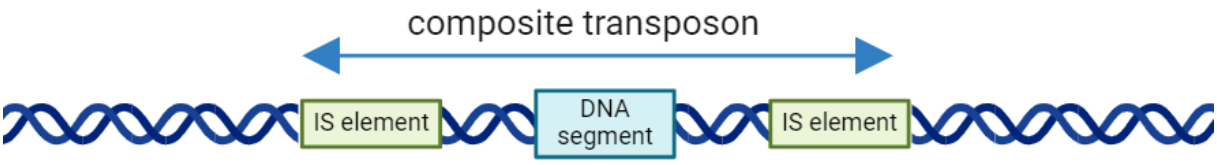


Fig. 6. Structure of a CT. The cargo/passenger genes (in blue) are flanked by an identical IS at both sites (created with Biorender.com).

IS elements are abundant and are categorised as the largest number of self-moving transposable elements within bacterial genomes. The process of identifying and categorizing ISs into families is achieved by considering a range of attributes such as their transposase amino acid sequences, their overall genetic structure, and the genes that certain ISs possess (Siguier *et al.*, 2014). The classification of IS elements is essential to manage the extensive variety and abundance of ISs. It not only aids in the recognition of the numerous IS sequences found in bacterial genomes, but also contributes to our comprehension of the impact on their host genomes and offers insights into their control and transposition mechanisms. The primary factor used for classification is the degree of amino acid sequence similarity of the transposase (Mahillon & Chandler, 1998; Siguier *et al.,* 2014). To identify and classify ISs, the ISfinder database is often used. This is a tool providing essential information on the distribution, classification, and characteristics of ISs in bacterial and archaeal genomes. It supports the identification, analysis, and annotation of these genetic elements, enhancing our understanding of their roles in prokaryotic evolution (Siguier *et al.,* 2006). The ISfinder online database comprises more than 4,000 recorded IS elements categorized into approximately 26 families, some of which can be further subdivided for better classification. This categorization is dynamic as new ISs are

constantly added (Siguier *et al.*, 2014). However, the ISfinder database remains incomplete because genomes are being sequenced more rapidly than they are fully annotated in the database (Robinson *et al.*, 2012). Notably, the quantity of ISs in a particular IS family according to ISfinder does not always indicate how common this particular IS occurs in the natural environment (Siguier *et al.*, 2015).

2.3.1 The mechanisms of IS transposition

As mentioned above, ISs can be classified based on their transposase. Varieties of transposases encompass DDE, DEDD, HUH, Tyrosine, and Serine transposases (Razavi *et al.*, 2020). These designations are indicative of the amino acids found within their catalytic sites and are associated with their transposition mechanisms. The majority of currently recognized IS elements are equipped with DDE transposases. D, E, H, and U represent the amino acids aspartic acid, glutamic acid, histidine, and a bulky hydrophobic amino acid, respectively (Siguier *et al.*, 2015). Since IS*6100*, the IS we focus on in this study, contains the *tnpA* gene encoding a DDE transposase, this paragraph will focus on the DDE transposition mechanism (Gai *et al.*, 2010; Nicolas *et al.*, 2015).

Transposition starts with the action of the transposase Tnp*A*, which belongs to the DDE/D superfamily of nucleotidyl transferases and is notably large, with approximately 1000 amino acids (Nicolas *et al.*, 2015). These enzymes share structural and catalytic similarities with RNaseH and other enzymes involved in nucleic acid processing (Rice *et al.*, 1996). The transposition mechanism for IS*6100* is replicative, i.e., as explained above, it duplicates itself and leaves a copy at the target site (Muñoz-López & García-Pérez, 2010; Nicolas *et al.*, 2015). The DDE/D transposase helps moving the DNA segments by binding specifically to the TIR to cut the DNA leading to the creation of free 3’-OH groups. By cutting only one strand, a cointegrate for the replicative transposition is created. To protect these free OH groups, the transposase covers them, and then a joining reaction integrates the transposon into another DNA sequence. Additionally, the DNA-binding domain of DDE/D transposases is crucial for sequence-specific recognition of terminal TIRs (Nesmelova & Hackett, 2010).

2.3.2 Impact of IS on bacterial community adaptation

ISs are widespread MGEs in bacteria. The long-term behaviour of IS elements in bacteria, as well as their influence on it, remain inadequately understood. This includes whether they primarily act as genomic parasites or serve as significant catalysts for adaptation through natural selection (Consuegra *et al.*, 2021). Due to recent research on transposons and ISs, it is portrayed that transposons and ISs are crucial components in prokaryotic adaptation (Siguier *et al.*, 2015). IS elements are responsible for approximately 35% of the mutations that became prevalent over 50,000 generations in populations of *Escherichia coli* while maintaining the original point-mutation rate of the ancestor (Consuegra *et al.*, 2021). The mutations they induce can have detrimental, neutral, or advantageous effects and generate diversity, which is necessary for the process of natural selection (Consuegra *et al.*, 2021; Wielgoss *et al.*, 2013). These mutations can be caused by random movement of ISs and can lead to an increased IS copy number. An example of a damaging mutation, as reported by Siguier *et al.* (2015), occurs when bacteria live inside a host rich in nutrients and do not need some of the genes crucial to live on their own. This results in occurrence of more harmful mutations that easily spread in the population. There are many cases of this phenomenon within intracellular endosymbionts. Meanwhile, other evolution experiments have pointed out that particular IS-mediated mutations can be beneficial. Treves *et al.* (1998) demonstrated evolutionary changes in *E. coli* populations, where IS insertions caused mutations in the regulatory region of the *acs* gene. These genetic changes resulted in overexpression of the *acs* gene, which improved the process of acetate scavenging. This is important in glucose-limited environments, where acetate can serve as an energy source.

Furthermore, one of the most remarkable attributes of ISs is their ability to expand within genomes and regulate the process of genome simplification or reduction by aiding in the removal of DNA. The ability of IS elements to create deletions is also likely to result in the eventual eradication, either entirely or partially, of the IS element themselves. This is believed to have occurred in ancient endosymbionts that no longer possess IS elements (Siguier *et al.*, 2015). Additionally, the insertion of genetic material into a gene almost certain disrupts its normal function, resulting in its inactivation or activation. However, an interesting result can occur when the inserted material

positions itself in front of a gene, in the promoter region, potentially leading to increased expression of the gene. The consequence of this alteration in gene activity, whether through inactivation or activation, can significantly impact the overall fitness of the host organism (Glansdorff *et al.*, 1981; Reif & Saedler, 1975; Vandecraen *et al.*, 2017). This not only in the gene region they insert into, but also activate or deactivate the transcription of neighbouring genes (Vandecraen *et al.*, 2017). They can achieve this through two different methods. The first involves the presence of internal promoters whose transcripts escape into adjacent DNA influencing the activity of nearby genes or secondly, the formation of hybrid promoters (Glansdorff *et al.*, 1981; Siguier *et al.*, 2015). One example of the transposition of an IS element that resulted in a beneficial effect is the involvement of IS*S12* in the development of resistance to toxic toluene in *Pseudomonas putida* S12. Insertion of an extra copy of IS*S12* in front of the *srpS* gene involved in toluene resistance, results into upregulation of the *srp* promoter activity and increased resistance (Wery *et al.*, 2001).

As ISs can form CTs, this aspect is particularly significant for bacterial community adaptation, as this way any gene or gene cluster can become incorporated into a MGE. Through IS mediated gene recruitment, bacteria can acquire additional functional genes, such as those associated with antibiotic resistance, xenobiotic degradation, or other catabolic processes (Bennett, 2008; Nojiri *et al.*, 2004; Tansirichaiya *et al.*, 2016). For example, the *tcbAB* gene cluster, which encodes chlorobenzene dioxygenase and dehydrogenase on plasmid p51 makes part of a CT designated as Tn*5280* featuring the same IS element positioned at both ends of the DNA fragment carrying the *tcbAB* gene cluster. In addition, plasmid pP51 carries a chlorocatechol gene cluster. It is hypothesized that the IS element on pP51 was involved in recruiting the *tcbAB* gene cluster by the plasmid as such combining *tcbAB* with a chlorocatechol catabolic gene cluster. This combination is suitable for the use of chlorobenzenes as a carbon source by the organism carrying pP51 (Van der Meer *et al.*, 1991). This confirms that new pathways arise through incorporation of catabolic genes or fragments of genes from various organisms into a compatible host involving IS element as a gene recruiter (Top & Springael, 2003). Another important insertion element linked with xenobiotic degradation is IS*1071*, especially for chlorinated aromatics and pesticide

biodegradation (Dunon *et al.*, 2013). It makes part of a CT containing genes responsible for the catabolism of various xenobiotic organic substances present in wastewater and agricultural soil, including herbicides like linuron, in many xenobiotic degrading bacteria, indicating that IS*1071* contributed in their recruitment. Moreover, IS*1071* associated CTs are often located on conjugative plasmids belonging to the highly promiscuous IncP-1 plasmid group (Dunon *et al.*, 2018). IncP-1 plasmids are extensively studied plasmids that show a broad host range enabling transfer between and recruitment of genes from various taxa (Dunon *et al.*, 2013; Nojiri *et al.*, 2004).

2.4 IS*6100*

IS*6100* is part of the IS*6* family, deriving its name from the ISs found within transposon Tn*6* (Siguier *et al.*, 2015). IS*6100*, with a size of 880 bp, is longer compared to other typical members of the IS*6* family. The IS*6* family transposases exhibit lengths spanning from 213 amino acids in IS*15* to 254 amino acids in IS*6100* (Varani *et al.*, 2021). All IS*6* family members possess similar highly conserved short TIRs of 14 to 20 bp, with IS*6100* having 14 bp and typically create 8 bp TSDs (Fig. 7) (Mahillon & Chandler, 1998). At this moment (March 2024), the IS*6* family includes approximately 161 members according to ISfinder (Siguier *et al.*, 2006; Varani *et al.*, 2021). IS*6* family elements are up to now only found in prokaryotes. They are widespread both in bacteria and archaea and have been found in approximately 80 species of bacteria and archaea. They encompass nearly all of the traditional archaeal lineages including methanogens, halophiles, thermoacidophiles, and hyperthermophiles, while bacterial phyla house IS*6* family elements such as *Actinobacteria, Proteobacteria,* and *Cyanobacteria* (Varani *et al.*, 2021). *Mycobacterium* taxa together with *Pseudomonas* are also established as the most common hosts for IS*6100* (Mahillon & Chandler, 1998). However, IS*6100* can be found in a wide range of host organisms suggesting that the element is crucial in spreading genes (Gai *et al.*, 2010).

Fig. 7. Overall structure of IS*6100*. The members of the IS*6* family have conserved TIRs flanking the transposase gene and encode for transposases with lengths ranging from 213 amino acids to 254 amino acids, with IS*6100* having 14 bp TIR and a transposase of 254 amino acids. The TSDs are usually 8 bp (created with Biorender.com).

The IS*6* family members translocate by a replicative transposon mechanism (Mahillon & Chandler, 1998). They produce cointegrates as explained in the section on transposable elements (Varani *et al.*, 2021). IS*6* family elements do not encode for a resolvase enzyme that resolves the donor DNA molecule from the target molecule as found in Tn*3*-related elements. It has been observed that the resolution of cointegrates mediated by IS*6* relies on a functional host *recA* gene in several organisms, indicating that it occurs through the host's own homologous recombination pathway (Brown *et al.*, 1984; Trieu-Cuot & Courvalin, 1985). There is limited knowledge regarding the regulation of transposase expression. In *Streptomyces lividans*, the ability of IS*6100* to transpose is greatly enhanced when positioned after a powerful promoter. Unlike other prokaryotic transposable elements, IS*6100* seems to lack efficient mechanisms to shield itself from external activation to transpose. This was proven by external transcription of the IS*6100* transposase gene by a copy of thiostrepton-inducible promoter ptipA, which is a genetically engineered promoter outside of the IS and can increase the transposition by factor 100 (Smith & Dyson, 1995).

2.4.1 Research on IS*6100* and its cargo genes for bacterial community adaptation

IS*6100* was initially obtained as a component of the complex transposon Tn*610* in *Mycobacterium fortuitum* (Smith & Dyson, 1995). The IS*6* group of bacterial and archaeal ISs, initially detected in the early 1980s, was at the outset particularly associated with antibiotic resistance genes, but it was subsequently discovered that they are also linked to CTs containing genes for catabolism of anthropogenic compounds (Kato *et al.*, 1994; Varani *et al.*, 2021). Typically, when forming CTs, the flanking IS elements can exist as either direct repeats being exact copies of the IS

element, or inverted copies, which are mirrored copies of the IS arranged in opposite orientations. Nonetheless, in the context of functional IS*6* CTs, the flanking IS elements are consistently observed as direct repeats (Varani *et al.*, 2021). Notably, IS*6100* has been found to play a role in the transfer of numerous functional genes, including on plasmid pTET3 carrying resistance genes to streptomycin, chinomycin and tetracycline, the *Pseudomonas aeruginosa* R1003 plasmid, and as part of the *Xanthomonas campestris* transposon Tn*5393b* associated with streptomycin resistance. This diverse range of functional genes suggests that IS*6100* facilitates HGT, allowing genetic material to move between different bacteria. This, in turn, provides bacteria with the ability to cope with specific environmental stresses (Mahillon & Chandler, 1998; Qin *et al.*, 2022; Varani *et al.*, 2021).

As the anthropogenic activities have a great impact on the environment, IS*6100* also helped adapting the bacteria for degradation of xenobiotics. Wei *et al.* (2009) found that the Tn*mph* CT contains the *mph* gene flanked by 2 identical copies of IS*6100*. This gene has a role in organophosphorus degradation and can integrate in different genomic sites by the help of the ISs. Another study revealed that IS*6100* flanked the gene that degrades carbazole, a compound recognised for its toxic and mutagenic properties, in P*seudomonas* and *Sphingomonas* bacteria (Gai *et al.*, 2010). Additionally, IS*6100* plays a role in nylon degradation as it was found alongside nylon degrading genes on plasmid pOAD2 in *Flavobacterium* strain K172 and on plasmid pNAD2 in the *Pseudomonas sp.* strain NK87 (Kato *et al.*, 1994). Furthermore, IS*6100* has been identified as carrying along the *linA* gene, responsible for degrading the chlorinated insecticide hexachlorocyclohexane (Dogra *et al.*, 2004). The hexachlorocyclohexane catabolic *Sphingomonas haloaromaticamans* MM1 contains 7 *lin* genes that together encode a complete pathway for hexachlorocyclohexane catabolism. The *lin* genes are dispersed across several plasmids in *S. haloaromaticamans* MM1 and on these plasmids they are consistently associated with IS*6100* forming multiple apparent CT structures. The consistent association of IS*6100* with the plasmids makes this bacterium an excellent reference for IS*6100* research (Tabata *et al.*, 2011). IS*6100* appears also to have been involved in patchwork assembly of suitable genes to create a pathway for the catabolism of 2-chloronitrobenzene, an intermediate compound in pharmaceuticals, dyes, and

pesticides, in *Pseudomonas stutzeri* ZWLR2-1 (Liu *et al.*, 2011). Moreover, IS*6100* often flanks the conserved *ceh* and *cfd* gene cluster, responsible for the hydrolysis of the carbamate insecticide carbofuran, in various carbofuran degrading bacteria, indicating its role in HGT in distributing the carbofuran catabolic genes (Jiang *et al.*, 2022).

2.4.2 Research methodologies related to IS*6100*

Various methods exist for examining the ecology of IS elements. Most research is based on the genetic analysis of cultured bacterial strains and hence depends on cultivation using selective media to isolate the strain of interest. The isolated strains are subjected to whole genome sequencing, to determine whether MGEs, including ISs, are present and whether they are linked to particular functional genes (Cerbino *et al.*, 2023). Nonetheless, it is crucial to recognize that not all bacteria can be cultured, that important microorganisms may go unnoticed, and this method is constrained to capture only a small portion of the overall microbial diversity. Indeed, the optimal growth conditions for the majority of environmental bacteria cannot be replicated in laboratory environments, leading to their undetectability (Lewis *et al.*, 2021). There is limited knowledge regarding the presence of IS*6100* as well as its cargo genes recruited in CT structures in environmental settings and complex communities. To obtain a comprehensive community-level understanding of the abundance of IS*6100* and of IS*6100* cargo genes in CT it is necessary to employ cultivation-independent analytical techniques where total community DNA is directly extracted from environmental samples. The integration of LR-PCR with high-throughput sequencing of amplicons enables the detection of cargo genes sequence information, which is particularly valuable for microbial ecosystems characterised by a high diversity, where the relevant bacteria carrying IS*6100* is likely only a small fraction (Dunon *et al.*, 2022). Dunon *et al.* (2018) developed a LR-PCR technique for IS*1071* and this method was optimised for IS*6100* with an amplicon threshold of 25 kb, exceeding the 5 kb threshold of conventional PCR (Mawarda and Springael, unpublished results The LR-PCR technique employs primers designed to bind to the *tnpA* gene of IS*6100*, with primer sequences oriented outwards from the IS element, enabling the amplification of the DNA segment located between two IS*6100* elements and hence the cargo of IS*6100*

associated CTs. Dunon *et al.* (2013) has successfully performed LR-PCR on IS*1071* with the same strategy and this strategy has already been confirmed to work on IS*6100* by Fuchu *et al.* (2008). They were able to confirm the presence of *lin* genes for hexachlorocyclohexane degradation flanked by IS*6100*, by performing PCR followed by sequencing (Fuchu *et al.*, 2008).

After that the LR-PCR amplicons are obtained, sequencing analysis is used to identify the cargo genes. This targeted metagenomics concept relies on studying genomes found in a diverse community of organism (Zhang *et al.*, 2021). For sequencing analysis, nanopore sequencing suits for long reads (tens of kilobases or even longer), which make the assembly easier and spans much longer repeat regions, while Illumina sequencing is better for short (maximally 300 bp), but more accurate reads, followed by bioinformatics to explore the accessory genes (Giani *et al.*, 2020). For function annotation a gene is compared with a specific functional database to comprehend the relevant metabolic pathway, like the NCBI database. However, a drawback is its inability to identify unknown genes with entirely distinct sequences from those already present in existing databases (Dunon *et al.*, 2018; Zhang *et al.*, 2021). Illumina sequencing is also performed to obtain the microbial community diversity by targeting the conserved 16S rRNA gene (Degnan & Ochman, 2011).

For precise quantification of the IS abundance, real-time PCR is necessary. While numerous techniques exist for the quantification of nucleic acids, real-time qPCR currently stands out as the most sensitive and accurate method. In the past a common approach involved hybridization methods like southern blotting, however qPCR is faster, needs less input DNA and doesn't require post-PCR manipulations resulting in less contamination risk (Hoebeeck *et al.*, 2007). Although it remains impossible to determine whether the cells containing the sequences are alive or dead (Cangelosi & Meschke, 2014). Notably, Dunon *et al.* (2018) have successfully performed the qPCR method on IS*1071* for abundance determination and this protocol was optimised for IS*6100* (Mawarda and Springael, unpublished results). It also requires the use of *tnpA* genes of IS*6100* for primer design. The measurement happens by the detection of fluorescence emitted by the SYBR Green dye bound to dsDNA and a melting curve analysis is carried out to avoid non-specific PCR products (Lee *et al.*, 2006).

3 Materials and methods

3.1 Environmental samples

For our study, 66 Pakistani soil samples and 73 Belgian soil samples were taken from various areas in the respective countries. The Belgian samples were collected or originated from distinct locations in Belgium and represented soils that were subjected to different xenobiotic exposures, e.g. pesticides, mineral oils, PFAS, and Zinc (Zn). In Table 1, the soil samples are displayed with the location where they were originally from, indicating the major pollutant they were exposed to and the sampling date at the contaminated site. Soil samples TBT, Vlarema VCL, VOCL, PFAS, MO4, MO5, MO7, MO8 and Zn were collected in April 2023 from soil heaps, stored at a remediation company and were awaiting soil cleaning. Samples UB, SB and LB were samples taken from three operational on farm biopurification systems (BPS) at Inagro, Rumbeke. Inagro is a research and advisory centre to enhance the competitiveness and sustainability of farming practices in Flanders. The BPS are used for treating pesticide contaminated wastewater, derived from pesticide spraying equipment. Most of the other soil samples were taken from agricultural fields with cultivation of various crops that had been treated with pesticides in the last month before sampling (Table 1), this at 4 different locations (Rumbeke, Kortemark, Zwevezele and Koksijde) in West-Flanders. Finally, sand samples were taken from the beach of Koksijde. From each soil, 3 replicate samples were taken. In case of the soil heaps, replicates were taken at different positions of the heaps, at least at 1 m distance from each other. From the BPS, replicate samples were taken 50 cm to 1 m away from each other, depending on the size of the BPS. In case of the agricultural fields and beach, replicates were taken 40 to 80 m from each other. Samples were always taken in a sterile way from the top 10 cm and collected in plastic containers, which had been sterilized using 70% ethanol. The samples were stored in the laboratory at 4°C and within 5 days, sieved and homogenised through a 2 mm sieve. Subsequently, the sieved soils were stored at 4°C until further analyses.

Table 1. Overview of the sampled soils from Belgium. *In case of soil samples TBT, Vlarema VCL, VOCL, PFAS, MO4, MO5, MO7, MO8 and Zn the location is the original location were the soils originated from before being stored at the remediation farm, while the date represents the date at which the soils were collected from the contaminated site before being stored at the remediation farm.

	Major pollutant	Location*	Date*
TBT	Tributyltin	Antwerp	Feb 2023
Vlarema VLR	Diverse	Mixture of different locations	Different dates
VOCL	VOCL	Ronse	Jan 2023
MO4	Mineral oil	Affligem	Apr 2021
MO5	Mineral oil	Bottelare	Sep 2021
PFAS	PFAS	Baasrode	Sep 2022
MO7	Mineral oil	Gistel	May 2023
Zn	Zn	Oudenburg	Dec 2022
MO8	Mineral oil	?	Feb 2023
Inagro BPS UB	Pesticides	Rumbeke	Apr 2023
Inagro BPS LB	Pesticides	Rumbeke	Apr 2023
Inagro BPS SB	Pesticides	Rumbeke	Apr 2023
Inagro spinach field SF	Pesticides	Rumbeke	Apr 2023
Inagro oinion field OF	Pesticides	Rumbeke	Apr 2023
Field farmer 1 Potato 1PF	Pesticides	Kortemark	Apr 2023
Field farmer 1 Corn 1CF	Pesticides	Kortemark	Apr 2023
Field farmer 1 wheat 1WF	Pesticides	Kortemark	Apr 2023
Field farmer 2 Potato 2PF	Pesticides	Zwevezele	Apr 2023
Field farmer 2 Sugarbeet 2SF	Pesticides	Zwevezele	Apr 2023
Field farmer 3 Wheat 31W	Pesticides	Koksijde	Apr 2023
Field farmer 3 Wheat 32W	Pesticides	Koksijde	Apr 2023
Field farmer 3 Potato 3PF	Pesticides	Koksijde	Apr 2023
Beach BS	?	Oost-Duinkerke	Apr 2023
Field Inagro IN	Pesticides	Rumbeke	July 2023

The Pakistani soil samples were collected in January 2022 from 22 different cities spanning the North Punjab province, extending from Lahore (31.5196° N, 74.3889° E) to Islamabad (33.6844° N, 73.0479° E). Samples were taken along the historic Grand Trunk Road, dating back to the 16th century and playing a major transportation role

connecting various important cities. To this day, the road is utilized for commercial and personal travel alongside industrial and agricultural sectors and spans a distance of 400 km from Lahore in the south to Islamabad in the north. Samples along this road were taken at different latitudes as shown in Fig. 8 near the cities of Chaklala, Dina, Gakhar, Gujar Khan, Gujranwala, Gujrat, Islamabad, Jehlum, Kala Shah Kaku, Kamoki, Kharian, Lahore Cantt, Lahore City, Lalamusa, Mandra, Muridke, Rawalpindi, Rawat, Sara e alamgir, Shahdara, Sohawa and Wazirabad. Near each city, samples were taken from an agricultural field, near the roadside, and from an industrial area. The spatial information and land use of every Pakistani sample is displayed in Table A.2. The area experiences both intense summer and moderate winter conditions, accompanied by limited and irregular rainfall. As the samples were taken in the coldest month, the temperature ranged from around 10 to 20°C in the daytime and 0 to 10°C at night. The soil samples were topsoil, collected from a depth of 0-20 cm. Upon collection, the samples were sieved and homogenised using a 2 mm sieve, and subsequently stored at 4ºC until further analysis in the laboratory.

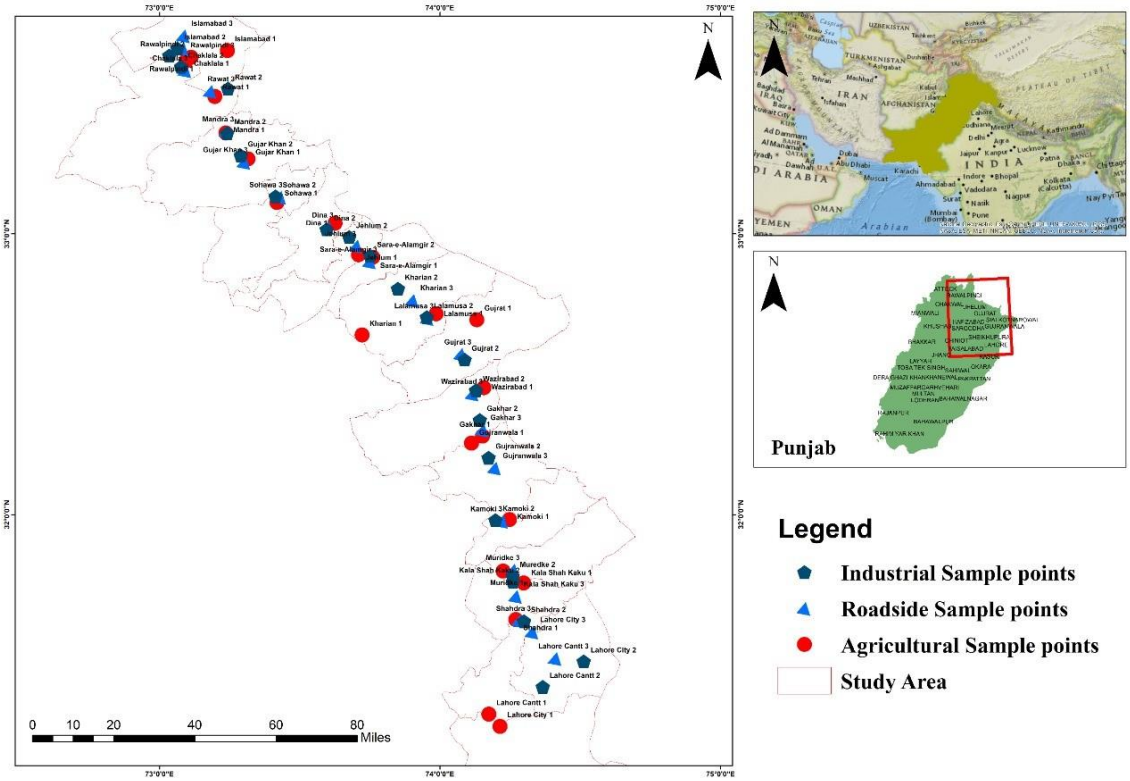


Fig. 8. Soil samples collected in Pakistan near cities located along the Grand Trunk Road spanning a distance of 400 km from Lahore in the south to Islamabad in the north. In each of the 22 cities one soil sample was collected from an industrial, roadside and agricultural area.

3.2 Characterization of the soil samples

For the Belgian soil samples the soil physico-chemical analyses i.e. moisture content, nitrate and ammonium, ICP-MS, pH, Olsen P, mineral N, texture, total C and N content are yet to be determined and measured. Since this data will not further be used in this thesis, they are not included. The chemical characteristics of the Pakistani soil samples were determined by the research group of Quaid-i-Azam University in Islamabad, Pakistan. The measured parameters are pH, electrical conductivity (EC), total nitrogen, polychlorinated biphenyl (PCB) concentrations (dichlorobiphenyl (DiPCB), TriPCB, TetraPCB, PentaPCB, HexaPCB, HeptaPCB), Cr, As, Cu, Pb, Cd and Total Petroleum Hydrocarbons (TPH), as displayed in Table A.1 in the appendix. For determining PCB concentrations, soil samples were extracted by solid-liquid extraction in a Soxhlet apparatus using dichloromethane for 24h, followed by concentration in a rotary evaporator and solvent exchange to *n*-hexane. The extracts underwent cleanup using an alumina-silica column, followed by concentration under ultrapure nitrogen gas. PCB analysis was performed by GC-MS/MS using an Agilent 7890 Gas Chromatograph equipped with a CP-Sil 8 CB capillary column and combined with an Agilent 7000 A Triple Quadrupole MS/MS quantifying 32 PCB congeners with specific settings. Next, pH and EC were measured using a pH-conductivity meter while total nitrogen and carbon were analyzed using an elemental analyzer. By using a Laser Diffraction Particle Size Analyzer, the soil texture was determined. Concentrations of total and available heavy metals i.e. Cd, Cu, Cr, Pb, and As were determined using inductively coupled plasma mass spectrometry (ICP-MS, Agilent Technologies). Furthermore, TPHs in soil were determined following the US EPA 3550 methodology, where the soil was ultrasonically extracted with hexane and Na_2SO_4, and the extract filtered and centrifuged. The TPHs were determined gravimetrically after hexane evaporation.

3.3 Used bacteria and growth conditions

S. haloaromaticamans MM1 (*JCM 19685 - JCM Catalogue*, 2007) was used to obtain plasmids that served as a reference for the LR-PCR targeting the cargo of IS*6100* associated CTs. Additionally, the IS*6100* gene in the plasmid is used as a template to generate the standard for the qPCR targeting IS*6100*. DNA from *Aminobacter niigataensis* MSH1 (Nielsen *et al.*, 2021) was used as a standard for qPCR targeting

the 16S rRNA gene. *S. haloaromaticamans* MM1 was obtained from Prof. Y. Nagata (University of Tokyo, Tokyo, Japan) and contains 3 plasmids, i.e., pSP1, pSP3 and pSP4 carrying various copies of IS*6100* as borders of presumed CTs (Tabata *et al.*, 2011) and as such, forming targets of the LR-PCR for the CTs flanked by IS*6100*, as shown in Table 2. Strain *S. haloaromaticamans* MM1 was cultured in 50 mL NB broth in a 100 mL Erlenmeyer at 37°C for 3 days. *A. niigataensis* MSH1 was similarly cultured, but in 50 mL LB broth. The bands that were successfully amplified by the LR-PCR method are indicated with an asterix in Table 2.

Table 2. Target regions on plasmids pSP4, pSP1 and pSP3 of the bacterial reference *S. haloaromaticamans* MM1 for the LR-PCR targeting the cargo of IS*6100* flanked CTs. An asterix indicates the bands that were successfully amplified by LR-PCR.

Plasmid size	Size target regions
Plasmid pSP4 (33kb)	2.6 kb
	3.06 kb
	4 kb*
	1.27kb
	5 kb*
Plasmid pSP1 (172 kb)	2.8 kb
	4.4 kb
	6.9 kb
	20 kb*
	107 kb
	3.2 kb
	2.6 kb
	10 kb
Plasmid pSP3 (54 kb)	12.3 kb*

3.4 Plasmid extraction

Following the manufacturer's instructions, the Plasmid Mini Kit (QIAGEN®) was used to extract the plasmids from S. *haloaromaticamans* MM1. The plasmid DNA was finally dissolved in 25 µL Tris-EDTA (TE) buffer and the DNA concentration was measured using the Qubit 4 Fluorometer (Thermo Fisher Scientific, Belgium) broad range assay kit for double-stranded DNA. The plasmid DNA was stored at -20°C until use.

3.5 High-molecular-weight (HMW) environmental DNA (eDNA) extraction from soil samples

HMW eDNA for LR-PCR purposes was extracted from soil using a two-step procedure. In the first step, cells were extracted from the soil using a density-gradient centrifugation method protocol adapted from Pratscher (2021). In the second part, DNA is extracted from the cells following the MasterPure™ DNA Purification Kit (Biosearch Technologies, UK). To start, 3 grams of soil was resuspended in 3 mL phosphate buffered saline (PBS) with 0.1% tween 20 solution (>40% GC) in a 15 mL Falcon tube. This solution was made by mixing 10 µL of Tween-20 in 10 mL of PBS solution, which corresponds to a final concentration of 0.1% Tween-20 in the PBS solution used to resuspend the soil. The suspensions were vigorously vortexed at maximum speed of 2600 rpm for 2 min at room temperature, followed by sonication at 20 kHz inside a water bath for 8 min at room temperature. After centrifugation (Rotina 420R, Hettich zentrifugen, Germany) at 700 g at 4°C for 1.5 min, approximately 1.4 mL of the supernatant was carefully transferred to a 2 mL Eppendorf tube. The suspension was centrifuged (MIKRO 200R, Hettich zentrifugen, Germany) at 250 g at 4°C for 1 min and 1.4 mL of the supernatant was again carefully transferred to new 2 mL Eppendorf tube. Subsequently, 0.4 mL of 80% (w/v) Histodenz solution was gently injected at the bottom of the tube using a plastic syringe with a long needle without mixing the layers. The tubes were then centrifuged at 4°C at 14000 g for 40 min with low acceleration and low deceleration to separate microbial cells from soil particles. Following centrifugation, the tube exhibited four distinct phases, as depicted in Fig. 9. A thin cloudy layer containing the cells, due to their lower density, floated on top of the Histodenz phase, but stayed below the low-density PBS phase. This cloudy layer was carefully collected into a 2 mL Eppendorf tube while avoiding taking the Histodenz phase. The cell suspension was centrifuged for 5 min at 20000 g at 4°C and the supernatant was discarded. If the pellet was high showing a white compact mass of material clearly observed by the eye, the pellet was washed three times to remove any residual Histodenz in 700 µL PBS by briefly pipetting up and down and centrifuging for 5 min at 20000 g at 4°C. Otherwise the washing steps were skipped if the pellet was too small and barely noticeable. The supernatant was discarded, and if not directly used for DNA extraction, stored in 200 µL PBS buffer at -20°C. When the stored DNA

was subsequently required for the proceeding DNA extraction, the sample was taken out of the freezer and centrifuged 5 min at 20000 g at 4°C, whereafter the supernatant was also discarded and ready for the further steps.

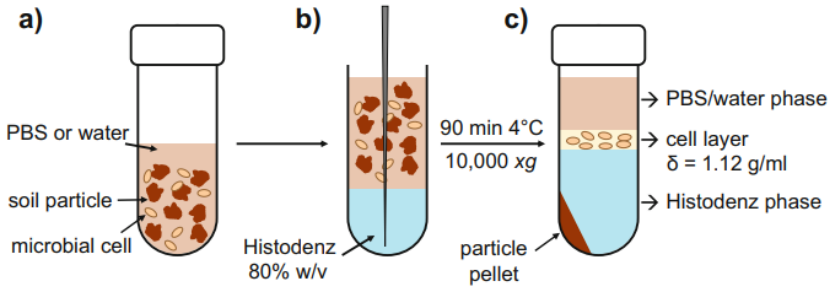


Fig. 9. Isolation of microbial cells from environmental samples using density-gradient centrifugation technique with Histodenz as taken from Pratscher (2021). a) The homogenized environmental sample suspension. b) Suspension floating on top of the injected Histodenz solution. c) Phases observed in the sample following gradient centrifugation.

DNA from the cells was extracted using the MasterPure™ DNA Purification Kit (Biosearch Technologies, UK) according to the manufacturer's recommendations with some modifications. The kit contains all solutions required for the HMW eDNA extraction including Metaphosphoric acid (MPC) protein precipitation reagent, tissue and 2X T&C cell lysis solution, lysozyme, TE buffer, Proteinase K and RNase A. The first steps were changed to cell lysis initiated by resuspending (by pipetting up and down) the cell pellet in 290 µL degrading solution. This degrading solution contained 300 µL TE buffer, 2 µL lysozyme and 1 µL RNase A. The samples were put in the incubator at 37°C for 30 min. Afterwards, 295 µL of the second degrading solution, which contained 300 µL tissue and 2X T&C cell lysis solution and 1 µL Proteinase K (50 µg/ml), was added to each tube and mixed with wide-bore pipette tips. Each tube was then incubated at 65°C for 15 min. An additional modification after adding 350 µL MPC protein precipitation reagent, was mixing 10 times with wide-bore pipette tips. Additionally, thereafter the supernatant was transferred to 1.5 mL Eppendorf tubes with wide-bore pipette tips. Next, additional steps were included after adding 500 µL of 70% ethanol, the tube was centrifuged again at 20000 g for 5 min at 4°C. After removing the supernatant, the pellet was spun down (Mini star centrifuge, U.S.) at 2000 g and all the residual supernatant was discarded. Lastly, the pellet was air-dried for 5 min or longer until all liquid was removed. This pellet was dissolved in 25 µL TE buffer by gently

tapping the tube. To dissolve the DNA even better, the tubes were spun down and placed in the incubator at 50°C for 10 min. The DNA concentration was quantified using the Qubit 4 Fluorometer (Thermo Fisher Scientific, Belgium) broad range assay kit for double stranded DNA, after which the samples were used for LR-PCR. It is advisable to conduct the LR-PCR without delay to minimize the risk of DNA degradation. While storing the samples at -20°C is feasible, it may have adverse effects on the final results.

3.6 LR-PCR targeting the cargo genes of IS*6100* associated CTs

LR-PCR targeting the cargo of IS*6100* associated CTs was performed in a total volume of 20 µL, containing 6.2 µL DNA template and 13.8 µL master mix. The LR-PCR used specific primers that bind specifically to the IS*6100 tnpA* gene and that are oriented outward from the IS element allowing the amplification of the DNA region situated between two IS*6100* elements. The master mix for 1 reaction contained 10 µL of Platinum™ SuperFi II Green PCR Master Mix (Thermo Fisher Scientific, Vilnius), 0.4 µL of 10 µM primer IS*6100*-F (5'-CGGCAGGTGAAGTATCTCAA-3') and 0.4 µL of 10 µM IS*6100*-R (5'-GGTACAGGTAGGTCCACTTG-3'), 0.2 µL 1% bovine serum albumin (BSA), and 0.8 µL dimethyl sulfoxide (100 v/v). The optimized LR-PCR program included an initial denaturation step of 2 min at 98°C, 35 cycles of 10 sec of denaturation at 98°C and elongation of 15 min at 72°C and a final extension of 5 min at 72°C. The lengths of the PCR products were visualized by Agarose gel electrophoresis (AGE) (0.8%) containing 1x GelRed™ (Biotium, U.S.). The gel was run in TE buffer (pH 8.3) for 1 hour 30 min at 90 mV and 400mA including a 1 kb ladder (GeneRuler 1 kb plus, Thermo Fisher Scientific, Vilnius) with a size range of 75 bp to 20000 bp on agarose gels. Electropherograms were pictured using the InGenius SynGene Bio Imaging System (model LW135M, SynGene, India) with the GeneSnap software (SynGene, India). The LR-PCR products were stored at -20°C until purification for nanopore sequencing.

3.7 Purification of the LR-PCR amplicons using AMPureXP beads and nanopore sequencing

Purification of the LR-PCR products was performed with AMPureXP beads (Beckman Coulter, U.S.) following the manufacturer's recommendations, although some adjustments were made to increase the DNA concentration. Multiple LR-PCR

amplicons of the same sample were pooled together as shown in Table 3 to increase the concentration during purification. The AMPureXP beads were added to the pooled samples solution containing the LR-PCR amplicons in a ratio of 1:1 (v/v) depending on the pooled samples' volume. The beads were washed 3 times in 175 µL or more depending on the amount of pellet that had to be submerged, in 80% ethanol. In the end, 28 µL of the supernatant was transferred to a new Eppendorf tube. The DNA concentration was analysed by the Qubit 4 Fluorometer (Thermo Fisher Scientific, Belgium) with a broad-range assay kit, designed for double-stranded DNA. The purified amplicons were stored at -20°C until nanopore sequencing. Nanopore sequencing was performed using the portable MinION device (MN21108, Oxford Nanopore Technologies, Oxford, UK) following the manufacturer's protocols (Oxford Nanopore Technologies, Oxford, UK) at the Laboratory of Gene Technology of Prof. Rob Lavigne. The library preparation of the LR-PCR amplicons was performed using the Ligation sequencing amplicon-native barcoding kit 24 V14 (SQK-NKBD114.24) following the manufacturer's instructions. Subsequently, the library was loaded onto the MinION flow cell (FLO-MIN114) and subjected to sequencing for 72 hours using the 400 bp fast base calling model v4.2.0 with minKNOW v23.22.7 software. The Nanoplot bioinformatics tool was utilized to generate histograms illustrating the read length distribution per sample, facilitating initial comparison of the sequencing outcomes with the AGE results.

Table 3. Amount of pooled samples for purification with AMPureXP beads and their concentrations. The samples were pooled after running LR-PCR in a 20 µL of PCR reaction mixture of which 4 µL was used to visualize the amplicons on AGE. HMW eDNA extraction resulted into a 20 µL DNA extract and for each LR-PCR 6 µL of the eDNA solution was used as template in the LR-PCR, i.e., three reactions per HMW eDNA extract. For example for TBT.1 soil, two eDNA extractions were performed allowing 6 LR-PCRs. *concentration was measured after eDNA extraction and before the LR-PCR and was calculated by adding together the concentrations of each replicate per soil sample

	Amount pooled (16 µL per sample)	Concentration DNA in the extracts before LR-PCR and purification* (ng/µL)	Concentration DNA in amplicons after LR-PCR and purification (ng/µL)
TBT.1	6	6.1	226
PFAS.1	11	28.56	160.2
MO7.1	6	34.4	312
Inagro BPS UB.1	6	44.22	326
Inagro BPS LB.1	6	14.72	278
Inagro BPS SB.1	9	129.6	246
Reference	10	172.6	516
Field farmer 1 Potato 1PF.2	6	26.5	256

3.8 Real-time quantitative PCR

The abundance of the bacterial 16S rRNA gene and IS*6100* in the eDNA extract of the soil samples was determined by SYBR green based real-time qPCR on a Rotor-Gene model RG-3000 centrifugal real-time cycler (Corbett Research, Australia). Due to the potential impact of inhibitors present in DNA extracts on the qPCR outcomes, qPCR was performed on 10-fold and 100-fold diluted extracts of the DNA extracts. As standard for the 16S rRNA gene qPCR, the almost complete 16S rRNA gene of *A. niigataensis* MSH1 was amplified from genomic DNA using primer set 27F (5'-AGAGTTTGATCMTGGCTCAG-3') and 1493R (5'-ACGGCTACCTTGTTACGACTT-3'). The master mix had a reaction volume of 25 µL with 18.875 µL nuclease-free water, 2.5 µL 10x Dreamtaq Green Buffer (Thermo Fisher Scientific, Vilnius), 1.25 µL 1% BSA, 0.5 µL 10 mM dNTP mix (Thermo Fisher Scientific, Vilnius), 0.75 µL of each primer, 0.125 µL Dreamtaq DNA polymerase (Thermo Fisher Scientific, Vilnius) and 2.5 µL template DNA. The PCR program started with an initial denaturation lasting 5 min at 95°C, followed by 30 cycles. Each cycle involved denaturation at 95°C for 1 min, annealing at 52°C for 1 min, elongation at 72°C for 1 min 30 sec, and a final elongation phase of 8 min at 72°C.

As a standard for the qPCR targeting IS*6100 tnpA*, a conserved 100 bp region of IS*6100* on a plasmid of *S. haloaromaticamans* MM1 was amplified using primer qPCR-IS*6100*-Forward (5'-CCGATCACGGAAAGCTCAAG-3') and qPCR-IS*6100*-Reverse (5'-GGCTCGCATGACTTCGAATC-3'). The master mix had a reaction volume of 25 µL and consisted of 2µL template DNA, 15.625 µL nuclease-free water, 2.5 µL 10x Dreamtaq Green Buffer (Thermo Fisher Scientific, Vilnius), 0.25 µL 1% BSA, 2 µL 10mM dNTP mix (Thermo Fisher Scientific, Vilnius), 1.25 µL of each primer, 0.125 µL Dreamtaq DNA polymerase (Thermo Fisher Scientific, Vilnius). The PCR program started with an initial denaturation at 95°C for 10 min. Then, 40 cycles consisting of denaturation at 95°C for 30 sec, annealing at 63°C for 35 sec, elongation at 72°C for 1 min, and a final elongation phase of 4 min at 72°C. The TOPO® TA Cloning® Kit (Invitrogen by Thermo Fisher Scientific, Belgium) was used following manufacturer's instruction to insert the PCR product in a plasmid vector. Next, the cloned PCR product inserted into a plasmid vector was stored at -80°C for further use as qPCR IS*6100* standard.

The qPCR mixture for actual 16S rRNA gene quantification consisted of 3 µL of template DNA, 7.5 µL of Absolute qPCR SYBR Green Mix 2X (Thermo Fisher Scientific, Vilnius), 0.45 µM of each primer 27F and 1493R, and 3.6 µL nuclease-free water with a 15 µL reaction volume in a PCR tube. The qPCR program for the 16S rRNA comprised an initial denaturation step of 15 min at 95°C, followed by 40 cycles of 10 sec denaturation at 95°C, 10 sec annealing at 55°C, and 15 sec elongations at 72°C, with a final elongation step of 45 sec at 72°C. The qPCR mixture targeting the IS*6100 tnpA* gene included 3 µL of template DNA, 7.5 µL of Absolute qPCR SYBR Green Mix 2X (Thermo Fisher Scientific), 0.45 µM of each primer qPCR-IS*6100*-Forward (5'-CCGATCACGGAAAGCTCAAG-3') and qPCR-IS*6100*-Reverse (5'-GGCTCGCATGACTTCGAATC-3'), and 3.6 µL nuclease-free water with a 15 µL reaction volume in a PCR tube. The qPCR program for the *tnpA* gene was identical to the program used for quantifying the 16S rRNA gene.

Standard curves using the partial *A. niigataensis* MSH1 16S rRNA gene in case of 16S rRNA gene quantification and the IS*6100 tnpA* standard in case of IS*6100* quantification were generated by performing qPCR on tenfold dilution series (ranging from 10^9 to 10^1 copies/µL) of the amplicons for 16S rRNA and linearized plasmids for the IS*6100 tnpA* gene containing the specific gene of interest. Only qPCR values not affected by inhibition were retained for further use, recognized by comparing the outcome of the differentially diluted samples. The number of IS*6100* gene copies determined by qPCR were normalized by dividing the value with the number of 16S rRNA gene copies and normalized and expressed by the log10 value.

3.9 16S rRNA gene amplicon sequencing

After using the DNeasy Power Soil Kit (Qiagen) to extract total eDNA from the soil samples in accordance with the manufacturer's instructions, 16S rRNA gene amplicon sequencing was carried out for analysing the diversity and structure of the bacterial community. The targeted bacterial 16S rRNA gene region was the hypervariable V3-V4 region, amplified using primers 338F (5'ACTCCTACGGGAGGCAGCAG3't) and 806R (5' GGACTACHVGGGTWTCTAAT-3'). The PCR program used an initial phase of denaturation lasting 3 min at 95°C, followed by 25 cycles consisting of denaturation at 95°C for 30 sec, annealing at 55°C for 30 sec, elongation at 72°C for 30 sec, and finally, a concluding elongation period of 5 min at 72°C. The amplicons were paired-end sequenced by Nanoball sequencing on the DNBSEQ platform at BGI, Hong Kong. The demultiplexed sequences underwent analysis using the microbiome Bioinformatics platform Quantitative Insights into Microbial Ecology 2 (QIIME 2, 2023.9.2). The DADA2 pipeline v.1.18.0 was used for sequence denoising including quality-filtering, de-replicating, merging and checking for chimeras default with settings at a Phred score of at least 25, resulting in the generation of an amplicon sequence variant (ASV) table, phylogenetic tree and taxonomy information.

3.10 Data analysis

To analyse the bacterial community composition in the Pakistani soil samples, Rstudio (version 4.3.0) was used. Firstly, rarefaction was done to standardize the sampling effort across different samples to a sequencing depth of 28461. Additionally, sequences belonging to the taxa that were not bacteria were excluded for further

analyses. A criterion for determining significance in each statistical analysis was established, with significance defined as $P<0.05$. For alpha diversity, which includes metrics such as species richness, Shannon diversity index, and phylogenetic diversity, a comparative analysis was conducted across different cities and land uses using the Phyloseq and Picante v.1.8.1 programs. The data normality assumption was verified by the Shapiro test, while homogeneity of the variance was tested by the Bartlett test. The statistical analysis for significant differences was tested either by one-way analysis of variance followed by post hoc Tukey HSD test. Alternatively, if assumptions for these tests were violated, non-parametric alternatives such as the Kruskal-Wallis test and pairwise Wilcox test were employed. Visualization was facilitated through the ggplot2 3.4.2 package. As for beta diversity, the Bray-Curtis distance was employed to quantify dissimilarities among samples from different land uses and cities, visualized by Principle Coordinates Analysis (PCoA). The analysis utilized the PERMANOVA adonis function from the vegan 2.6-4 package to assess whether there were statistically significant variations in the bacterial community structure across different land use and city categories. Furthermore, DeSEQ2 analysis was carried out to identify the bacterial taxa showing significant variations in relative abundance across various land uses. Additionally, the unique bacterial composition of each land use was determined.

For correlation analyses, the multicollinearity from each environmental parameter was examined. On each environmental variable, a logarithmic transformation and standardization to z-scores was performed. Next, the Variance Inflation Factor and Spearman's rank correlation coefficient were calculated to remove collinear data. Variables with a variation inflation factor exceeding 6 and a correlation coefficient greater than 0.7 were not included in subsequent analyses. Additionally, to address zero values in our dataset, a Hellinger transformation was used on the bacterial community data. Forward selection determined which variables should be incorporated into the variation partitioning analysis. Only environmental variables showing significant influence were considered ($P<0.05$). The relationships between these variables were visualized by redundancy analysis (RDA) to reveal how soil chemical parameters correlate with different land use types. Subsequently, Variation Partitioning was conducted to determine which parameter i.e. land use, spatial, or environmental

factors contribute more to bacterial community structure variation based on the R^2 and P-value. This was visualized by a Venn diagram and both marginal as conditional effects of each environmental parameter after forward selection as well as the land use and spatial factor were calculated by looking at the R^2 and P-values. The conditional effect represents the individual contribution of a specific environmental factor, while the marginal effect captures the impact of a particular parameter together with others.

Through Partial Least Squares Regression (PLSR) and iterative PLSR (iPLSR), the relationship between the environmental parameters and IS*6100* abundance data was analyzed, allowing for the identification of key physiochemical factors that may influence the abundance of IS*6100.* While PLSR constructs latent variables to maximize covariance between predictors and responses, iPLSR iteratively selects variables based on their contribution to prediction accuracy, potentially resulting in a more concise model. The calibrated PLSR model was trained based on the entire dataset to best capture the underlying relationship between the predictors and the response variables. The cross-validation model involved systematically excluding each environmental variable once and assessing how well the model performed on the remaining data. By averaging the performance of the model over multiple iterations of this process, an estimate of its ability to predict outcomes on unseen data was made. The model was finetuned by selecting latent variables that explain the variation. The number of latent variables was chosen by minimizing the prediction error without overfitting the data. The performance of the model was determined by R^2 and root-mean-square error for calibration and cross-validation. For iPLSR, environmental variables were added to the model iteratively based on their impact on improving prediction performance.

For the statistical analysis of the qPCR data on IS*6100* abundances, significant differences in relative abundances were analysed by checking for the same assumptions as used for assessing the alpha diversity of bacterial community composition data and were determined by the same statistical tests as used for assessing the alpha diversity of bacterial community composition data.

4 Results and discussion

4.1 Pakistani samples: bacterial community analysis and IS*6100* relative abundances.

4.1.1 Bacterial community structure and diversity

4.1.1.1 Alpha diversity

In this study, the impact of land use types, spatial factors and environmental factors, on soil bacterial community structure and IS*6100* relative abundances were investigated in the Pakistani soil samples taken near 22 cities along the Grand Trunk Road in the north Punjab province of Pakistan. After rarefaction to a sequencing depth of 28461 and removing sequences of taxa which are not bacteria, in total 15047 taxa across the 66 samples were identified in the dataset. Analysis of the alpha diversity showed that the type of land use drove bacterial alpha diversity variation (Kruskal, $P<0.05$). The alpha diversity includes (i) the species richness referring to the amount of bacterial species or taxa identified in a microbiome, (ii) the Shannon diversity index assessing both richness and evenness to determine the similarity of abundances among different species in the community and (iii) the phylogenetic diversity which considers evolutionary/phylogenetic relationships among species in the community (Wagner *et al.*, 2018). Other studies previously reported that land use impacts the physico-chemical characteristics of the soils significantly, leading to modifications of soil properties such as pH, which affect the diversity of soil bacterial community (Bai *et al.*, 2022; Peddle *et al.*, 2024; Xue *et al.*, 2023). Additionally, Xue *et al.* (2023) reported that the impact on the microbial ecology is influenced by the intensity of land use. For instance, in our study, the agriculture soil communities showed a significant higher alpha diversity compared with the industrial and roadside soil communities as shown in Fig. 10 for the Shannon diversity index and phylogenetic diversity (Wilcox post hoc, $P<0.05$). This was also confirmed in the study of Walters & Martiny (2020), where land use significantly impacted alpha diversity and specifically agricultural soils displayed some of the highest levels of alpha diversity compared to other habitats on the globe. The alignment of our results with these prior findings underscores the importance of considering land management practices in

understanding and managing microbial diversity within agricultural and natural ecosystems (Walters & Martiny, 2020). The higher diversity in agriculture can be explained referring to previous studies, that suggested that disturbances linked with anthropogenic activity of different sources affect bacterial species composition and reduce species diversity (Berga *et al.*, 2012; Stephanou *et al.*, 2021). Agricultural practices such as nutrient supplementation and soil tillage, can increase bacterial diversity by redistributing the availability of resources (Stephanou *et al.*, 2021) triggering microbial metabolism and by mixing the bacterial microhabitats (Szoboszlay *et al.*, 2017). In industrial areas and at roadsides, increased concentrations of certain pollutants might decrease bacterial diversity, as found that Cu and Cd correlate positively with industrial samples as described by the RDA model in section 4.1.2.2 (Stephanou *et al.*, 2021).

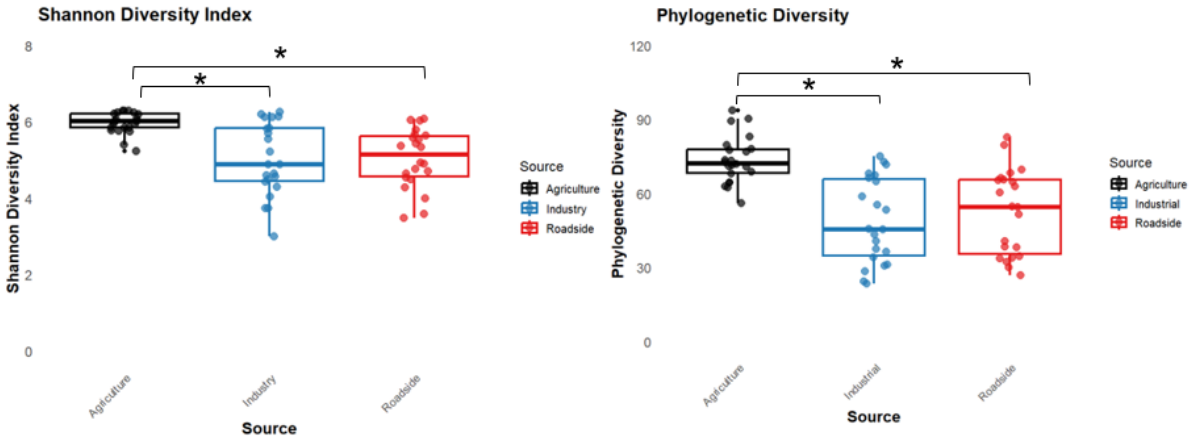


Fig. 10. Shannon diversity index and phylogenetic diversity according to land use as calculated for the Pakistani samples. The agricultural soils showed a significant higher Shannon diversity index and phylogenetic diversity compared to the industrial and roadside soils, indicating that land use drives bacterial alpha diversity variation. The asterix indicates the land use with a significant higher alpha diversity (P<0.05).

Meanwhile, the bacterial community alpha diversity also differed according to the location (denominated according to the nearest city) of the samples. When zooming in on a subset of cities, it is observed that the Kala Shah Kaku location showed a lower species richness and Shannon diversity index compared to the Gujar Khan, Gujranwala, Islamabad, Lalamusa, Rawalpindi and Rawat locations (Wilcox post hoc, P<0.05) (Fig. 11). Kharian also showed a higher Shannon diversity index compared to Kala Shah Kaku (Wilcox post hoc, P<0.05). Additionally, Dina had

a lower Shannon diversity index compared to Rawat and Islamabad (Wilcox post hoc, $P<0.05$). These conclusions need to be interpreted carefully because of the low statistical power with only 3 samples for every location/city. Differences in microbial diversity across cities can be attributed to the combined effects of localized environmental conditions and anthropogenic activities including soil conditions, climate, geographical location and specific environmental stressors like Hg which was not measured in our study (Sazykin *et al.*, 2023; Lopes *et al.*, 2021). Additionally, soil contamination by heavy metals, like the measured Cr and As, is of pressing concern in Pakistan, affecting ecosystems, public health and economic domains for which persistent exposures induce alterations in the overall composition of the soil microbial population (Khan, 2001; Sheik *et al.*, 2012). No significant differences were observed for the phylogenetic diversity index, even when subsets were made of the cities. This could be explained by the higher sensitivity of the metrics of Shannon diversity index and species richness in measuring species abundance and taxonomic composition. Whereas, phylogenetic diversity measures evolutionary distinctiveness or relatedness of species and is less able to capture subtle differences in species abundance, composition, or diversity within an ecosystem. Additionally, the lack of significant differences in our results could be attributed to the limited dataset and hence the limitations of statistical power in our dataset (Santini *et al.*, 2017).

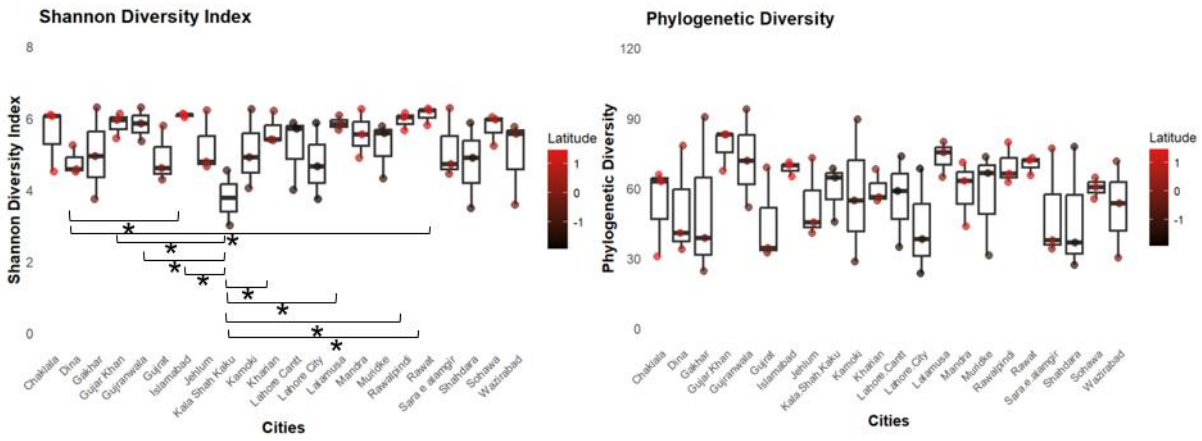


Fig. 11. Shannon diversity index and phylogenetic diversity according to the location (city) as calculated for the Pakistani soil samples. Locations showing a significant lower Shannon diversity index compared to other cities (P<0.05) are indicated with an asterix. Kala Shah Kaku location has a significant lower Shannon diversity index compared to Gujar Khan, Gujranwala, Islamabad, Kharian, Lalamusa, Rawalpindi and Rawat locations. Additionally, Dina had a lower Shannon diversity index compared to Rawat and Islamabad. No significant differences were observed based on the phylogenetic diversity.

4.1.1.2 Beta diversity

The beta diversity quantifies the variation in species composition and abundance among different communities (Wagner *et al.*, 2018). In this study, the Bray-Curtis metric was used to compute the dissimilarities across different locations (according to the nearest city) and land uses. The results showed that there were significant distinctions between bacterial community composition for both land use (ADONIS, R^2 = 0.083, P= 0.0001) and location (city) (ADONIS, R^2= 0.049, P = 0.0001). Land use contributed more to the dissimilarities than the location, based on the p-value and R^2-value, which measure how well the independent variables explain the variability. This result was also confirmed through Variation Partitioning Analysis in section 4.1.3, where the relative impact is discussed of these factors i.e. land use and location. Furthermore, some bacterial communities clustered in association with land use, while some clustered corresponding to the location as visualized by PCoA in Fig. 12. For instance, the bacterial communities of Gakhar, Gujrat, Kamoki, Muridke and Sara e Alamgir grouped together based on land use, with the industrial samples grouping together at the top of the second axis explaining 8.27% of the variation, the roadside samples clustering predominantly at the lower part of the second axis and left

side of the first axis and the agricultural samples situated towards the right side of the first axis. Notably, samples obtained from agricultural areas exhibited the greatest similarities in beta diversity in comparison to samples collected from industrial and roadside locations. Indeed, a clear cluster of samples from agriculture areas is observed, with the exception of the Dina, Kala Shah Kaku and Kharian locations. As mentioned before, the Dina and Kala Shah Kaku locations showed a lower Shannon diversity index, which likely contributes to this divergence. On the contrary, Islamabad, Lalamusa, Rawalpindi and Sohawa grouped together based on the location on the right side of the first axis explaining 15.9% of the variation, regardless of the specific land use. Mayerhofer *et al.* (2021) studying bacterial community diversity across soils in Switzerland, also found that differences in community composition were explained by both biogeographic region and the land use type. While the bacterial community composition of bacteria is sensitive to human activities e.g. land use practices and contaminants, the authors found that locations that were closer together tended to have more similar types of bacteria. Apparently, the location of the city explained why it had similar bacteria, confirming the significant influence of the location on the bacterial composition. On the contrary, regarding land use, the specific characteristics of the land play a role in determining the types of bacteria found there, making it a significant determinant of bacterial community composition (Mayerhofer *et al.*, 2021). Comparable land use types typically tend to have similar soil chemical characteristics resulting in similar microbial communities in a particular region, explaining the clusters of land use observed in our data. However, alterations and differences in soil chemical properties can induce changes in the composition of bacteria (Tian *et al.*, 2017). Different land management practices and degrees of urbanization can create unique environmental niches, resulting in variations in microbial community composition across locations with similar land use types (Mayerhofer *et al.*, 2021; Yan *et al.*, 2016).

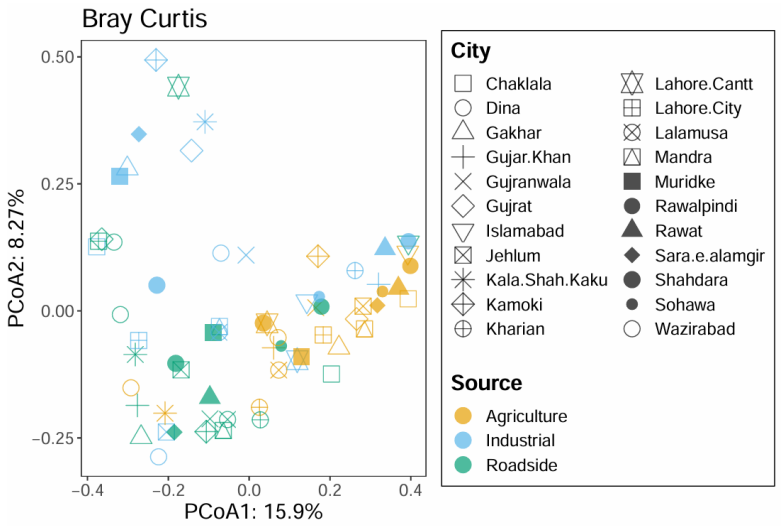


Fig. 12. PcoA plot based on Bray-Curtis dissimilarity metric showing bacterial community compositions clustering based on location (city) and land use. Some samples clustered based on land use, while others clustered based on the city.

4.1.1.3 Bacterial Community Composition across land use

Since land use type drove soil bacterial community structure, bacterial taxa showing significant differences in relative abundance between different land uses were determined by performing DeSEQ2 analysis. When comparing the agricultural and industrial soils, 45 ASVs were found to be more abundant in agricultural soil compared to industrial (DeSEQ2, Wald test, P<0.05). These ASVs were associated with genera like *Blastococcus, Brevundimonas, Constrictibacter, Dietzia, Gordonia, Indioceanicola, Luteimonas, Lysobacter, Microbacterium, Mycobacterium, Pseudomonas, Pseudorhizobium, Skermanella, Sphingoaurantiacus, Sphingomonas* and *Williamsia*. Meanwhile, 65 bacterial ASVs, associated with genera such as *Aeromicrobium, Aggregatilinea, Agromyces, Arenimicrobium, Flavisolibacter, Geodermatophilus, Ilumatobacter, Jiangella, Kribbella, Lysobacter, Methylobacterium, Methyloceanibacter, Microvirga, Nocardia, Nocardioides, Phascolarctobacterium, Pontibacter, Priestia, Promicromonospora, Rubrobacter, Skermanella, Sphingomicrobium, Streptomyces* and *Sulfotelmatobacter,* showed significantly lower abundances in agricultural soils compared to industrial soils (DeSEQ2, Wald test, P<0.05). The *Actinobacter* phylum e.g. *Dietzia* genus was most abundant in agricultural soils compared to industrial soils, while the *Actinobacter* phylum e.g. *Aeromicrobium* genus was most abundant in industrial soil compared to agricultural. Although the reason behind the variation in bacterial abundance across

soil types is unclear, some ecological roles of some bacteria in soils can be proposed. *Actinobacter* has diverse enzymes including ligninase, β-glucosidase and cellobiohydrolase for decomposing plant residues and as a result is of relevance for agricultural areas where this decomposition usually takes place (Bao *et al.*, 2021). The reason why *Pseudomonas* are more abundant in agricultural soil compared to industrial soil could be associated with their role of protecting plants/crops from pathogens. Its presence also enhances the accessibility of nitrogen and phosphorus to plants by fixing nitrogen and making phosphorus ions soluble. The latter function can also be performed by *Microbacterium* that was also more present in agricultural soil. It was also reported that *Pseudomonas* promotes plant growth in wheat and soybean crops by producing IAA and ACC deaminase, which are key plant hormones (Suman *et al.*, 2022). Additionally, the abundance of *Pseudomonas* and *Sphingomonas* has been associated with pesticide hexachlorocyclohexane application in agricultural soil as they can use it for their energy and nutrients (Gupta *et al.*, 2022). Meanwhile in our study, *Streptomyces* was not only more present in industrial soils compared to agriculture, similarly it was more abundant in roadside soils compared to agriculture as described in the next paragraph. Their widespread presence is attributed to having enzymes including cellulases, proteases, amylases and xylanases. These enzymes play crucial roles in various industrial applications and the presence of *Streptomyces* in industrial soils may be correlated with its potential contribution to processes relevant to industrial activities (Nazari *et al.*, 2022).

In the comparison between agricultural and roadside soils, 30 bacterial ASVs were found to have significantly higher abundance in agricultural soils (DeSEQ2, Wald test, $P<0.05$), whereas 68 bacterial ASVs showed significantly lower abundance in agricultural soils (DeSEQ2, Wald test, $P<0.05$). ASVs that were more abundant in agricultural soils than in roadside soils, belonged to genera such as *Blastococcus, Devosia, Dietzia, Georgenia, Gillisia, Lysobacter, Mesobacillus, Microvirga, Ornithinimicrobium, Paracoccus, Planococcus, Pseudomonas, Pseudorhizobium, Rubellimicrobium, Salinicoccus, Sphingomicrobium, Thermomonas and Youngiibacter.* In contrast, certain bacterial ASVs, affiliated with the genera *Aeromicrobium, Agromyces, Amycolatopsis, Bradyrhizobium, Catellatospora,*

Cellulomonas, Devosia, Jiangella, Kribbella, Lysinibacillus, Methyloceanibacter, Microlunatus, Microvirga, Mycobacterium, Neobacillus, Nitrospira, Nocardioides, Peribacillus, Priestia, Sphingomicrobium, Streptomyces and *Truepera* were significantly less abundant in agricultural soils than in roadside soils. The reason why bacteria are more abundant in one soil type compared to the other remains unclear, but some functions can be explained. The genus most abundant in the roadside soils compared to the agricultural soils was *Kribbella,* which can produce xylanase enzymes used in breakdown of organic matter (Putri & Setiawan, 2019). For agricultural soils the phylum with the highest relative abundance compared to roadside soils was again *Actinobacter*, followed by *P. thermomonas.* The latter organism is resistant to high temperatures which may be advantageous during hot periods in Pakistan as they were also found in hot springs in Jordan (Al-Daghistani *et al.*, 2021).

Comparing industrial and roadside soils, we identified 18 bacterial ASVs that exhibit higher abundance in industrial soil (DeSEQ2, Wald test, $P<0.05$), including ASVs belonging to *Devosia, Kocuria, Mesobacillus, Nocardioides, Ornithinicoccus, Pontibacter* and *Rubellimicrobium*. 14 ASVs were less abundant in industrial soils compared to roadside soils (DeSEQ2, Wald test, $P<0.05$), belonging to genera like *Constrictibacter, Kribbella, Luteimonas, Mycobacterium Skermanella, Sphingomicrobium* and *Sphingomonas. Nocardioides* was the most prevalent in industrial soils compared to roadside soils. This organism is known to degrade a variety of pollutants e.g. aromatic compounds, haloalkanes, hydrocarbons, polymeric polyesters, and nitrogen heterocycles used in various industries (Ma *et al.*, 2023). On the contrary, C*. Thermomicrobiales* is the most abundant in the roadside soil, which could be explained by its enhanced stress tolerance caused by pollutants from vehicle emissions and fluctuating temperatures (Dasgupta *et al.*, 2023). Regardless of the land use where the soil was taken from i.e. agriculture, industry, or along the roadside, *Actinobacteria* and *Proteobacteria* were consistently the most abundant types of bacteria present, which is in line with another study on soils from Pakistan. The study tried to understand how soil microbes respond to stress, to identify beneficial microbes for promoting plant growth and crop productivity by taking samples in the Indus Basin region and provinces of Punjab and Sindh in Pakistan (Marghoob *et al.*, 2022).

Furthermore, unique distinct bacterial taxa associated with each land use type were identified. Agricultural soil exhibited a greater number of unique bacterial ASVs compared to roadside and industrial soils. In total, 4932 ASVs unique to agricultural soils were identified. These ASVs belonged to species such as *Aggregatilinea lenta, Agromyces humi, Akkermansia muciniphila, Cesiribacter roseus, Klebsiella* genus, *Microvirga flavescens*, *Planifilum composti* and *Thiogranum longum, Variovorax paradoxus*. The increased presence of these bacteria remains unexplained, but it is possible that they have evolved specific traits to thrive in agricultural environments. *C. roseus* is capable of hydrolyzing aesculin, casein, starch, and gelatine. This ability suggests that it may contribute to the decomposition of organic matter in soil, which is essential for nutrient cycling and soil fertility (Liu *et al.*, 2011). *P. composti* was found in compost soil originally, which indicates its role in the degradation of organic matter and the conversion of complex compounds into simpler, more stable forms (Han *et al.*, 2013). *T. longum*'s capacity to oxidize inorganic sulphur compounds contributes to sulphur availability in soils. By improving sulphur availability, *T. longum* enhances the growth and productivity of sulphur-demanding crops, ultimately leading to increased yields and improved plant health and poses a selective advantage (Mori *et al.*, 2015). *M. flavescens* is capable of fixing atmospheric nitrogen. They convert atmospheric nitrogen into ammonia, which can be taken up by plants as a nitrogen source. This ability is particularly important in agriculture, because it reduces the need for costly synthetic nitrogen fertilizers (Soumare *et al.*, 2020; Zhang *et al.*, 2019). Roadside soil contained 3159 unique bacterial taxa, such as Brevibacterium samyangense, *Cesiribacter andamanensis, Desulforudis audaxviator, Desulfosporosinus hippie, Desulfotruncus arcticus, Devosia limi, Dietzia cinnamea*, and *Domibacillus enclensis* and *Thermomonas Brevis*. Although it cannot be explained why these bacteria are more present, these bacteria may be associated with functions to resist to or break down pollutants commonly found in road environments, like polyaromatic hydrocarbons, suggesting their importance in polluted ecosystems (Singh & Hiranmai, 2021). The industrial site soils showed 2996 unique bacterial ASVs including *Actinotalea ferrariae, Aequorivita vladivostokensis, Chryseolinea serpens, Fusicatenibacter saccharivorans, Granulicoccus phenolivorans*, *Sandaracinus amylolyticus* and *Truepera radiovictrix*. The reason for the higher abundance of these

bacteria cannot be definitively explained. However, it is possible that they have developed adaptations to industrial pollutants, enabling them to contribute to the degradation and detoxification of these pollutants (Bala *et al.*, 2022). *G. phenolivorans* is a bacterium known for its ability to degrade phenolic compounds which are common pollutants in industrial soil (Maszenan *et al.*, 2007). Phenol contamination links to plastics industry and various organic synthesis methods, as well as in dyes, pharmaceuticals, and other sectors. During phenol production and its utilization in various processes, phenol can readily enter soil through multiple pathways (Du *et al.*, 2017). A comprehensive list of all the unique bacterial ASVs for every land use type is not presented in this thesis.

4.1.2 The impact of environmental factors on bacterial community structure in the Pakistani soil samples

4.1.2.1 Multicollinearity

The environmental parameters (i.e. pH, EC, total nitrogen, DiPCB, TriPCB, TetraPCB, PentaPCB, HexaPCB, HeptaPCB, Cr, Cu, Pb, Cd, As, and TPH) that drive the variability in bacterial community structure were determined by forward selection. The influence of the selected environmental characteristics on the land use types was then visualized with an RDA plot. First, the multicollinear variables were removed. The Variance Inflation Factor and Pearson Correlation showed there was multicollinearity between diPCB, triPCB, tetraPCB, pentaPCB, hexaPCB, heptaPCB, latitude and longitude. These collinear variables were removed from the dataset for further analyses, leaving only triPCB since it has a higher importance as a pollutant in this study. Indeed, higher levels of triPCB were found in urban and industrial areas in the Punjab province, where our samples were taken, compared to other PCBs (Syed *et al.*, 2013).

4.1.2.2 Forward selection and RDA Analysis

Subsequently, forward selection was performed, indicating that Cu, total nitrogen, As, Cd and total carbon concentrations significantly explained the variation of bacterial communities in different land uses ($P<0.05$). The conditional effect, which is the pure contribution of each environmental parameter, was calculated together with the marginal effect collecting the impact of a particular environmental parameter in

combination with others. The parameter explaining most of the variation in composition of the bacterial community was Cu, although it only had a contribution of 3.8% (*P*=0.0009) for the marginal effect and 2% (*P*=0.0009) for the conditional effect (Table 4). Cd, As and total nitrogen also significantly explained a small part of the variation in the bacterial community composition, Cd contributed only 1.4% (*P*=0.005) to the marginal effect and 1% (*P*=0.006) to the conditional effect, As contributed 1.7% (*P*=0.003) to the marginal effect and 1% (*P*=0.009) to the conditional effect, whereas total nitrogen contributed 1.6% (*P*=0.002) to the marginal effect and 1.1% (*P*=0.004) to the conditional effect. Total carbon had the lowest contribution with 1.2% (*P*=0.009) for the marginal effect and 0.6% (*P*=0.051) for the conditional effect. The measured parameters hence offered only a minor explanation for the variability observed, highlighting the complexity of the soil ecosystem and suggesting that other factors not accounted for in the analysis might have a bigger impact on bacterial community composition.

Table 4. Marginal and conditional effect of environmental parameters explaining the variation in bacterial community composition between soils. Cu contributes the most, while total carbon the least.

	Marginal effect		**Conditional effect**	
	R^2_{Adj}	**P-value**	**R^2_{Adj}**	**P-value**
Cu	0.038	0.0009	0.02	0.0009
Cd	0.014	0.005	0.01	0.006
As	0.017	0.003	0.009	0.009
Total nitrogen	0.016	0.002	0.011	0.004
Total carbon	0.012	0.009	0.006	0.051

Cu explaining most of the variation is an important nutrient in soil, but excessive levels can become toxic to microorganisms. For instance, while it is an essential element for bacterial growth, it has been reported to inhibit nitrogen and phosphorus recycling in soil (Ahmad *et al.*, 2021). As such, high levels of Cu can negatively impact bacterial diversity as shown by Guo *et al.* (2023) that Cu pollution significantly negatively affects bacterial communities in paddy soils along a polluted river in southern China. In contrast to Cu, Cd is not an essential metal and toxic at lower concentration, affecting

bacterial diversity negatively (Guo *et al.*, 2023). Although Cd is indeed one of the most prevalent heavy metal pollutants globally, the measured concentrations in our samples were not that high compared to other heavy metals (Yu *et al.*, 2021). However, the correlation with toxicity parameters appears to be more closely associated with the fraction of Cd available in the soil rather than the total concentration of Cd (Vig *et al.*, 2003). As is globally acknowledged as one of the major risks to human safety and food security. More than 100 million individuals are at risk because of elevated As levels in groundwater used for drinking water, with Pakistan being among the most affected regions, which makes it an important characteristic in our study (Rehman *et al.*, 2022). A study on bacterial community composition and the functional genes related to As in different forest types showed that in soils contaminated with As, the presence of As significantly influences the composition of bacterial communities as a result of microorganisms adapting to long-term exposure to this element (Bei *et al.*, 2023). Genes involved in As reduction were more present in bacteria in the contaminated soils, since it is a toxic element and not essential (Bei *et al.*, 2023). For the total nitrogen content, higher levels of total nitrogen content led to a reduction in bacterial diversity, while moderate levels enhanced it. Additionally, maintaining optimal levels of microbial biomass carbon and nitrogen is essential for the productivity and sustainability of terrestrial ecosystems. These factors regulate important soil biogeochemical processes and have a significant impact on bacterial diversity. Consequently, they are considered noteworthy parameters to monitor and manage in soil management practices (Ullah *et al.*, 2020).

In a similar soil study the bacterial composition was determined in agriculture, forest, industrial and schoolyard samples in Cyprus. The key factors shaping bacterial composition were identified by forward selection and pH, total organic carbon, calcium, cobalt, EC, iron, Pb, sodium, kalium and vanadium were identified as significant factors explaining the variation in bacterial community composition. The study found that heavy metals, particularly aluminium, cobalt, Pb and vanadium were the important elements influencing the composition of bacterial communities in schoolyard and industrial settings. In contrast, bacterial communities in agricultural and forest settings appeared to be more strongly associated with the presence of calcium (Stephanou *et al.*, 2021).

While pH had no significant impact on the bacterial composition in our study, it has been confirmed in numerous studies that pH plays a crucial role in determining the composition and diversity of soil bacterial communities in terrestrial ecosystems (Bei *et al.*, 2023; Zhao *et al.*, 2022). This because pH can favour the survival of specific bacterial groups adapted to particular hydrogen ion concentrations. Additionally, changed soil characteristics e.g. nutrient availability, cation exchange capacity, organic carbon content, and soil moisture levels are often directly or indirectly influenced by pH (Amin *et al.*, 2020). Consequently, The structure of bacterial communities may change as a result of these modifications. However, another research on Pakistani soils conducted by Amin *et al.* (2020) also showed that the pH variation was not significantly correlated with bacterial composition. Their soil types could not be categorized as acidic or alkaline, similar to our study since our pH values ranged between 6.234 and 7.91 and can be considered neutral as shown in Table A.1.

Subsequently, many studies verified that temperature has a major impact on how bacterial communities are composed. Higher temperatures boost metabolic rates, while lower temperatures can slow down bacterial growth rates. This results in a different bacterial community composition with samples in winter and autumn being more complicated than the samples in summer and spring, since dominant bacteria with a high abundance become less abundant when it gets colder because of the slower growth rate, which gives other types of bacteria a better chance to grow, increasing the diversity of bacteria in autumn and winter (Landesman *et al.*, 2019; Zhang *et al.*, 2020). This could have influenced the results of the bacterial composition in this study since the samples were taken during winter in January and it would be interesting to examine if there is a difference in the bacterial composition and ecosystem functioning in summer (Zhang *et al.*, 2020).

The 5 selected environmental parameters were then used to build an RDA model to map how the selected environmental parameters affected the bacterial communities in every land use type. The RDA plot (Fig. 13) revealed that RDA1 on the x-axis accounted for 8.7% of the total variation, while RDA2 on the y-axis explained 3.1% of the observed variability. The majority of the bacterial communities from agricultural soil showed a positive correlation with increasing As levels. As already mentioned, As is a

major pollutant globally (Rehman *et al.*, 2022). Naturally occurring As found in geological formations, together with anthropogenic activities like pesticide use are the primary causes of As contamination and could explain why agriculture soil is positively correlated with it (Qaswar *et al.*, 2022). The Punjab region in Pakistan has high concentrations of As. Moreover, reports indicate that more than 45% of the groundwater samples in Pakistan are contaminated with As (Rehman *et al.*, 2022). This water is used for irrigation contaminating agricultural land and explaining the positive correlation between agriculture and As (Brammer & Ravenscroft, 2009). This might result in the emergence of As resistant bacteria that potentially can perform important roles in soil geochemical processes (Li *et al.*, 2021). For example, *Variovorax* can oxidize Arsenite, suggesting its role in remediating As contamination and was found to be uniquely present in agricultural soil in our study (Bahar *et al.*, 2013). On the other hand, bacterial community composition in agricultural soils also positively correlated with increasing total nitrogen concentration. Nitrogen fertilization is a widespread practice in agriculture that results in soils enriched with diverse nitrogen sources (Pan *et al.*, 2022). This could promote the development of specific bacterial taxa capable of using these nitrogen sources and increasing their prevalence in the soil (Pereira e Silva *et al.*, 2013; Dai *et al.*, 2018). As observed in this study *Klebsiella* was predominantly found in agricultural soils compared to roadside and industrial soils as they are able to fix atmospheric nitrogen, converting it into a form that plants can readily use for growth and development (Suman *et al.*, 2022).

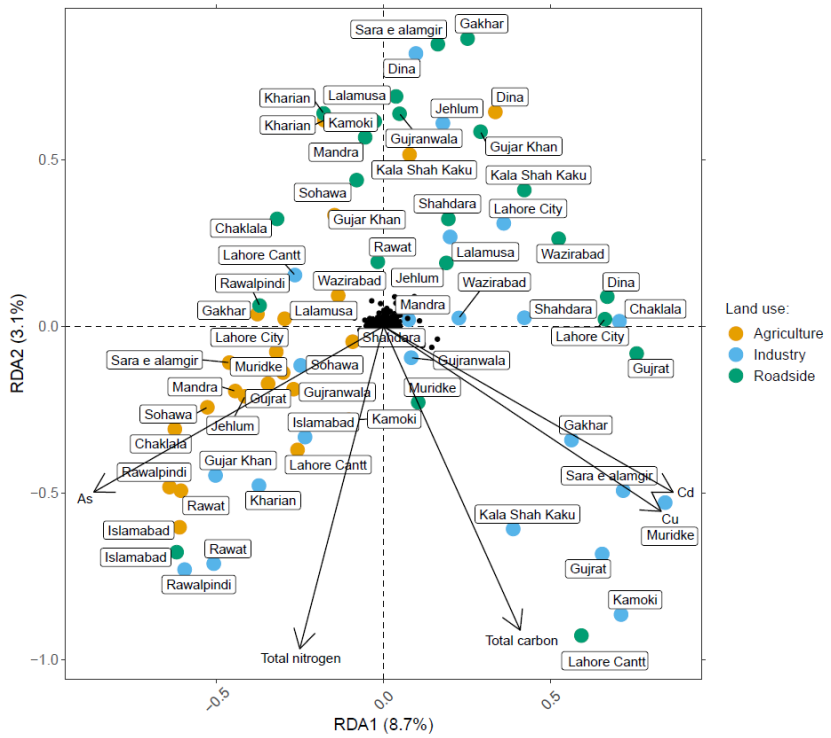


Fig. 13. RDA plot of the bacterial community compositions for every land use type correlated with measured environmental parameters. The different points represent the different sampling locations (cities) and a correlation was observed between the agricultural soils and As an total nitrogen, while some industrial soils correlated with Cu and Cd.

Furthermore, certain bacterial communities from industrial soils exhibited positive correlations with increased concentrations of Cd and Cu. However, this correlation was only obvious for some samples. Other research indicated that industrial soils in Pakistan have been found to be polluted with heavy metals like Cu and Cd (Din *et al.*, 2022; Waseem *et al.*, 2014). The industrial products containing Cd are plant dyes, metal pipes, solders or mixtures of metals (Din *et al.*, 2022). It was confirmed by a global study of Kubier *et al.* (2019) that Cd contamination in soil and groundwater is coming from mining, industry, waste deposit and disposal of sewage sludge, which contributes to the industrial area contamination. On the contrary, it was found for Cu that the sewage and industrial waste disposed in the area indicates anthropogenic influence. It comes from unregulated solid waste disposal or household liquid waste discharge (Malik *et al.*, 2010). The higher levels of Cu could be attributed to the manufacturing of wires, steel equipment and alloys where Cu is used and could end up poisoning the landfills. Indeed, soil around the city Lahore in North Punjab Pakistan is contaminated with heavy metals like Cd and Cu, primarily as a result of long-term

sewage sludge or water application (Younas *et al.*, 1998). Muridke is also strongly associated with Cu and Cd since it is the location of the Grand Trunk Road industrial zone which is the biggest industrial zone of Punjab (Abbas *et al.*, 2007). The presence of these heavy metals facilitates the development of bacteria that are resistant to Cu and Cd (Guo *et al.*, 2023). In this study, *Streptomyces* was found to more abundant in industrial soils compared to agriculture which is capable of resisting heavy metals (Mawang *et al.*, 2021).

Some other factors could explain more of the variation for these soils, which were not included in the analysis. For example, Zn was found as a significant pollutant on roadsides in Pakistan, but was not measured in our dataset (Faiz *et al.*, 2009)

4.1.3 Relative impact of environmental parameters, land use, and spatial factor on bacterial community structure in Pakistani samples

As the environmental factors Cu, As, Cd, total carbon and total nitrogen have limited explanatory power regarding the community structure, the contribution of land use and spatial factors were examined to determine the most significant parameter affecting these communities using Variation Partition Analysis. Land use explained most of the variance in bacterial community structure (7.2%) while the spatial factor contributed 4.4% and the environmental factor contributed 3% (Fig. 14). This observation suggests that variations in bacterial community structure were shaped by land use and spatial factors, rather than being significantly influenced by environmental parameters. Overall, 14% of the variance is explained by the three factors combined, while 86% remains unexplained.

The conditional effect of each factor was calculated together with the marginal effect using Variation Partition Analysis, to see the pure contribution of each factor, but also the combined impact of the factors. As shown in Table 5, the analysis further supported that land use drove the bacterial community variation more than spatial and environmental factors, as the R^2_{adj} values for both the marginal (R^2_{adj}= 0.043, *P*=0.001) and conditional effect (R^2_{adj}= 0.04, *P*=0.001) of land use were higher. It was followed for the conditional effect by the spatial factor with a significant influence of 3.3% (*P*=0.001) and lastly the environmental factors with a contribution of 1.5% (*P*=0.001).

For the marginal effect, spatial factor contributed 3% (*P*=0.001), while the environmental factors 1.5% (*P*=0.001).

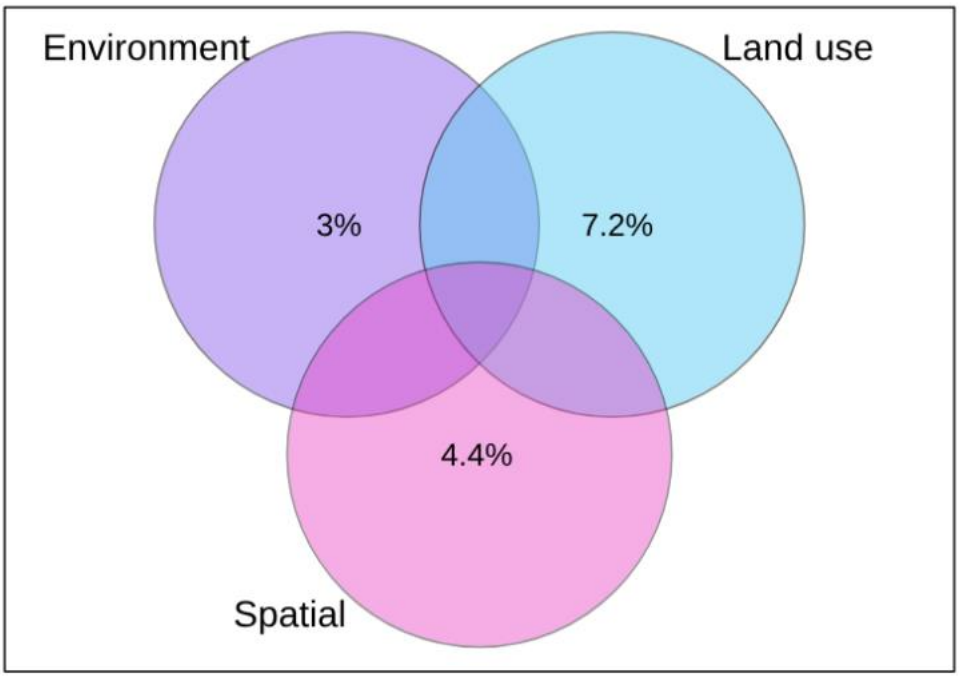


Fig. 14. Venn diagram showing the contribution of each factor on bacterial composition. Land use explains most of the variability, followed by the spatial factor, while the environmental factors explain the least.

Table 5. The marginal and conditional effect of the three factors: land use, spatial and environment for explaining the variation in bacterial community composition. Land use contributes the most, while the environmental factors the least.

	Marginal effect		**Conditional effect**	
	R^2_{Adj}	**P-value**	**R^2_{Adj}**	**P-value**
Land use	0.043	0.001	0.040	0.001
Spatial	0.03	0.001	0.033	0.001
Environment	0.015	0.001	0.015	0.001

The results showed that land use drove the variance in soil bacterial compositions more compared to spatial and environmental factors. This finding aligns with other research showing that while contributions of variables such as latitude and soil chemistry are significant, land use is highlighted as the predominant driver of bacterial

community composition (Barnett *et al.*, 2020; Fu *et al.*, 2023; Stephanou *et al.*, 2021). It was found in previous studies that various land use types host different communities of organisms, which indicates that particular organisms and biological processes are associated with specific land use practices. Altering land use practices directly impacts these bacterial compositions in natural ecosystems, making them important drivers of variation (Louisson *et al.*, 2023). When land use practices change in time, this has a lasting impact on the composition of the soil microbiome and influences the types and distributions of microorganisms present in the soil (Romdhane *et al.*, 2022). Furthermore, Barnett *et al.* (2020) confirmed that differences in how closely related these organisms are across different locations, are mostly influenced by how the land is used. This suggests that specific factors in land use make it more likely for closely related groups of organisms to be found together and help shape the community composition significantly (Barnett *et al.*, 2020).

The spatial scale across sampling sites influences the impact magnitude of the spatial factor in bacterial communities. In studies focusing on a larger scale e.g. whole China, there were more significant differences in bacterial communities compared to small scale studies. The large spatial scale affects the spatial variability in bacterial communities, potentially leading to an overestimation of the impact of the spatial factor in our study (Liu *et al.*, 2023).

A significant amount of the variability remains unexplained (86%), indicating that other factors may play a role in shaping bacterial communities and highlights the complexity of ecosystems (Plassart *et al.*, 2019). Other climatic factors like precipitation and radiation may have been neglected (Liu *et al.*, 2023). For example, exploring slope as another factor would be interesting, because it was found to be the most important factor affecting bacterial richness and diversity in the Karst forest (Peng *et al.*, 2019). In other studies, elevation was found as the main driver of variation in bacterial composition (Zhao *et al.*, 2022). Changes in slope or elevation result in different temperature, moisture and soil nutrients influencing the bacterial community (Peng *et al.*, 2019; Zhao *et al.*, 2022). Additionally, salinity was found to influence the relative abundance of phylotypes in Pakistan and would be an interesting parameter to add in further studies (Marghoob *et al.*, 2022).

4.1.4 The relative abundance of IS*6100* in correlation with land use, spatial factor and environmental factors in Pakistani samples

IS*6100* was detected in all soil samples ranging from 2.11 to 8.77 log(*tnpA*/10^6 16S rRNA) showing its widespread occurrence along the studied area. The distribution of IS*6100* across all samples shows its omnipresence within the bacterial communities as it is not restricted to specific conditions or locations. This observation is consistent with findings from studies on IS*1071*, where a similar widespread distribution of the element was reported in soils from different land uses, while relatively high abundances were found in ecosystems exposed continuously to organic xenobiotics (Debrabandere, 2023; Dunon *et al.*, 2013). IS*6100* being present in all samples suggests that it might be involved in crucial biological processes within the microbial populations or participating in processes related to human activities with their effects on the environment, as the same result was found for IS*1071* by Dunon *et al.* (2018). Relative abundances of IS*6100* were first correlated with land use, location/city, and latitude divided in north (Chaklala, Dina, Mandra, Islamabad, Gujar Khan, Sohawa and Rawat), central (Kharian, Gakhar, Gujrat, Jehlum, Sara e Alamgir, Wazirabad and Lalamusa) and south (Gujranwala, Kamoki, Kala Shah Kaku, Lahore Cantt, Muridke, Lahore City and Shahdara). The highest abundance was detected in the agricultural soil from Lalamusa in central Pakistan, while the lowest abundance was found in the agricultural soil in Sohawa in the North of Pakistan. However, no significant differences were observed across locations (Kruskal, $P>0.05$), neither across land uses (Kruskal, $P>0.05$) and not across latitude (Kruskal, $P>0.05$) (Fig. 15). As such in contrast to community composition, neither land use nor location seems to steer IS*6100* relative abundances. Finally, it was examined whether IS*6100* relative abundances were determined by biotic, i.e. community composition and abiotic environmental parameters. To this end, PLSR and iPLSR were performed to examine which environmental variables explain the variability of IS*6100* abundance. The R^2 of the model had a value of only 0.095, indicating that not much variability in the IS*6100* relative abundance is explained by the constructed model including environmental factors, species richness, Shannon diversity index and phylogenetic diversity. There are 2 variables explaining the abundance of IS*6100* significantly, which are TriPCB (coefficient= -0.05, $P<0.05$) and TetraPCB (coefficient=-0.058, $P<0.05$).

Both coefficients are negative, suggesting that as the values of TriPCB and TetraPCB increase, the relative abundance of IS*6100* decreases. On the contrary, 7 variables were considered as less important for explaining IS*6100* abundance e.g. Cr, Cu, As, TPH, longitude, Pb, and species richness. Specific taxa correlated with the relative abundance of IS*6100* were not precisely identified, although it could highlight certain taxa that play a significant role in its prevalence.

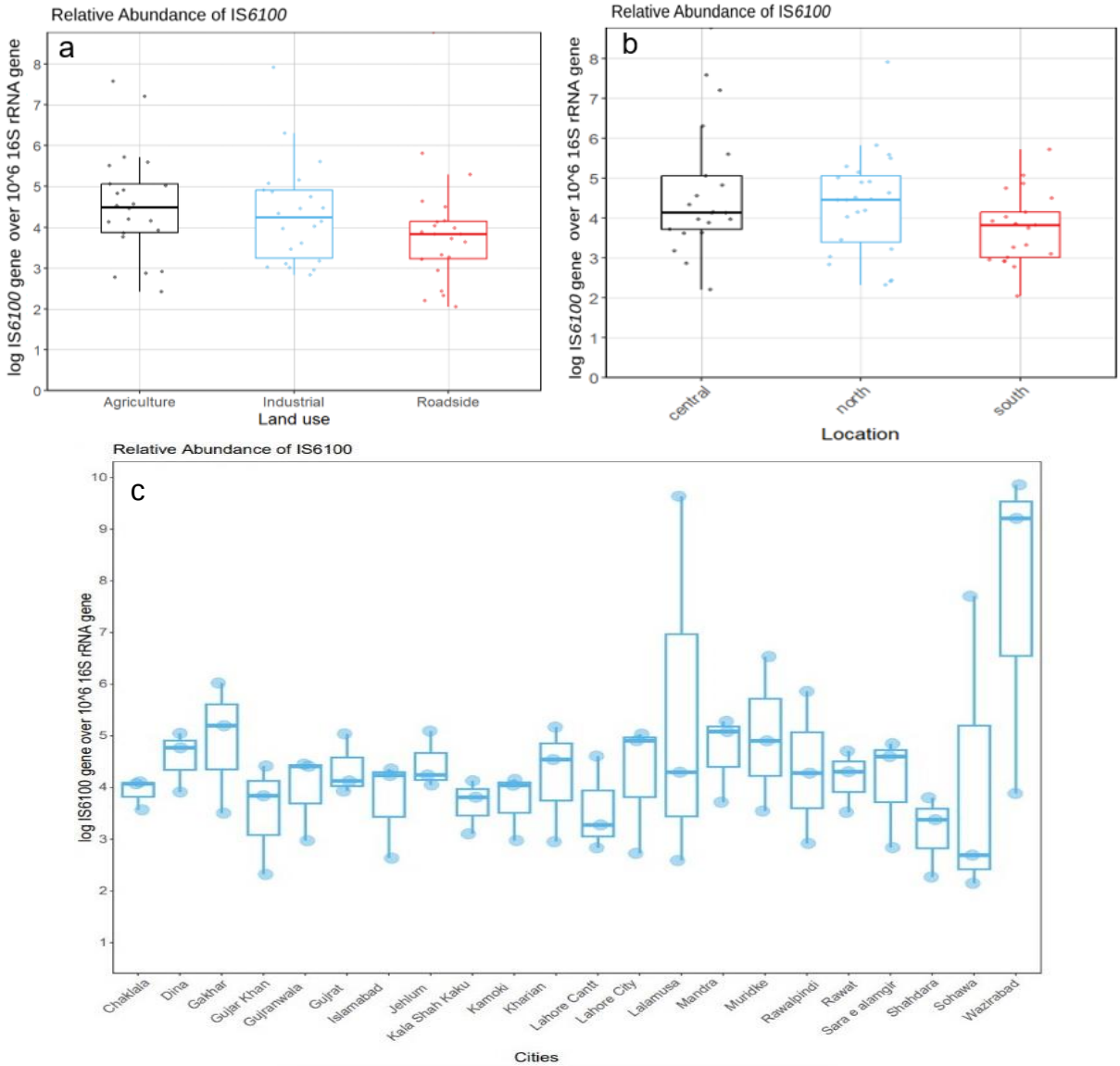


Fig. 15. Differences in abundance based on location, land use and latitude expressed in log(*tnpA*/10^6 16S rRNA). The agricultural soil from Lalamusa in central Pakistan had the maximum abundance, while the agricultural soil from Sohawa in Pakistan's north had the lowest abundance. a) Abundance of IS*6100* based on land use, where no significant differences were observed. b) Abundance of IS*6100* based on latitude, where no significant differences were observed. c) Abundance of IS*6100* based on cities, where no significant differences were observed.

Clearly IS*6100* abundance did not change across different land uses and locations and was not associated with environmental parameters and alpha diversity. If there was a correlation, it was extremely low for TriPCB and TetraPCB. Other studies found that the relative abundance of IS elements is regulated by the genome size and the frequency of deleterious insertion targets within the genome. These sites are locations where the insertion of an IS element could disrupt essential genes or regulatory regions, posing risks to the organism (Touchon & Rocha, 2007). The study by Touchon & Rocha (2007) indicated that IS family specificity, pathogenicity, human association, and gene transfer, do not play significant roles in determining IS abundance in prokaryotic genomes.

Additionally, there could also be a correlation with factors that were not measured, potentially causing a significant difference that was not yet investigated. Previous studies showed that most IS*6100* cargo genes carry genes play a role in antibiotic resistance, carbazole degradation, breakdown of hexachlorocyclohexane and carbofuran pesticide hydrolysis. The concentration of these xenobiotics may significantly correlate with IS*6100* abundance, however we did not measure them in this study (Dogra *et al.*, 2004, Gai *et al.*, 2010, Jiang *et al.*, 2022). To address this, a comparative approach could be implemented by applying distinct pesticide treatments in two separate fields, one known to be associated with IS*6100* e.g. carbofuran and another type of pesticide to see if there is a difference in relative abundance of IS*6100* in response to varying pesticide exposures.

The alpha diversity did not correlate with the IS*6100* abundance as presence of IS*6100* may not be directly associated to the diversity of bacterial communities. Multiple species potentially perform similar ecological roles in an ecosystem, as a result, other species can perform their function when there is a loss or gain in diversity. For instance, if there is a disturbance in the ecosystem, for example use of a pesticide, different species that are already present may compensate or take on additional functions to maintain ecological balance, including functions related to pesticide breakdown facilitated by genes carried by IS*6100* (Philippot *et al.*, 2021). As a result, the relative abundance of IS*6100* may remain relatively stable, regardless of fluctuations in alpha diversity, as other microbial species adapt and contribute to the ecosystem functioning.

Moreover, the impact of the selective forces of IS*6100* can be mitigated by the presence of other MGEs and ISs. These genetic elements may have similar ecological functions in facilitating genetic adaptation and variation leading to a smaller proliferation of IS*6100* genes, consequently the relative abundance does not change either. Possibly, other MGEs play a bigger role in xenobiotic degradation in the environment and not IS*6100*. There are not much studies on the relative abundance of MGEs across larger areas. Although a study by Nardelli *et al.* (2012) investigated how antimicrobial resistance genes were acquired, expressed, and distributed in integrons through a gradient of anthropogenic activities disturbing a Patagonian island. The precise factors leading to the spread of class one integrons among natural communities were still not fully understood. However, it was clear that integrons offer an advantage to the host cell by enabling the acquisition of gene cassettes, potentially enhancing the cell's ability to thrive in challenging environments. The study didn't find a clear pattern of differences in abundance across different levels of urbanization. Instead, it emphasized that the distribution of these genes is influenced by multiple factors and can vary between different habitats and regions (Nardelli *et al.*, 2012). Since no other studies focused on the differences in abundance of IS*6100* it can be even possible that it does not play a role in adaptation at all in the studied areas. However, even though IS*6100* might have helped with adaptation, it is possible that the soil's characteristics prevented the adapted organism from proliferating quickly or dispersing widely. The general circumstances of the soil may have imposed limitations that limited the extent to which the adapted organism could survive or dominate within the soil microbial community, even while IS*6100* mediated adaptations may have offered a certain degree of fitness advantage.

The similar relative abundance of IS*6100* can also be explained by the expanding human population, which leads to increased disturbance of the soil. The same trend is present in all different cities in Pakistan as there are a lot of anthropogenic activities because of the high population. It ranks as the sixth most populated country globally, experiencing an annual population growth rate of around two percent (Syed *et al.*, 2022). All these places have a lot of anthropogenic activities and as approximately 67% of the population of Pakistan lives in rural regions, this results in a high anthropogenic activity in agricultural soils that is comparable with industrial and

roadside pollution (Rahman *et al.*, 2011). Additionally, based on the results on microbial community analysis where the beta diversity was also similar across the cities, this could influence the comparable abundance of IS*6100* across different cities. The combination of increasing human population and the consequent microbial community similarities across urban environments could contribute to the lack of significant differences in IS*6100* abundance between cities.

Remarkably, although the relative abundance remained consistent, it does not imply a lack of ecological relevant functions of IS*6100* cargo genes. It could be that different functions are carried out in different samples, leading to a similar abundance although the cargo is different. Determining the functions would add valuable information to our understanding of microbial community dynamics and the role of IS*6100*. Moreover, the high abundance in the agriculture sample from Lalamusa in Central Pakistan, suggests a potential association with important functions specific to that environment. The identification of cargo genes facilitated by IS*6100* CTs could determine these functions to investigate what sets this sample apart from the others.

It is important to mention that no pristine environments were included in the Pakistani samples to see a difference between areas with high and low anthropogenic activity. Including a pristine environment would have facilitated a more comprehensive understanding of the impact of anthropogenic factors on the distribution and abundance of bacterial taxa and IS*6100*, thereby enhancing the ecological insights derived from the study.

4.2 Belgian samples

4.2.1 Effects of environmental stresses on the relative abundance of IS*6100* in Belgian samples

The relative abundances of IS*6100* in Belgian soils was examined including soils waiting at a soil treatment company for soil cleaning, while being contaminated by different pollutants (organic and inorganic) and originating from different locations in Flanders. Three soils were collected from three BPSs located at Inagro in Rumbeke. In addition, soils were sampled at around 9 agricultural fields located in West-Flanders. All fields had been treated with one or more pesticide solutions in the month prior to

sampling. Finally, one soil was sampled from a beach in West-Flanders. From each location/soil, three samples were taken and measured separately. The results showed that IS*6100*, as in the Pakistan soils, is omnipresent in the Belgian soils, since in every Belgian sample IS*6100* was detected above the detection limit ranging from 1.182 to 4.734 log(*tnpA*/10^6 16S rRNA) and corrected by the dilution factor. There were significant differences in relative abundance between different soils (Kruskal, $P<0.05$). Interestingly, the three soils originating from the BPSs used for the remediation of pesticide contaminated wastewater, showed the highest IS*6100* relative abundances which were significantly higher compared to soil contaminated with PFAS that showed the lowest relative abundance (Conover, $P<0.05$) (Fig. 16). In BPSs, there is a constant exposure to pesticides and hence likely a continuous selection of the adapted bacteria, which leads to higher relative abundances of MGEs involved in adaptation. This observation in BPS was also found for IncP-1 plasmids and IS*1071,* adding IS*6100* as a potential MGE that contributes to the microbial adaptation towards pesticide catabolism in the BPS environment. IS*6100* has previously been shown to carry genes for pesticide degradation like genes for lindane catabolism in lindane treated soils (Nagata *et al.*, 2007). The lowest relative abundance of IS*6100* was detected in the PFAS contaminated soil (designated as PFAS). Concentrations of PFAS in that soil were extremely low and PFAS is known to be highly resistant to microbial degradation (Berhanu *et al.*, 2023). The agricultural field soils that were exposed to pesticides at least one month prior to sampling did not show extreme high IS*6100* abundances as compared with the BPS soils. However, the timing of sample collection, whether shortly after spraying or after an extended period, may have had an impact. While in case of the BPS soils, there is a continuous exposure to pesticides implying a continuous selective pressure, that is not the case for the agricultural fields. It would be of interest to conduct a study with daily sampling after spraying a field with pesticides to investigate how the relative abundance of IS*6100* (and other MGEs) fluctuates immediately after pesticide application compared to over an extended period. Finally, no real trend was found for all the contaminated soils derived from the remediation company. For instance, different mineral oil contaminated soils were sampled but there the IS*6100* relative abundances varied a lot between 2.128 and 3.395 log(*tnpA*/10^6 16S rRNA). Overall, these soil samples did not show significantly higher IS*6100* relative abundances compared to soils derived from the agricultural soils

despite the relatively high contamination levels. Although, remarkably VLR and MO8 soils exhibited significantly higher IS*6100* abundance compared to other mineral oil-contaminated soils, as well as soils contaminated with TBT and PFAS. Therefore, while IS*6100* might contribute to pollution adaptation in the BPS environment, this is currently unclear in the other soils. Hence, soil samples from pristine areas with minimal pollution, like nature reserves should also be included in our analysis to compare the difference in abundance. As discussed for the Pakistan soil samples, IS*6100* might have only a minor contribution to adaptation being one of many MGEs. Moreover, conditions for adapted bacteria that used IS*6100* for adaptation were potentially not ideal to proliferate to sufficiently high numbers (like in the case of the BPS). As for the Pakistani soils, examining the functionality of the genetic cargo that is carried by IS*6100* CTs in the different soils might shed more light on the involvement of IS*6100* in the local community adaptation. Additionally, the relative abundance of IS*6100* needs to be linked to the bacterial composition or environmental factors to determine intricate associations within complex environments.

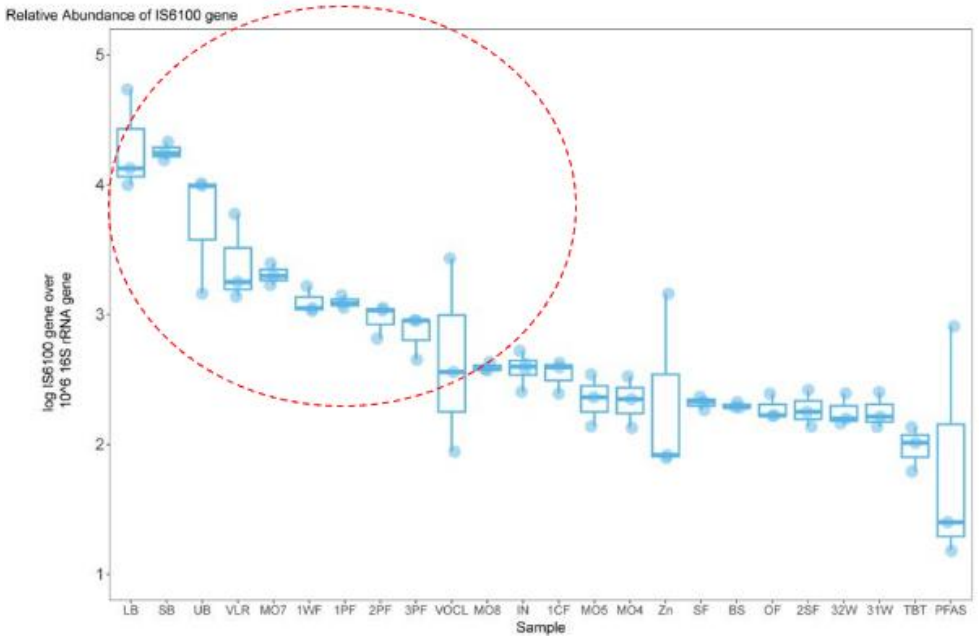


Fig. 16. Relative abundances of IS*6100* in the Belgian soils sampled from different environments exposed to various pollution types. BPS soil samples show the highest relative abundance, while the PFAS soil the lowest relative abundances. The soil samples showing a significant higher relative abundance of IS*6100* compared to PFAS are circled in red (Conover, $P<0.05$). Samples are ordered by decreasing *tnpA* relative abundance (log(*tnpA*/10^6 16S rRNA)). The median was 2.578 log(*tnpA*/10^6 16S rRNA) in the IN sample from a field sprayed with pesticides.

4.2.2 Amplification of the cargo of IS*6100* associated CTs

LR-PCR was utilized to amplify cargo genes linked with IS*6100* CTs from the Belgian soils. The LR-PCR amplicons were analysed by AGE to visualize and analyse variations in amplicon length sizes among soils. The LR-PCR method previously reported for IS*1071* enabled amplifying cargo genes of IS*1071* associated CTs up to 25 kb, which was adapted to recover IS*6100* associated cargo (Dunon *et al.*, 2022; Mawarda and Springael, unpublished results). The new developed LR-PCR technique selectively amplified sections positioned between two IS*6100* elements using primers that are directed outwards from IS*6100*. When coupled with long-read sequencing of the amplicons, the method makes it possible to study the functionality of the genetic cargo of IS*6100* in the samples. First, the method was performed on DNA extracted from reference *S. haloaromaticamans* MM1 that carries several plasmids containing IS*6100* CT cargo of different lengths. Expected amplicon sizes of 20 kb, 10 kb and 4 kb (Table 2) were recovered suggesting that the LR-PCR worked well (Fig. 17). Subsequently, LR-PCR was performed on several of the Belgian soil samples. Different replicate samples from the same soil were assessed. HMW eDNA was extracted from the soils as outlined in the Materials and Method section and the amplicons visualized by AGE. Size differences across different samples on the AGE will give a first insight into the variability of the IS*6100* CT cargo genes in the different soils and replicates. The HMW eDNA extraction and LR-PCR and AGE was repeated for some soil samples in case not all replicates showed LR-PCR amplicons on the AGE.

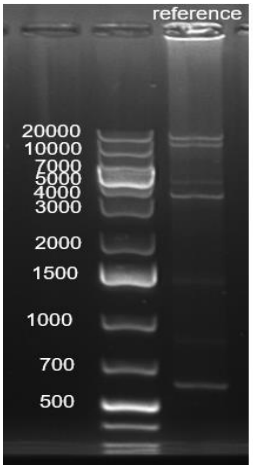


Fig. 17. AGE profile of amplicons of the LR-PCR targeting IS*6100* CT cargo as obtained using DNA of reference strain *S. haloaromaticamans* MM1 as the template. The expected band lengths of this reference are shown in Table 2.

Most soils and their replicate samples exhibited clear bands in the AGE, indicating that the cargo genes were amplified successfully (Fig. 18). For some soils/replicates the LR-PCR was repeated a second time or even a third time. This was for instance the case for the LB soil that in a first trial only showed amplicon bands for one replicate (Fig. 18.a) and in the second trial for two out of three replicates (Fig. 18.b). The three replicates showed that the LR-PCR procedure was quite reproducible per replicate but also that not all replicate samples taken from the same soil necessary show the same profile and hence contain the same IS*6100* CT cargo. The latter was also the case for soil SB. Soils UB, LB and SB represent all soils from BPS environments and inhomogeneity across samples taken from BPS was also observed when amplifying IS*1071* CT cargo (Dunon *et al.* 2018). Also various other soils showed substantial differences in AGE profiles between replicate samples, i.e., soils MO7, PFAS, VLR, Zn, VOCL, MO4, and MO5. Remarkably, the BPS samples UB, LB, SB showed a wide range of intense bands. BPS are exposed continuously to various pesticides during the largest part of the year which likely creates a significant pressure that drives the evolution and proliferation of bacteria capable of degrading pesticides of multiple chemical structures requesting different pathways (Sniegowski *et al.*, 2011). This is in contrast with samples WF and BS from a wheat field and beach sand having only a faintly perceptible single band. Meanwhile, the beach has unclear defined anthropogenic activities. MO7 showed repeatably clear bands up to 10 kb, while the other soils contaminated with mineral oil e.g. MO4, MO5, MO8 have less intense bands, which also have a smaller size. Three different potato fields were sampled represented by soils 1PF, 2PF and 3PF. All three soils showed different patterns, although one replicate of soil 2PF was similar to the patterns observed for the three 1PF replicates. Remarkably 1PF.2 showed a band around 20kb, which is particularly interesting to investigate. The next step is to look at the gene functions encoded by the presumed LR-PCR IS*6100* CT cargo amplicons. To this end, eight samples were selected for nanopore sequencing. Samples with a large diversity of bands were specifically interesting.

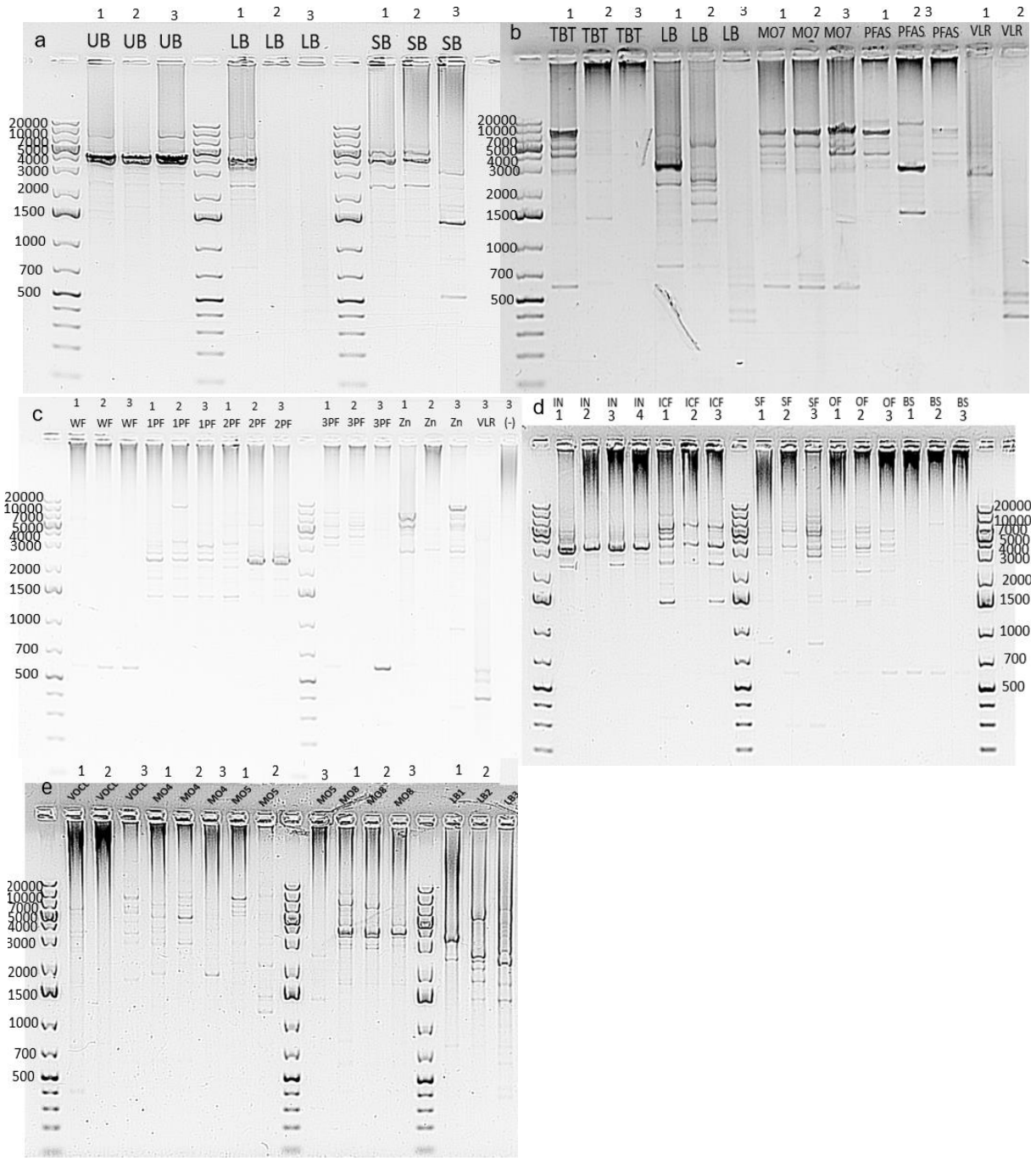


Fig. 18. AGE profiles of the amplicons obtained with LR-PCR targeting IS*6100* CT cargo in the Belgian samples across different panels. The used soil sample is indicated above the lanes. The different numbers in the lanes indicate different replicate samples taken from the same soil. The BPS samples UB, LB, SB show a high diversity of bands with a high intensity, while samples including BS and WF have only 1 faint band.

Based on the qPCR and LR-PCR, diverse soils/replicates were selected for nanopore sequencing of the LR-PCR IS*6100* CT cargo amplicons. The selected soils were UB.1, LB.1, SB.1, MO7.1, 1PF.2, TBT.1 and PFAS.1 with the number at the end indicating the replicate sample. Soils UB.1, LB.1, SB.1, 1PF.2 and MO7.1 showed multiple IS*6100* cargo bands ranging from 0.5 to 10 kb (Fig. 18) as well as a high relative abundances of IS*6100* using qPCR. Soils TBT.1 and PFAS.1 showed a low relative abundance of IS*6100* and multiple bands of larger length in the LR-PCR IS*6100* cargo profiles (Fig. 18). In addition to the selected soils, also the reference strain was selected for sequencing. To enable long-read sequencing, sufficient amplicon DNA is required with a DNA concentration up to 268 ng/µl per sample in case read lengths of 25 kb need to be sequenced. A minimal volume of 11,5 µL per sample is needed. To this end, the HMW eDNA extraction of the selected soils was repeated once (soil UB.1) or twice (other soils) and used as template for 3 identical LR-PCR targeting the IS*6100* CT cargo. Results are shown in Fig. 19.

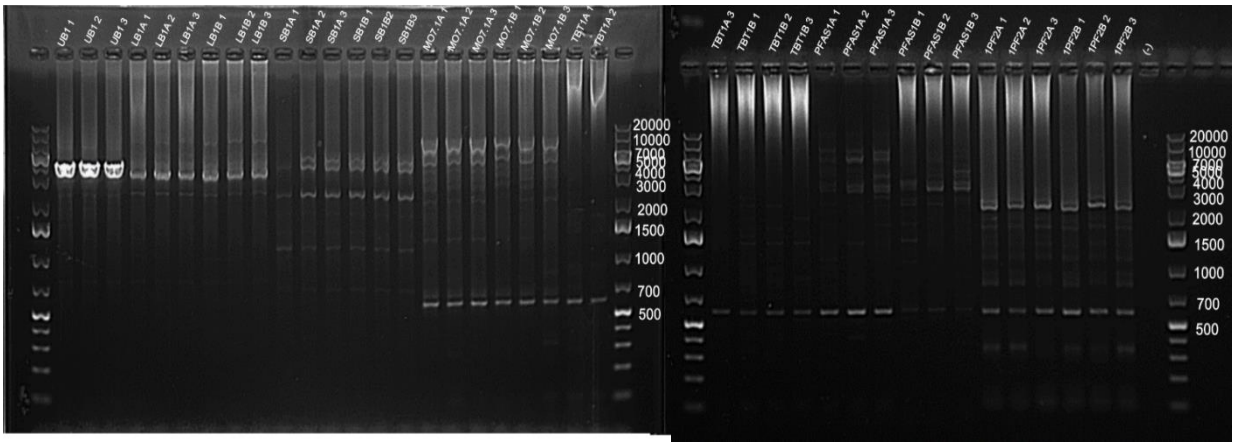


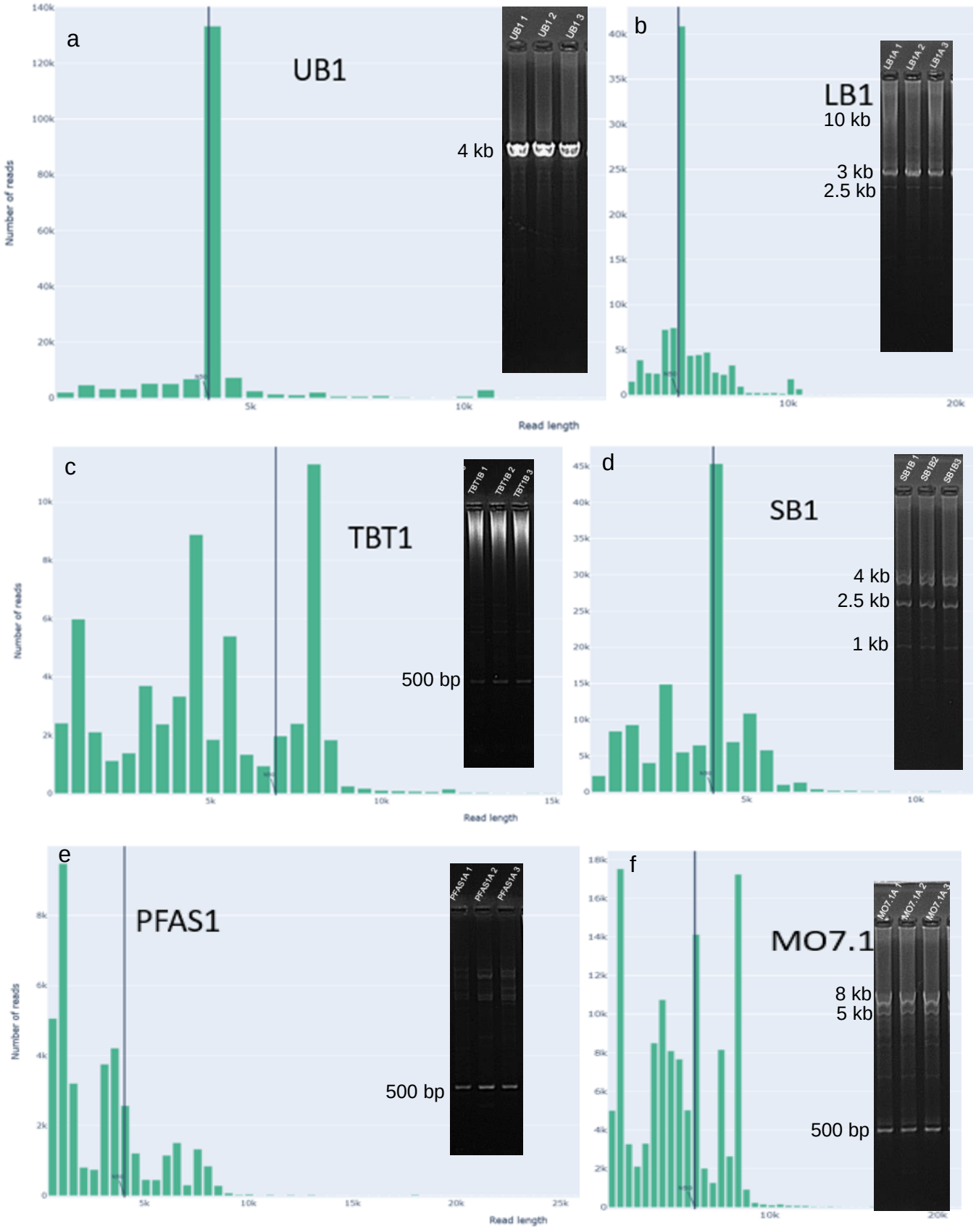
Fig. 19. AGE profiles showing LR-PCR amplicons targeting IS*6100* CT cargo from soils UB.1, LB.1, SB.1, MO7.1, PFAS.1, 1PF.2 and TBT.1. HMW eDNA was extracted 1X (in case of UB.1) or twice (other soils). Multiple extracts are indicated with A and B. Three identical LR-PCRs were performed on every DNA sample indicated by the numbers 1, 2 and 3 at the end of the lane indications. Samples UB.1, LB.1, SB.1, 1PF.2 and MO7.1 have clear and diverse bands of IS*6100* CT. On the contrary TBT.1 and PFAS.1 show a low diversity of bands.

The obtained LR-PCR profiles (Fig. 19) were compared with those obtained before (Fig.18) and for most of the samples the same profiles were obtained despite starting from new DNA extracts, showing again the reproducibility of the procedure. Exceptions were soil TBT.1 and PFAS.1 that did not show the expected multiple larger amplicons

of 5 to 10 kb observed earlier (Fig. 18). LR-PCR amplicons were pooled for each soil and the pooled LR-PCR products were purified. An additional HMW eDNA extraction and LR-PCR was performed for PFAS.1 and SB.1 as the concentration was still too low. Minimum concentrations for long-read sequencing of amplicon lengths up to 25 kb were finally achieved for samples UB.1 (326 ng/µl), LB.1 (278 ng/µl), MO7.1 (312 ng/µl) and the bacterial reference (515 ng/µl). SB.1 had 246 ng/µl and 1PF.2 had 256 ng/µl, which is sufficient as the observed bands had only a length of maximum 10 kb for which 107.15 ng/µl is required, while TBT.1 (226 ng/µl) and PFAS.1 (160.2 ng/µl) showed very small bands and still met the requirement for 10 kb.

Nanopore sequencing can sequence long reads, advantageous for aligning distinct loci within a large genome and genomes that contain repetitive sequences (Yasir *et al.*, 2022). The correlation between the observed bands lengths on AGE with the dominantly sequenced read lengths in nanopore sequencing, were a first indication that nanopore sequencing was successful (Fig. 20). Band lengths of higher intensity on the AGE should correlate with the most sequenced read lengths. This was indeed the case. For instance, for UB.1 the most intense band on the AGE gel was around 4 kb and a read length of 4 kb also showed the highest number of reads in the nanopore sequencing. For LB.1 the most intense AGE band is around 3 kb with less intense bands at 2500 bp and 10 kb which is in line with the most prevalent read lengths in the nanopore sequences on Fig. 20.b. For TBT.1 the bands on the AGE were less diverse and with a low intensity, which is in line with the low amount of long-read sequences compared to UB.1 and LB.1. The intense band at 500 bp on the AGE was in line with the peak in read number for that length in nanopore sequencing. Interestingly, two other peaks in read number for amplicon lengths of around 4.5 and 6 kb were not visible on the AGE. In Figure 20.c there is a smear of DNA visible at the upper part of the lanes for the TBT.1 samples, which might indicate that too much DNA was loaded which resulted in a poor resolution, making the higher bands undetectable (Doggett *et al.*, 1992). Hence, these reads could be linked with the bands observed at around 4 to 10 kb detected earlier in soil TBT.1 (Fig. 18). For all other soils, the most frequently registered read lengths corresponded with the dominant AGE band lengths. Finally, the reference strain was extracted and amplified a couple of times. Multiple LR-PCR products were pooled together to reach the desired high concentrations.

Some replicate LR-PCRs of the reference did not show the bands of 10 and 20 kb (data not shown) and it was therefore decided to perform extraction and LR-PCR again, although it still resulted in the underrepresentation of these high bands in the nanopore sequencing in comparison with the other smaller amplicons. The absence of 10 and 20 kb bands in the LR-PCR could be due to shearing, resulting in fragmentation of the DNA during the protocol of plasmid extraction. The bands observed in AGE at 600 bp and faintly at 800 bp are visible as peaks in the nanopore reads (Figure 20.g). The nanopore read peaks of around 1.5 kb and 3.5 kb and 4.5 kb are also visible on the AGE. It was concluded that for all samples there is a congruence between the read length frequency obtained from nanopore sequencing and the band intensity and size observed on AGE. This alignment between the two independent methods suggests a successful sequencing and confirmation of the accuracy and reliability of the sequencing results. In a next step, it needs to be confirmed that the sequences indeed relate to IS*6100* CT cargo. That can be checked by analysing the ends of the sequences for IS*6100* tag sequences that are expected to become co-amplified. Furthermore, identifying the functions of the CT cargo genes is needed for further insight into the ecology of IS*6100*.


a
UB1
4 kb
UB1 1
UB1 2
UB1 3
b
LB1
10 kb
3 kb
2.5 kb
LB1A 1
LB1A 2
LB1A 3
c
TBT1
500 bp
TBT1B 1
TBT1B 2
TBT1B 3
d
SB1
4 kb
2.5 kb
1 kb
SB1B 1
SB1B2
SB1B3
e
PFAS1
500 bp
PFAS1A 1
PFAS1A 2
PFAS1A 3
f
MO7.1
8 kb
5 kb
500 bp
MO7.1A 1
MO7.1A 2
MO7.1A 3
Number of reads
Read length

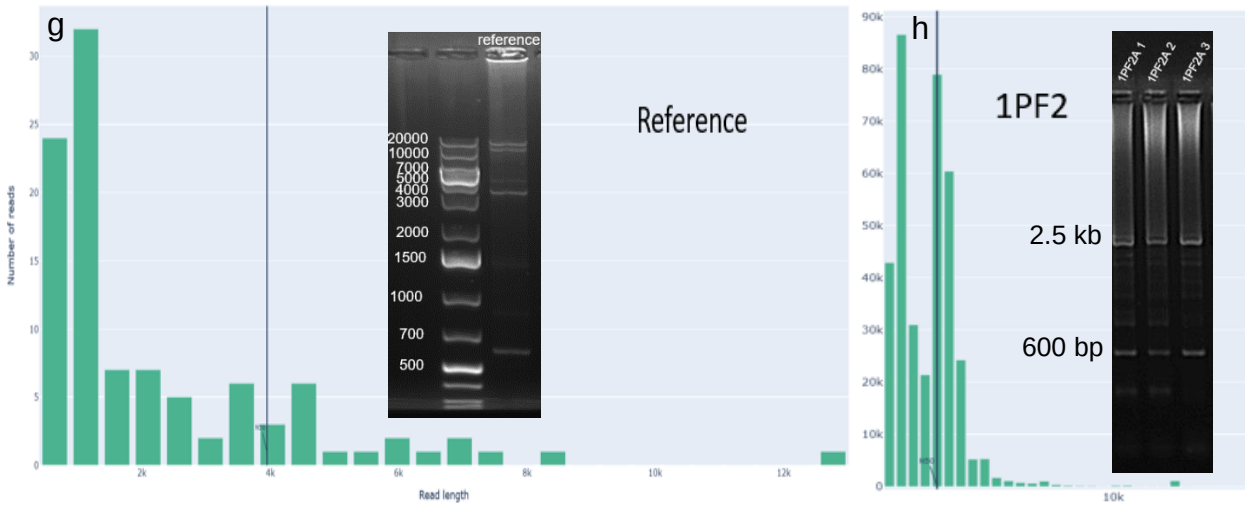


Fig. 20. The sequence reads and AGE profile (from Fig. 19) of 8 selected samples UB.1, LB.1, TBT.1, SB.1, PFAS.1, MO7.1, reference, 1PF.2. These sequencing reads are compared with the bands on the AGE. The presence and intensity of each band is in line with the number of reads for each read length suggesting a well performed sequencing. Every panel shows the results of a different soil sample.

5 Conclusions and future perspectives

In the first part of this thesis, focusing on Pakistani samples, the bacterial diversity and the abundance of IS*6100* were correlated with environmental factors, land use and spatial factor. Land use significantly drove the alpha diversity variation since it has a substantial impact on the physico-chemical features of the soils, altering the diversity of the soil bacterial population. Compared to industrial and roadside communities, agricultural soil had higher alpha diversity, possibly due to the higher disturbance. Additionally, based on cities/locations it was found that the city Kala Shah Kaku and Dina had a lower alpha diversity, which could be attributed to the localized environmental conditions and anthropogenic activities, although the phylogenetic diversity showed no significant differences. Moreover, for the beta diversity, results also indicated significant differences in bacterial community composition based on land use and location, with land use again contributing more to the dissimilarities. Some samples clustered based on cities, while others clustered based on land use, with the samples collected from agricultural soil showing a cluster with a more similar bacterial composition. In other studies it was also found that land use characteristics determine the types of bacteria found, with comparable land use types having more similar microbial communities due to similar soil chemical characteristics. However, alterations in soil properties due to land management practices and urbanization induced changes in bacterial composition. Furthermore, unique bacterial community analyses for every land use revealed that certain taxa may serve as indicators of bacterial adaptation to particular environmental stressors inherent to each land use. While environmental factors explaining most of the soil bacterial community variation were Cu, total nitrogen, As, Cd and total carbon, the contributions were low. An RDA model was built to illustrate how bacterial communities in every land use type were impacted by the selected environmental parameters. The bacterial communities found in agricultural soil exhibited a positive correlation with increasing levels of As. This trend could be linked to the fact that 45% of groundwater samples in Pakistan, which are contaminated with As, are often utilized for irrigation in agricultural land. Additionally, the composition of bacterial communities in agricultural soils displayed a

positive correlation with increasing concentrations of total nitrogen, since the common agricultural practice of nitrogen fertilization leads to soil enrichment with various nitrogen sources. Furthermore, several bacterial communities identified in industrial soils demonstrated positive correlations with elevated concentrations of Cd and Cu, probably due to the heavy metals contamination issue in Pakistan, where Cd and Cu are prevalent due to prolonged sewage sludge or water application. However, it was found that the environmental parameters explained the least of the soil bacterial composition, while land use contributed the most. Still a significant amount (86%) of the variability remained unexplained, suggesting that other factors that were not measured could shape the bacterial composition, for example Zn, the slope, temperature… It would be interesting to include these measurements in a future experiment to explain more of the variation.

In every sample of the Pakistani soils IS*6100* was present, indicating that it is widespread. However, there were no differences in abundance between different cities/locations, land uses and latitudes. It is possible that IS*6100* is not involved in adaptation in the examined soils and that only a background IS*6100* abundance was measured. Otherwise, the similar abundance of IS*6100* in the different soils may be attributed to compensatory functions of other microbial species and the potential influence of other MGEs diluting the effect of IS*6100*. Otherwise, IS*6100* might be involved in adaptation to unknown stresses that vary between soils but finally all contributing to a similar abundance. Last but not least, maybe IS*6100* contributed to adaptation, but conditions in the soil did not allow for strong proliferation of the adapted organism. That is why the function of CT cargo genes should be identified by sequencing the LR-PCR so more insights could be obtained into the ecology of IS*6100*. Furthermore, the environmental factors negatively correlating with the abundance of IS*6100* were DiPCB and TriPCB as PCB is a toxic contaminant, while other environmental factors, species richness, Shannon diversity index, and phylogenetic diversity explain minimal variability in IS*6100* abundance. There might be a correlation with unmeasured parameters, which could result in a noteworthy difference that hasn't been examined yet. For future research it would be interesting to include the measurement of carbazole, hexachlorocyclohexane and carbofuran

(xenobiotics associated with the cargo genes associated with IS*6100*) to see if there is a higher explanatory power between IS*6100* abundance and those xenobiotics. Additionally, a pristine sample should be added in the analyses to distinguish between samples with high and low levels of anthropogenic activities. Furthermore, specific taxa associated with relative abundance of IS*6100* still need to be identified.

The second part of this thesis focused on Belgian samples, to assess both the abundance and the presence of cargo genes associated with IS*6100*. IS*6100* was again detected in every Belgian sample, with significant differences among samples with exposures to different contaminants. Soil samples continuously exposed to pesticides in BPS samples showed higher IS*6100* abundance compared to those with less clear contaminants like beach sand, indicating a potential correlation between continuous selective pollution and the occurrence of IS*6100*. This also signifies IS*6100* as a potential MGE facilitating bacterial adaptation within the BPS environment for pesticide degradation. In the PFAS contaminated soil, IS*6100* was found to have the lowest relative abundance. Moreover, agricultural soil did not show high IS*6100* abundances as the BPS samples did. In contrast to the consistent exposure to pesticides observed in BPS soils, agricultural fields do not experience continuous selective pressure from pesticides as it is mostly only sprayed with pesticides twice a year. A study focusing on daily sampling immediately after pesticide application could provide valuable insights into the dynamics of MGEs like IS*6100* in response to pesticide exposure. This approach would investigate how the relative abundance of MGEs fluctuates shortly after pesticide application compared to observations over an extended period. Also for the Belgian samples no pristine environment was sampled, although it would be interesting to compare what the prevalence of IS*6100* would be. Moreover, a closer look at the way the genetic cargo carried by IS*6100* CTs functions in various soil conditions could provide additional insight into how IS*6100* contributes to bacterial community adaptation. Additionally, just what was done for the Pakistani soil samples, the IS*6100* abundances from the Belgian soil samples still needs to be correlated with its bacterial community diversity and environmental parameters to better understand which factors drive the variation.

Regarding cargo identification, AGE was used to view and analyse the differences in amplicon length sizes between samples in the LR-PCR amplicons. Every sample had

distinct bands in the AGE, proving that the cargo genes had been successfully amplified with the developed LR-PCR method. The three replicates of each sample demonstrated high reproducibility of the LR-PCR process per replicate, but they also demonstrated that different replicate samples from the same soil do not always exhibit the same profile and can contain different IS*6100* CT cargo. Notably, the BPS samples UB, LB, SB showed a diversity of intense bands, probably due to the continuous selective force of the pesticides. The intensity and size observed for the bands in the AGE analysis were found to be comparable to the quantity of reads obtained through nanopore sequencing, indicating congruence between the two methodologies and were a first indication that sequencing of this cargo was successful. Future studies need to identify the cargo genes of the IS*6100* CT and need to examine the correlation between IS*6100* CTs and its environmental parameters. This way, our understanding on the ecological role and potential influence of IS*6100* on soil bacterial community adaptation may be improved.

References

Abbas, S. T., Sarfraz, M., Mehdi, S. M., Hassan, G., & Obaid-Ur-Rehman. (2007). Trace elements accumulation in soil and rice plants irrigated with the contaminated water. *Soil and Tillage Research*, *94*(2), 503–509. https://doi.org/10.1016/J.STILL.2006.10.004

Rahman, A., Hayat, Y., Habib, Z., & Iqbal, J. (2011). Rural-Urban disparities in Khyber Pakhtunkhwa Pakistan. *Sarhad J. Agric*, *27*(3), 477–483.

Ahmad, W., Alharthy, R. D., Zubair, M., Ahmed, M., Hameed, A., & Rafique, S. (2021). Toxic and heavy metals contamination assessment in soil and water to evaluate human health risk. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-94616-4

Al-Daghistani, H. I., Mohammad, B. T., Kurniawan, T. A., Singh, D., Rabadi, A. D., Xue, W., Avtar, R., Othman, M. H. D., & Shirazian, S. (2021). Characterization and applications of *Thermomonas hydrothermalis* isolated from Jordan's hot springs for biotechnological and medical purposes. *Process Biochemistry*, *104*, 171–181. https://doi.org/10.1016/J.PROCBIO.2021.03.010

Alamro, H., Alzamel, M., Iliopoulos, C. S., Pissis, S. P., & Watts, S. (2021). IUPACpal: efficient identification of inverted repeats in IUPAC-encoded DNA sequences. *BMC Bioinformatics*, *22*(1). https://doi.org/10.1186/S12859-021-03983-2/FIGURES/7

Alengebawy, A., Abdelkhalek, S. T., Qureshi, S. R., & Wang, M. Q. (2021). Heavy Metals and Pesticides Toxicity in Agricultural Soil and Plants: Ecological Risks and Human Health Implications. *Toxics*, *9*(3). https://doi.org/10.3390/TOXICS9030042

Amin, A., Ahmed, I., Khalid, N., Khan, I. U., Ali, A., Dahlawi, S. M., & Li, W. J. (2020). Insights on comparative bacterial diversity between different arid zones of Cholistan Desert, Pakistan. *3 Biotech*, *10*(5). https://doi.org/10.1007/S13205-020-02204-6

Bahar, M. M., Megharaj, M., & Naidu, R. (2013). Kinetics of arsenite oxidation by *Variovorax sp.* MM-1 isolated from a soil and identification of arsenite oxidase gene. *Journal of Hazardous Materials*, *262*, 997–1003. https://doi.org/10.1016/J.JHAZMAT.2012.11.064

Bai, Z., Zheng, L., Bai, Z., Jia, A., & Wang, M. (2022). Long-term cultivation alter soil bacterial community in a forest-grassland transition zone. *Frontiers in Microbiology*, *13*. https://doi.org/10.3389/FMICB.2022.1001781/FULL

Bala, S., Garg, D., Thirumalesh, B. V., Sharma, M., Sridhar, K., Inbaraj, B. S., & Tripathi, M. (2022). Recent Strategies for Bioremediation of Emerging Pollutants: A Review for a Green and Sustainable Environment. *Toxics*, *10*(8). https://doi.org/10.3390/TOXICS10080484

Bao, Y., Dolfing, J., Guo, Z., Chen, R., Wu, M., Li, Z., Lin, X., & Feng, Y. (2021). Important ecophysiological roles of non-dominant *Actinobacteria* in plant residue decomposition, especially in less fertile soils. *Microbiome*, *9*(1). https://doi.org/10.1186/S40168-021-01032-X/FIGURES/7

Barkay, T., Miller, S. M., & Summers, A. O. (2003). Bacterial mercury resistance from

atoms to ecosystems. *FEMS Microbiology Reviews*, *27*, 355–384. https://doi.org/10.1016/S0168-6445(03)00046-9

Barnett, S. E., Youngblut, N. D., & Buckley, D. H. (2020). Soil characteristics and land-use drive bacterial community assembly patterns. *FEMS Microbiology Ecology*, *96*(1). https://doi.org/10.1093/FEMSEC/FIZ194

Bei, Q., Yang, T., Ren, C., Guan, E., Dai, Y., Shu, D., He, W., Tian, H., & Wei, G. (2023). Soil pH determines arsenic-related functional gene and bacterial diversity in natural forests on the Taibai Mountain. *Environmental Research*, *220*, 115–181. https://doi.org/10.1016/J.ENVRES.2022.115181

Benler, S., Faure, G., Altae-Tran, H., Shmakov, S., Zheng, F., & Koonin, E. (2021). Cargo Genes of Tn*7*-Like Transposons Comprise an Enormous Diversity of Defense Systems, Mobile Genetic Elements, and Antibiotic Resistance Genes. *MBio*, *12*(6). https://doi.org/10.1128/MBIO.02938-21/SUPPL_FILE/MBIO.02938-21-SF009.EPS

Bennett, P. M. (2008). Plasmid encoded antibiotic resistance: Acquisition and transfer of antibiotic resistance genes in bacteria. *British Journal of Pharmacology*, *153*. https://doi.org/10.1038/SJ.BJP.0707607

Berga, M., Székely, A. J., & Langenheder, S. (2012). Effects of Disturbance Intensity and Frequency on Bacterial Community Composition and Function. *PLoS ONE*, *7*(5). https://doi.org/10.1371/JOURNAL.PONE.0036959

Berhanu, A., Mutanda, I., Taolin, J., Qaria, M. A., Yang, B., & Zhu, D. (2023). A review of microbial degradation of per- and polyfluoroalkyl substances (PFAS): Biotransformation routes and enzymes. *Science of The Total Environment*, *859*. https://doi.org/10.1016/J.SCITOTENV.2022.160010

Böhm, M. E., Huptas, C., Krey, V. M., & Scherer, S. (2015). Massive horizontal gene transfer, strictly vertical inheritance and ancient duplications differentially shape the evolution of *Bacillus cereus* enterotoxin operons *hbl, cytK* and *nhe. BMC Evolutionary Biology*, *15*(1). https://doi.org/10.1186/S12862-015-0529-4

Brammer, H., & Ravenscroft, P. (2009). Arsenic in groundwater: A threat to sustainable agriculture in South and South-east Asia. *Environment International*, *35*(3), 647–654. https://doi.org/10.1016/J.ENVINT.2008.10.004

Briffa, J., Sinagra, E., & Blundell, R. (2020). Heavy metal pollution in the environment and their toxicological effects on humans. *Heliyon*, *6*(9). https://doi.org/10.1016/J.HELIYON.2020.E04691

Brown, A. M. C., Coupland, G. M., & Willetts, N. S. (1984). Characterization of IS*46*, an insertion sequence found on two IncN plasmids. *Journal of Bacteriology*, *159*(2), 472–481. https://doi.org/10.1128/JB.159.2.472-481.1984

Brown, N., & Evans, L. (1991). Transposition in prokaryotes: transposon Tn*501*. *Research in Microbiology*, *142*(6), 689–700. https://doi.org/10.1016/0923-2508(91)90082-L

Cangelosi, G. A., & Meschke, J. S. (2014). Dead or Alive: Molecular Assessment of Microbial Viability. *Applied and Environmental Microbiology*, *80*(19), 5884–5891. https://doi.org/10.1128/AEM.01763-14

Cerbino, G. N., Traglia, G. M., Ayala Nuñez, T., Parmeciano Di Noto, G., Ramírez, M.

S., Centrón, D., Iriarte, A., & Quiroga, C. (2023). Comparative genome analysis of the genus *Shewanella* unravels the association of key genetic traits with known and potential pathogenic lineages. *Frontiers in Microbiology*, *14*. https://doi.org/10.3389/FMICB.2023.1124225/FULL

Chen, I., & Dubnau, D. (2004). DNA uptake during bacterial transformation. *Nature Reviews Microbiology*, *2 (3)*, 241–249. https://doi.org/10.1038/nrmicro844

Consuegra, J., Gaffé, J., Lenski, R. E., Hindré, T., Barrick, J. E., Tenaillon, O., & Schneider, D. (2021). Insertion-sequence-mediated mutations both promote and constrain evolvability during a long-term experiment with bacteria. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-21210-7

Dai, Z., Su, W., Chen, H., Barberán, A., Zhao, H., Yu, M., Yu, L., Brookes, P. C., Schadt, C. W., Chang, S. X., & Xu, J. (2018). Long-term nitrogen fertilization decreases bacterial diversity and favors the growth of *Actinobacteria* and *Proteobacteria* in agro-ecosystems across the globe. *Global Change Biology*, *24*(8), 3452–3461. https://doi.org/10.1111/GCB.14163

Dasgupta, D., Ries, M., Walter, K., Zitnick-Anderson, K., Camuy-Vélez, L. A., Gasch, C., & Banerjee, S. (2023). Cover cropping reduces the negative effect of salinity on soil microbiomes. *Journal of Sustainable Agriculture and Environment*, *2*(2), 140–152. https://doi.org/10.1002/SAE2.12054

Debrabandere, K. (2023). *Contribution of the insertion sequence element IS1071 to microbial community adaptation to environmental stresses in the context of global change* [non-published master thesis]. KU Leuven.

De Sousa, J. M., Lourenç, M., & Gordo, I. (2023). Horizontal gene transfer among host-associated microbes. *Cell Host & Microbe*, *31*(4), 513–527. https://doi.org/10.1016/j.chom.2023.03.017

Degnan, P. H., & Ochman, H. (2011). Illumina-based analysis of microbial community diversity. *The ISME Journal*, *6*(1), 183–194. https://doi.org/10.1038/ismej.2011.74

Din, I. U., Muhammad, S., & Rehman, I. ur. (2022). Heavy metal(loid)s contaminations in soils of Pakistan: a review for the evaluation of human and ecological risks assessment and spatial distribution. *Environmental Geochemistry and Health*, *45*(5), 1991–2012. https://doi.org/10.1007/S10653-022-01312-X

Doggett, N. A., Smith, C. L., & Cantor, C. R. (1992). The effect of DNA concentration on mobility in pulsed field gel electrophoresis. *Nucleic Acids Research*, *20*(4), 859–864. https://doi.org/10.1093/nar/20.4.859

Dogra, C., Raina, V., Pal, R., Suar, M., Lal, S., Gartemann, K. H., Holliger, C., Van Der Meer, J. R., & Lal, R. (2004). Organization of lin Genes and IS*6100* among Different Strains of Hexachlorocyclohexane-Degrading *Sphingomonas paucimobilis*: Evidence for Horizontal Gene Transfer. *Journal of Bacteriology*, *186*(8). https://doi.org/10.1128/JB.186.8.2225-2235.2004

Du, C. M., Shang, C., Wang, T., Xiao, X. L., Ruan, J. J., Ma, D. Y., Tang, Y. T., & Qiu, R. L. (2017). Study of the Process and Mechanism of the Remediation of Phenol Contaminated Soil by Plasma Vibrated Bed. *Plasma Chemistry and Plasma Processing*, *37*(6), 1635–1653. https://doi.org/10.1007/S11090-017-9850-6/FIGURES/12

Dunon, V., Bers, K., Lavigne, R., Top, E. M., & Springael, D. (2018). Targeted

metagenomics demonstrates the ecological role of IS*1071* in bacterial community adaptation to pesticide degradation. *Environmental Microbiology*, *20*(11), 4091–4111. https://doi.org/10.1111/1462-2920.14404

Dunon, V., Holmsgaard, P. N., Dealtry, S., Lavigne, R., Sørensen, S. J., Smalla, K., Top, E. M., & Springael, D. (2022). Long-Range PCR Reveals the Genetic Cargo of IncP-1 Plasmids in the Complex Microbial Community of an On-Farm Biopurification System Treating Pesticide-Contaminated Wastewater. *Applied and Environmental Microbiology*, *88*(3). https://doi.org/10.1128/aem.01648-21

Dunon, V., Sniegowski, K., Bers, K., Lavigne, R., Smalla, K., & Springael, D. (2013). High prevalence of IncP-1 plasmids and IS*1071* insertion sequences in on-farm biopurification systems and other pesticide-polluted environments. *FEMS Microbiology Ecology*, *86*(3), 415–431. https://doi.org/10.1111/1574-6941.12173

Faiz, Y., Tufail, M., Javed, M. T., Chaudhry, M. M., & Naila-Siddique. (2009). Road dust pollution of Cd, Cu, Ni, Pb and Zn along Islamabad Expressway, Pakistan. *Microchemical Journal*, *92*(2), 186–192. https://doi.org/10.1016/J.MICROC.2009.03.009

Frost, L. S., Leplae, R., Summers, A. O., & Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, *3*(9), 722–732. https://doi.org/10.1038/nrmicro1235

Fu, X. Y., Cheng, Z. C., Ni, H. W., & Zhang, R. T. (2023). Latitude variations of soil bacterial community diversity and composition in three typical forests of temperate, northeastern of China. *Frontiers in Earth Science*, *10*. https://doi.org/10.3389/FEART.2022.1096931/BIBTEX

Fuchu, G., Ohtsubo, Y., Ito, M., Miyazaki, R., Ono, A., Nagata, Y., & Tsuda, M. (2008). Insertion sequence-based cassette PCR: cultivation-independent isolation of g-hexachlorocyclohexane-degrading genes from soil DNA. *Applied Genetics and Molecular Biotechnology*, *79*, 627–632. https://doi.org/10.1007/s00253-008-1463-9

Gai, Z., Wang, X., Liu, X., Tai, C., Tang, H., He, X., Wu, G., Deng, Z., & Xu, P. (2010). The Genes Coding for the Conversion of Carbazole to Catechol Are Flanked by IS*6100* Elements in *Sphingomonas sp.* Strain XLDN2-5. *PLoS ONE*, *5*(4). https://doi.org/10.1371/journal.pone.0010018

Ghanam, J., Chetty, V. K., Barthel, L., Reinhardt, D., Hoyer, P. F., & Thakur, B. K. (2022). DNA in extracellular vesicles: from evolution to its current application in health and disease. *Cell and Bioscience*, *12*(1). https://doi.org/10.1186/S13578-022-00771-0/METRICS

Giani, A. M., Gallo, G. R., Gianfranceschi, L., & Formenti, G. (2020). Long walk to genomics: History and current approaches to genome sequencing and assembly. *Computational and Structural Biotechnology Journal*, *18*, 9–19. https://doi.org/10.1016/J.CSBJ.2019.11.002

Glansdorff, N., Charlier, D., & Zafarullah, M. (1981). Activation of gene expression by IS*2* and IS*3*. *Cold Spring Harbor Symposia on Quantitative Biology*, *45*, 153–156. https://doi.org/10.1101/SQB.1981.045.01.024

Guo, Y., Cheng, S., Fang, H., Yang, Y., Li, Y., Shi, F., & Zhou, Y. (2023). Copper and

cadmium co-contamination affects soil bacterial taxonomic and functional attributes in paddy soils. *Environmental Pollution*, *329*. https://doi.org/10.1016/J.ENVPOL.2023.121724

Gupta, A., Singh, U. B., Sahu, P. K., Paul, S., Kumar, A., Malviya, D., Singh, S., Kuppusamy, P., Singh, P., Paul, D., Rai, J. P., Singh, H. V., Manna, M. C., Crusberg, T. C., Kumar, A., & Saxena, A. K. (2022). Linking Soil Microbial Diversity to Modern Agriculture Practices: A Review. *International Journal of Environmental Research and Public Health*, *19*(5). https://doi.org/10.3390/IJERPH19053141

Han, S. I., Lee, J. C., Lee, H. J., & Whang, K. S. (2013). *Planifilum composti sp*. *nov.,* a thermophile isolated from compost. *International Journal of Systematic and Evolutionary Microbiology*, *63*, 4557–4561. https://doi.org/10.1099/IJS.0.053199-0/CITE/REFWORKS

Hasegawa, H., Suzuki, E., & Maeda, S. (2018). Horizontal Plasmid Transfer by Transformation in *Escherichia coli*: Environmental Factors and Possible Mechanisms. *Frontiers in Microbiology*, *9*. https://doi.org/10.3389/FMICB.2018.02365

Hehemann, J. H., Correc, G., Barbeyron, T., Helbert, W., Czjzek, M., & Michel, G. (2010). Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature*, *464*, 908–912. https://doi.org/10.1038/NATURE08937

Hemme, C. L., Green, S. J., Rishishwar, L., Prakash, O., Pettenato, A., Chakraborty, R., Deutschbauer, A. M., Van Nostrand, J. D., Wu, L., He, Z., Jordan, I. K., Hazen, T. C., Arkin, A. P., Kostka, J. E., & Zhou, J. (2016). Lateral Gene Transfer in a Heavy Metal-Contaminated-Groundwater Microbial Community. *MBio*, *7*(2). https://doi.org/10.1128/MBIO.02234-15

Hoebeeck, J., Speleman, F., & Vandesompele, J. (2007). Real-time quantitative PCR as an alternative to southern blot or fluorescence in situ hybridization for detection of gene copy number changes. *Methods in Molecular Biology*, *353*, 205–226. https://doi.org/10.1385/1-59745-229-7:205

Huddleston, J. R. (2014). Horizontal gene transfer in the human gastrointestinal tract: potential spread of antibiotic resistance genes. *Infection and Drug Resistance*, *7*, 167–176. https://doi.org/10.2147/IDR.S48820

Ikuma, K., & Gunsch, C. K. (2012). Genetic bioaugmentation as an effective method for in situ bioremediation: functionality of catabolic plasmids following conjugal transfers. *Bioengineered*, *3*(4), 236–241. https://doi.org/10.4161/BIOE.20551

JCM 19685 - JCM Catalogue. (2007). https://www.jcm.riken.jp/cgi-bin/jcm/jcm_number?JCM=19685

Jiang, W., Zhang, M., Gao, S., Zhu, Q., Qiu, J., Yan, X., Xin, F., Jiang, M., & Hong, Q. (2022). Comparative Genomic Analysis of Carbofuran-Degrading Sphingomonads Reveals the Carbofuran Catabolism Mechanism in *Sphingobium sp.* Strain CFD-1. *Applied and Environmental Microbiology*, *88*(22). https://doi.org/10.1128/AEM.01024-22

Kato, K., Ohtsuki, K., Mitsuda, H., Yomo, T., Negoro, S., & Urabe, I. (1994). Insertion sequence IS*6100* on plasmid pOAD2, which degrades nylon oligomers. *Journal of Bacteriology*, *176*(4), 1197–1200. https://doi.org/10.1128/JB.176.4.1197-

1200.1994

Kawakami, K., Largaespada, D. A., & Ivics, Z. (2017). Special Issue: Transposable Elements Transposons As Tools for Functional Genomics in Vertebrate Models. *Trends in Genetics*, *33*(11). https://doi.org/10.1016/j.tig.2017.07.006

Keeling, P. J., & Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, *9*(8), 605–618. https://doi.org/10.1038/nrg2386

Khan, A. G. (2001). Relationships between chromium biomagnification ratio, accumulation factor, and *mycorrhizae* in plants growing on tannery effluent-polluted soil. *Environment International*, *26*, 417–423. https://doi.org/10.1016/s0160-4120(01)00022-8

Kojima, K. K., & Bao, W. (2023). IS*481EU* Shows a New Connection between Eukaryotic and Prokaryotic DNA Transposons. *Biology*, *12*(3). https://doi.org/10.3390/BIOLOGY12030365/S1

Kubier, A., Wilkin, R. T., & Pichler, T. (2019). Cadmium in soils and groundwater: A review. *Applied Geochemistry : Journal of the International Association of Geochemistry and Cosmochemistry*, *108*. https://doi.org/10.1016/J.APGEOCHEM.2019.104388

Landesman, W. J., Freedman, Z. B., & Nelson, D. M. (2019). Seasonal, sub-seasonal and diurnal variation of soil bacterial community composition in a temperate deciduous forest. *FEMS Microbiology Ecology*, *95*(2), 2. https://doi.org/10.1093/FEMSEC/FIZ002

Lee, C., Kim, J., Shin, G., & Hwang, S. (2006). Absolute and relative qPCR quantification of plasmid copy number in *Escherichia coli. Journal of Biotechnology*, *123*, 273–280. https://doi.org/10.1016/j.jbiotec.2005.11.014

Lee, S. D. (2006). *Brevibacterium samyangense sp. nov*., an *actinomycete* isolated from a beach sediment. *International Journal of Systematic and Evolutionary Microbiology*, *56*, 1889–1892. https://doi.org/10.1099/ijs.0.64269-0

Leisinger, T. (1983). Microorganisms and xenobiotic compounds. *Experientia*, *39*(11), 1183–1191. https://doi.org/10.1007/BF01990355

Lewis, W. H., Tahon, G., Geesink, P., Sousa, D. Z., & G Ettema, T. J. (2021). Innovations to culturing the uncultured microbial majority. *Nature Reviews Microbiology*, *19*, 225–240. https://doi.org/10.1038/s41579-020-00458-8

Li, Y., Zhang, M., Xu, R., Lin, H., Sun, X., Xu, F., Gao, P., Kong, T., Xiao, E., Yang, N., & Sun, W. (2021). Arsenic and antimony co-contamination influences on soil microbial community composition and functions: Relevance to arsenic resistance and carbon, nitrogen, and sulfur cycling. *Environment International*, *153*. https://doi.org/10.1016/J.ENVINT.2021.106522

Liu, H., Han, H., Zhang, C., Yu, X., Nie, W., Shao, Q., Yang, P., Li, X., Yang, Y., & Cao, H. (2023). Patterns of bacterial distance decay and community assembly in different land-use types as influenced by tillage management and soil layers. *Ecotoxicology and Environmental Safety*, *266*. https://doi.org/10.1016/J.ECOENV.2023.115595

Liu, M., Qi, H., Luo, X., Dai, J., Peng, F., & Fang, C. (2011). *Cesiribacter roseus sp. nov.*, a pink-pigmented bacterium isolated from desert sand. *International Journal*

of Systematic and Evolutionary Microbiology, *62*(1), 96–99. https://doi.org/10.1099/IJS.0.028423-0/CITE/REFWORKS

Lopes, L. de S., Mendes, L. W., Antunes, J. E. L., de Souza Oliveira, L. M., Melo, V. M. M., de Araujo Pereira, A. P., da Costa, A. F., de Paula Oliveira, J., Martínez, C. R., Figueiredo, M. do V. B., & Araujo, A. S. F. (2021). Distinct bacterial community structure and composition along different cowpea producing ecoregions in Northeastern Brazil. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-020-80840-x

Louisson, Z., Hermans, S. M., Buckley, H. L., Case, B. S., Taylor, M., Curran-Cournane, F., & Lear, G. (2023). Land use modification causes slow, but predictable, change in soil microbial community composition and functional potential. *Environmental Microbiome*, *18*(1). https://doi.org/10.1186/S40793-023-00485-X/FIGURES/6

Ma, Y., Wang, J., Liu, Y., Wang, X., Zhang, B., Zhang, W., Chen, T., Liu, G., Xue, L., & Cui, X. (2023). *Nocardioides*: "Specialists" for Hard-to-Degrade Pollutants in the Environment. *Molecules*, *28*(21). https://doi.org/10.3390/MOLECULES28217433

Madamwar, D., Tsuda, M., Solyanikova, I., Chen, S., Mishra, S., Lin, Z., Pang, S., Zhang, W., & Bhatt, P. (2021). Recent Advanced Technologies for the Characterization of Xenobiotic-Degrading Microorganisms and Microbial Communities. *Frontiers in Bioengineering and Biotechnology*, *9*. https://doi.org/10.3389/fbioe.2021.632059

Mahillon, J., & Chandler, M. (1998). Insertion sequences. *Microbiology and Molecular Biology Reviews : MMBR*, *62*(3), 725–774. https://doi.org/10.1128/MMBR.62.3.725-774.1998

Malik, R. N., Jadoon, W. A., & Husain, S. Z. (2010). Metal contamination of surface soils of industrial city Sialkot, Pakistan: A multivariate and GIS approach. *Environmental Geochemistry and Health*, *32*(3), 179–191. https://doi.org/10.1007/S10653-009-9274-1/TABLES/4

Marghoob, M. U., Rodriguez-Sanchez, A., Imran, A., Mubeen, F., & Hoagland, L. (2022). Diversity and functional traits of indigenous soil microbial flora associated with salinity and heavy metal concentrations in agricultural fields within the Indus Basin region, Pakistan. *Frontiers in Microbiology*, *13*. https://doi.org/10.3389/FMICB.2022.1020175/FULL

Maszenan, A. M., Jiang, H. L., Tay, J. H., Schumann, P., Kroppenstedt, R. M., & Tay, S. T. L. (2007). *Granulicoccus phenolivorans gen. nov.*, *sp. nov.,* a Gram-positive, phenol-degrading *coccus* isolated from phenol-degrading aerobic granules. *International Journal of Systematic and Evolutionary Microbiology*, *57*(4), 730–737. https://doi.org/10.1099/IJS.0.64671-0/CITE/REFWORKS

Mawang, C. I., Azman, A. S., Fuad, A. S. M., & Ahamad, M. (2021). *Actinobacteria:* An eco-friendly and promising technology for the bioaugmentation of contaminants. *Biotechnology Reports*, *32*. https://doi.org/10.1016/J.BTRE.2021.E00679

Mayerhofer, J., Wächter, D., Calanca, P., Kohli, L., Roth, T., Meuli, R. G., & Widmer, F. (2021). Environmental and Anthropogenic Factors Shape Major Bacterial Community Types Across the Complex Mountain Landscape of Switzerland. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/FMICB.2021.581430/FULL

Mcdonald, J. F. (1993). Evolution and consequences of transposable elements. *Current Opinion in Genetics and Development*, *3*, 855–864. https://doi.org/10.1016/0959-437X(93)90005-A

Mori, K., Suzuki, K. I., Yamaguchi, K., Urabe, T., & Hanada, S. (2015). *Thiogranum longum gen. nov., sp. nov.,* an obligately chemolithoautotrophic, sulfur-oxidizing bacterium of the family *Ectothiorhodospiraceaeisolated* from a deep-sea hydrothermal field, and an emended description of the genus *Thiohalomonas. International Journal of Systematic and Evolutionary Microbiology*, *65*(1), 235–241. https://doi.org/10.1099/IJS.0.070599-0/CITE/REFWORKS

Muñoz-López, M., & García-Pérez, J. L. (2010). DNA Transposons: Nature and Applications in Genomics. *Current Genomics*, *11*, 115–128. https://doi.org/10.2174/138920210790886871

Nagata, Y., Endo, R., Ito, M., Ohtsubo, Y., & Tsuda, M. (2007). Aerobic degradation of lindane (γ-hexachlorocyclohexane) in bacteria and its biochemical and molecular basis. *Applied Microbiology and Biotechnology*, *76*(4), 741–752. https://doi.org/10.1007/S00253-007-1066-X/FIGURES/4

Nardelli, M., Scalzo, P. M., Ramírez, M. S., Quiroga, M. P., Cassini, M. H., & Centrón, D. (2012). Class 1 Integrons in Environments with Different Degrees of Urbanization. *PLoS ONE*, *7*(6). https://doi.org/10.1371/JOURNAL.PONE.0039223

Nazari, M. T., Machado, B. S., Marchezi, G., Crestani, L., Ferrari, V., Colla, L. M., & Piccin, J. S. (2022). Use of soil *actinomycetes* for pharmaceutical, food, agricultural, and environmental purposes. *3 Biotech*, *12*(9). https://doi.org/10.1007/S13205-022-03307-Y

Nazarian, P., & Boedicker, T. F. (2018). Modeling Multispecies Gene Flow Dynamics Reveals the Unique Roles of Different Horizontal Gene Transfer Mechanisms. *Evolutionary and Genomic Microbiology*, *9*. https://doi.org/10.3389/fmicb.2018.02978

Nesmelova, I. V., & Hackett, P. B. (2010). DDE transposases: Structural similarity and diversity. *Advanced Drug Delivery Reviews*, *62*(12), 1187–1195. https://doi.org/10.1016/J.ADDR.2010.06.006

Nicolas, E., Lambin, M., Dandoy, D., Galloy, C., Nguyen, N., Oger, C. A., & Hallet, B. (2015). The Tn*3* family of Replicative Transposons . *Microbiology Spectrum*, *3*(4). https://doi.org/10.1128/MICROBIOLSPEC.MDNA3-0060-2014

Nielsen, T. K., Horemans, B., Lood, C., T'Syen, J., van Noort, V., Lavigne, R., Ellegaard-Jensen, L., Hylling, O., Aamand, J., Springael, D., & Hansen, L. H. (2021). The complete genome of 2,6-dichlorobenzamide (BAM) degrader *Aminobacter sp.* MSH1 suggests a polyploid chromosome, phylogenetic reassignment, and functions of plasmids. *Scientific Reports*, *11*(1). https://doi.org/10.1038/S41598-021-98184-5

Nojiri, H., Shintani, M., & Omori, T. (2004). MINI-REVIEW Divergence of mobile genetic elements involved in the distribution of xenobiotic-catabolic capacity. *Appl Microbiol Biotechnol*, *64*, 154–174. https://doi.org/10.1007/s00253-003-1509-y

Nugent, A., & Allison, S. D. (2022). A framework for soil microbial ecology in urban

ecosystems. *Ecosphere*, *13*(3). https://doi.org/10.1002/ECS2.3968

Ochman, H., Lawrence, J. G., & Grolsman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, *405*, 299–304. https://doi.org/10.1038/35012500

Pan, S. Y., He, K. H., Lin, K. T., Fan, C., & Chang, C. T. (2022). Addressing nitrogenous gases from croplands toward low-emission agriculture. *Npj Climate and Atmospheric Science*, *5*(1). https://doi.org/10.1038/s41612-022-00265-3

Parks, A. R., & Peters, J. E. (2018). Conservative Site-Specific Recombination. In *Molecular Life Sciences* (pp. 119–127). https://doi.org/10.1007/978-1-4614-1531-2_165

Peddle, S. D., Cando-Dumancela, C., Krauss, S. L., Liddicoat, C., Sanders, A., & Breed, M. F. (2024). Agricultural land-use legacies affect soil bacterial communities following restoration in a global biodiversity hotspot. *Biological Conservation*, *290*. https://doi.org/10.1016/J.BIOCON.2023.110437

Peng, W., Zhu, Y., Song, M., Du, H., Song, T., Zeng, F., Zhang, F., Wang, K., Luo, Y., Lan, X., & Zhang, J. (2019). The spatial distribution and drivers of soil microbial richness and diversity in a karst broadleaf forest. *Forest Ecology and Management*, *449*. https://doi.org/10.1016/J.FORECO.2019.03.033

Pereira e Silva, M. C., Schloter-Hai, B., Schloter, M., van Elsas, J. D., & Salles, J. F. (2013). Temporal Dynamics of Abundance and Composition of Nitrogen-Fixing Communities across Agricultural Soils. *PLoS ONE*, *8*(9). https://doi.org/10.1371/JOURNAL.PONE.0074500

Philippot, L., Griffiths, B. S., & Langenheder, S. (2021). Microbial Community Resilience across Ecosystems and Multiple Disturbances. *Microbiology and Molecular Biology Reviews*, *85*(2). https://doi.org/10.1128/MMBR.00026-20

Pimpinelli, S., Berlocot, M., Fantit, L., Dimitri, P., Bonaccorsi, S., Marchetti, E., Caizzit, R., Caggeset, C., & Gatrri, M. (1995). Transposable elements are stable structural components of *Drosophila melanogaster* heterochromatin. *National Academy of Sciences*, *92*(9), 3804–3808. https://www.jstor.org/stable/2367435

Plassart, P., Prévost-Bouré, N. C., Uroz, S., Dequiedt, S., Stone, D., Creamer, R., Griffiths, R. I., Bailey, M. J., Ranjard, L., & Lemanceau, P. (2019). Soil parameters, land use, and geographical distance drive soil bacterial communities along a European transect. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-018-36867-2

Pratscher, J. (2021). Extraction of Microbial Cells from Environmental Samples for FISH Approaches. *Methods in Molecular Biology*, *2246*. https://doi.org/10.1007/978-1-0716-1115-9_19

Putri, A. L., & Setiawan, R. (2019). Isolation and screening of *actinomycetes* producing cellulase and xylanase from Mamasa soil, West Sulawesi. *Earth Environmental Sciences*, *308*. https://doi.org/10.1088/1755-1315/308/1/012035

Qaswar, M., Yiren, L., Liu, K., Zhenzhen, L., Hongqian, H., Lan, X., Jianhua, J., Ahmed, W., Lisheng, L., Mouazen, A. M., & Huimin, Z. (2022). Interaction of Soil Nutrients and Arsenic (As) in Paddy Soil in a Long-Term Fertility Experiment. *Sustainability*, *14*(19). https://doi.org/10.3390/SU141911939

Qin, X., Zhai, L., Khoshnevisan, B., Pan, J., & Liu, H. (2022). Restriction of biosolids

returning to land: Fate of antibiotic resistance genes in soils after long-term biosolids application. *Environmental Pollution*, *301*. https://doi.org/10.1016/J.ENVPOL.2022.119029

Razavi, M., Kristiansson, E., Flach, C.-F., & Larsson, D. G. J. (2020). The Association between Insertion Sequences and Antibiotic Resistance Genes. *MSphere*, *5*(5). https://doi.org/10.1128/MSPHERE.00418-20

Rehman, H. ur, Ahmed, S., Ur Rahman, M., & Mehmood, M. S. (2022). Arsenic contamination, induced symptoms, and health risk assessment in groundwater of Lahore, Pakistan. *Environmental Science and Pollution Research*, *29*(33), 49796–49807. https://doi.org/10.1007/S11356-022-19405-6/TABLES/4

Reif, H. J., & Saedler, H. (1975). IS*1* is involved in deletion formation in the gal region of *E. coli* K12. *Molecular & General Genetics*, *137*(1), 17–28. https://doi.org/10.1007/BF00332538

Rice, P., Craigie, R., & Davies, D. R. (1996). Retroviral integrases and their cousins. *Current Opinion in Structural Biology*, *6*(1), 76–83. https://doi.org/10.1016/S0959-440X(96)80098-4

Robinson, D. G., Lee, M.-C., & Marx, C. J. (2012). OASIS: an automated program for global investigation of bacterial and archaeal insertion sequences. *Nucleic Acids Research*, *40*(22). https://doi.org/10.1093/nar/gks778

Romdhane, S., Spor, A., Banerjee, S., Breuil, M. C., Bru, D., Chabbi, A., Hallin, S., van der Heijden, M. G. A., Saghai, A., & Philippot, L. (2022). Land-use intensification differentially affects bacterial, fungal and protist communities and decreases microbiome network complexity. *Environmental Microbiomes*, *17*(1), 1–15. https://doi.org/10.1186/S40793-021-00396-9/FIGURES/5

Santini, L., Belmaker, J., Costello, M. J., Pereira, H. M., Rossberg, A. G., Schipper, A. M., Ceauşu, S., Dornelas, M., Hilbers, J. P., Hortal, J., Huijbregts, M. A. J., Navarro, L. M., Schiffers, K. H., Visconti, P., & Rondinini, C. (2017). Assessing the suitability of diversity metrics to detect biodiversity change. *Biological Conservation*, *213*, 341–350. https://doi.org/10.1016/J.BIOCON.2016.08.024

Sazykin, I., Khmelevtsova, L., Azhogina, T., & Sazykina, M. (2023). Heavy Metals Influence on the Bacterial Community of Soils: A Review. *Agriculture*, *13*(3). https://doi.org/10.3390/AGRICULTURE13030653

Sheik, C. S., Mitchell, T. W., Rizvi, F. Z., Rehman, Y., Faisal, M., Hasnain, S., McInerney, M. J., & Krumholz, L. R. (2012). Exposure of Soil Microbial Communities to Chromium and Arsenic Alters Their Diversity and Structure. *PLoS ONE*, *7*(6). https://doi.org/10.1371/JOURNAL.PONE.0040059

Shi, A., Fan, F., & Broach, J. R. (2022). Bioenergy/Biofuels/Biochemicals-Mini-Review Microbial adaptive evolution. *Journal of Industrial Microbiology and Biotechnology*, *49*(2). https://doi.org/10.1093/jimb/kuab076

Siefert, J. L. (2009). Defining the mobilome. *Methods in Molecular Biology*, *532*, 13–27. https://doi.org/10.1007/978-1-60327-853-9_2

Siguier P. *et al.* (2006) ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 34: D32-D36 (https://pubmed.ncbi.nlm.nih.gov/16381877/) and the database URL (http://www-is.biotoul.fr).

Siguier, P., Gourbeyre, E., & Chandler, M. (2014). Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiology Reviews*, *38(5)*, 865–891. https://doi.org/10.1111/1574-6976.12067

Siguier, P., Gourbeyre, E., Varani, A., Ton-Hoang, B., & Chandler, M. (2015). Everyman's Guide to Bacterial Insertion Sequences. *Mobile DNA III*, *3*(2), 555–590. https://doi.org/10.1128/9781555819217.CH26

Singh, S., & Hiranmai, R. Y. (2021). Monitoring and molecular characterization of bacterial species in heavy metals contaminated roadside soil of selected region along NH 8A, Gujarat. *Heliyon*, *7*(11). https://doi.org/10.1016/J.HELIYON.2021.E08284

Sinzelle, L., Izsvμk, Z., & Ivics, Z. (2009). Review Molecular domestication of transposable elements: From detrimental parasites to useful host genes. *Cellular and Molecular Life Sciences*, *66*(6), 1073 – 1093. https://doi.org/10.1007/s00018-009-8376-3

Smith, B., & Dyson, P. (1995). Inducible transposition in *Streptomyces lividans* of insertion sequence IS*6100* from *Mycobacterium fortuitum. Molecular Microbiology*, *18*(5), 933–941. https://doi.org/10.1111/J.1365-2958.1995.18050933.X

Sniegowski, K., Bers, K., Van Goetem, K., Ryckeboer, J., Jaeken, P., Spanoghe, P., & Springael, D. (2011). Improvement of pesticide mineralization in on-farm biopurification systems by bioaugmentation with pesticide-primed soil. *FEMS Microbiology Ecology*, *76*(1), 64–73. https://doi.org/10.1111/J.1574-6941.2010.01031.X

Soucy, S. M., Huang, J., & Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, *16*(8), 472–482. https://doi.org/10.1038/nrg3962

Soumare, A., Diedhiou, A. G., Thuita, M., Hafidi, M., Ouhdouch, Y., Gopalakrishnan, S., & Kouisni, L. (2020). Exploiting Biological Nitrogen Fixation: A Route Towards a Sustainable Agriculture. *Plants*, *9*(8), 1–22. https://doi.org/10.3390/PLANTS9081011

Stephanou, C., Omirou, M., Philippot, L., Zissimos, A. M., Christoforou, I. C., Trajanoski, S., Oulas, A., & Ioannides, I. M. (2021). Land use in urban areas impacts the composition of soil bacterial communities involved in nitrogen cycling. A case study from Lefkosia (Nicosia) Cyprus. *Scientific Reports*, *11*(1). https://doi.org/10.1038/S41598-021-87623-Y

Suman, J., Rakshit, A., Ogireddy, S. D., Singh, S., Gupta, C., & Chandrakala, J. (2022). Microbiome as a Key Player in Sustainable Agriculture and Human Health. *Frontiers in Soil Science*, *2*. https://doi.org/10.3389/FSOIL.2022.821589/BIBTEX

Summers, A. O., Wireman, J., Vimy, M. J., Lorscheider, F. L., Marshall, B., Levy, S. B., Bennett, S., & Billard, L. (1993). Mercury released from dental "silver" fillings provokes an increase in mercury- and antibiotic-resistant bacteria in oral and intestinal floras of primates. *Antimicrobial Agents and Chemotherapy*, *37*(4), 825–834. https://doi.org/10.1128/AAC.37.4.825

Sung, P., & Klein, H. (2006). Mechanism of homologous recombination: mediators and helicases take on regulatory functions. *Nature Reviews Molecular Cell Biology*,

7(10), 739–750. https://doi.org/10.1038/nrm2008

Syed, A., Raza, T., Bhatti, T. T., & Eash, N. S. (2022). Climate Impacts on the agricultural sector of Pakistan: Risks and solutions. *Environmental Challenges*, *6*. https://doi.org/10.1016/J.ENVC.2021.100433

Syed, J. H., Malik, R. N., Li, J., Zhang, G., & Jones, K. C. (2013). Levels, distribution and air–soil exchange fluxes of polychlorinated biphenyls (PCBs) in the environment of Punjab Province, Pakistan. *Ecotoxicology and Environmental Safety*, *97*, 189–195. https://doi.org/10.1016/J.ECOENV.2013.06.005

Szoboszlay, M. A., Dohrmann, A. B., Poeplau, C., Don, A., & Tebbe, C. C. (2017). Impact of land-use change and soil organic carbon quality on microbial diversity in soils across Europe. *FEMS Microbiology Ecology*, *93*. https://doi.org/10.1093/femsec/fix146

Tabata, M., Endo, R., Ito, M., Ohtsubo, Y., Kumar, A., Tsuda, M., & Nagata, Y. (2011). The lin Genes for-Hexachlorocyclohexane Degradation in *Sphingomonas sp.* MM-1 Proved to Be Dispersed across Multiple Plasmids. *Bioscience Biotechnology and Biochemistry*, *75*(3), 466–472. https://doi.org/10.1271/bbb.100652

Tansirichaiya, S., Mullany, P., & Roberts, A. P. (2016). PCR-based detection of composite transposons and translocatable units from oral metagenomic DNA. *FEMS Microbiology Letters*, *363*(18). https://doi.org/10.1093/femsle/fnw195

Telenti, A., Imboden, P., Marchesi, F., Matter, L., Schopfer, K., Bodmer, T., Lowrie, D., Colston, M. J., & Cole, S. (1993). Detection of rifampicin-resistance mutations in *Mycobacterium tuberculosis*. *The Lancet*, *341*(8846), 647–651. https://doi.org/10.1016/0140-6736(93)90417-F

Tian, Q., Taniguchi, T., Shi, W. Y., Li, G., Yamanaka, N., & Du, S. (2017). Land-use types and soil chemical properties influence soil microbial communities in the semiarid Loess Plateau region in China. *Scientific Reports*, *7*(1). https://doi.org/10.1038/srep45289

Top, E. M., & Springael, D. (2003). The role of mobile genetic elements in bacterial adaptation to xenobiotic organic compounds. *Current Opinion in Biotechnology*, *14*(3), 262–269. https://doi.org/10.1016/S0958-1669(03)00066-1

Touchon, M., & Rocha, E. P. C. (2007). Causes of insertion sequences abundance in prokaryotic genomes. *Molecular Biology and Evolution*, *24*(4), 969–981. https://doi.org/10.1093/MOLBEV/MSM014

Treves, D. S., Manning, S., & Adams, J. (1998). Repeated Evolution of an Acetate-Crossfeeding Polymorphism in Long-Term Populations of *Escherichia coli. Molecular Biology and Evolution*, *15*(7), 789–797. https://doi.org/10.1093/oxfordjournals.molbev.a025984

Trieu-Cuot, P., & Courvalin, P. (1985). Transposition behavior of IS*15* and its progenitor IS*15*-Δ: Are cointegrates exclusive end products. *Plasmid*, *14*(1), 80–89. https://doi.org/10.1016/0147-619X(85)90034-4

Trzilova, D., & Tamayo, R. (2021). Site-specific recombination – how simple DNA inversions produce complex phenotypic heterogeneity in bacterial populations. *Trends in Genetics*, *37*(1). https://doi.org/10.1016/J.TIG.2020.09.004

Ullah, S., He, P., Ai, C., Zhao, S., Ding, W., Song, D., Zhang, J., Huang, S., Abbas, T.,

& Zhou, W. (2020). How Do Soil Bacterial Diversity and Community Composition Respond under Recommended and Conventional Nitrogen Fertilization Regimes. *Microorganisms*, *8*(8). https://doi.org/10.3390/MICROORGANISMS8081193

Van der Meer, J. R., Zehnder, A. J. B., & De Vos, W. M. (1991). Identification of a novel composite transposable element, Tn*5280*, carrying chlorobenzene dioxygenase genes of *Pseudomonas sp.* strain P51. *Journal of Bacteriology*, *173*(22), 7077–7083. https://doi.org/10.1128/JB.173.22.7077-7083.1991

Vandecraen, J., Chandler, M., Aertsen, A., & Van Houdt, R. (2017). The impact of insertion sequences on bacterial genome plasticity and adaptability. *Critical Reviews in Microbiology*, *43*(6), 709–730. https://doi.org/10.1080/1040841X.2017.1303661

Varani, A., He, S., Siguier, P., Ross, K., & Chandler, M. (2021). The IS*6* family, a clinically important group of insertion sequences including IS*26*. *Mobile DNA*, *12*(1). https://doi.org/10.1186/S13100-021-00239-X/FIGURES/11

Vig, K., Megharaj, M., Sethunathan, N., & Naidu, R. (2003). Bioavailability and toxicity of cadmium to microorganisms and their activities in soil: a review. *Advances in Environmental Research*, *8*(1), 121–135. https://doi.org/10.1016/S1093-0191(02)00135-1

Wagner, B. D., Grunwald, G. K., Zerbe, G. O., Mikulich-Gilbertson, S. K., Robertson, C. E., Zemanick, E. T., & Harris, J. K. (2018). On the use of diversity measures in longitudinal sequencing studies of microbial communities. *Frontiers in Microbiology*, *9*. https://doi.org/10.3389/FMICB.2018.01037/BIBTEX

Walters, K. E., & Martiny, J. B. H. (2020). Alpha-, beta-, and gamma-diversity of bacteria varies across habitats. *PLoS ONE*, *15*(9). https://doi.org/10.1371/JOURNAL.PONE.0233872

Wang, X., Sial, M. U., Bashir, M. A., Bilal, M., Raza, Q. U. A., Raza, H. M. A., Rehim, A., & Geng, Y. (2022). Pesticides Xenobiotics in Soil Ecosystem and Their Remediation Approaches. *Sustainability*, *14*(6). https://doi.org/10.3390/SU14063353

Waseem, A., Arshad, J., Iqbal, F., Sajjad, A., Mehmood, Z., & Murtaza, G. (2014). Pollution Status of Pakistan: A Retrospective Review on Heavy Metal Contamination of Water, Soil, and Vegetables. *BioMed Research International*. https://doi.org/10.1155/2014/813206

Wei, M., Zhang, J. J., Liu, H., Wang, S. J., Fu, H., & Zhou, N. Y. (2009). A transposable class I composite transposon carrying *mph* (methyl parathion hydrolase) from *Pseudomonas sp*. strain WBC-3. *FEMS Microbiology Letters*, *292*(1), 85–91. https://doi.org/10.1111/J.1574-6968.2008.01468.X

Wery, J., Hidayat, B., Kieboom, J., & De Bont, J. A. M. (2001). An Insertion Sequence Prepares *Pseudomonas putida* S12 for Severe Solvent Stress. *Journal of Biological Chemistry*, *276*(8), 5700–5706. https://doi.org/10.1074/JBC.M007687200

Wielgoss, S., Barrick, J. E., Tenaillon, O., Wiser, M. J., Dittmar, W. J., Cruveiller, S., Chane-Woon-Ming, B., Medigue, C., Lenski, R. E., & Schneider, D. (2013). Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proceedings of the National Academy of Sciences of*

the United States of America, *110*(1), 222–227. https://doi.org/10.1073/PNAS.1219574110/SUPPL_FILE/PNAS.201219574SI.PDF

Xue, P., Minasny, B., McBratney, A., Wilson, N. L., Tang, Y., & Luo, Y. (2023). Distinctive role of soil type and land use in driving bacterial communities and carbon cycling functions down soil profiles. *Catena*, *223*. https://doi.org/10.1016/J.CATENA.2022.106903

Yan, B., Li, J., Xiao, N., Qi, Y., Fu, G., Liu, G., & Qiao, M. (2016). Urban-development-induced Changes in the Diversity and Composition of the Soil Bacterial Community in Beijing. *Scientific Reports*, *6*(1). https://doi.org/10.1038/srep38811

Yang, Z. N., Liu, Z. S., Wang, K. H., Liang, Z. L., Abdugheni, R., Huang, Y., Wang, R. H., Ma, H. L., Wang, X. K., Yang, M. L., Zhang, B. G., Li, D. F., Jiang, C. Y., Corvini, P. F. X., & Liu, S. J. (2022). Soil microbiomes divergently respond to heavy metals and polycyclic aromatic hydrocarbons in contaminated industrial sites. *Environmental Science and Ecotechnology*, *10*. https://doi.org/10.1016/J.ESE.2022.100169

Yasir, M., Turner, A. K., Lott, M., Rudder, S., Baker, D., Bastkowski, S., Page, A. J., Webber, M. A., & Charles, I. G. (2022). Long-read sequencing for identification of insertion sites in large transposon mutant libraries. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-07557-x

Younas, M., Shahzad, F., Afzal, S., Khan, M. I., & Ali, K. (1998). Assessment of Cd, Ni, Cu, and Pb pollution in Lahore, Pakistan. *Environment International*, *24*(7), 761–766. https://doi.org/10.1016/S0160-4120(98)00068-3

Yu, X., Zhao, J. T., Liu, X., Sun, L. X., Tian, J., & Wu, N. (2021). Cadmium Pollution Impact on the Bacterial Community Structure of Arable Soil and the Isolation of the Cadmium Resistant Bacteria. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/FMICB.2021.698834/FULL

Zhang, Lei, Shen, T., Cheng, Y., Zhao, T., Li, L., & Qi, P. (2020). Temporal and spatial variations in the bacterial community composition in Lake Bosten, a large, brackish lake in China. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-019-57238-5

Zhang, Lu, Chen, F. X., Zeng, Z., Xu, M., Sun, F., Yang, L., Bi, X., Lin, Y., Gao, Y. J., Hao, H. X., Yi, W., Li, M., & Xie, Y. (2021). Advances in Metagenomics and Its Application in Environmental Microorganisms. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/FMICB.2021.766364/BIBTEX

Zhang, X. J., Zhang, J., Yao, Q., Feng, G. Da, & Zhu, H. H. (2019). *Microvirga flavescens sp. nov.,* a novel bacterium isolated from forest soil and emended description of the genus *Microvirga*. *International Journal of Systematic and Evolutionary Microbiology*, *69*(3), 667–671. https://doi.org/10.1099/IJSEM.0.003189/CITE/REFWORKS

Zhao, Y., Luo, M., Zhou, Y., Jia, X., Kang, S., Yang, S., & Mu, Q. (2022). Spatial patterns of dominant bacterial community components and their influential factors in the southern Qinling Mountains, China. *Frontiers in Microbiology*, *13*. https://doi.org/10.3389/FMICB.2022.1024236/FULL

Appendices

Table A.1. Overview of the Pakistani samples with their environmental characteristics.

sample-id	Cities	pH	EC (uS)	Total_carbon (%)	Total_nitrogen (%)	DiPCB (ng/g)	TriPCB (ng/g)	TetraPCB (ng/g)	PentaPCB (ng/g)	HexaPCB (ng/g)	HeptaPCB (ng/g)	Cr (mg/kg)	Cu (mg/kg)	Pb (mg/kg)	Cd (mg/kg)	As (mg/kg)	TPH (mg/kg)
LA34	Kharian	7.34	306	1.65	0.07	5.93	3.95	26.97	114.5	15.43	18.39	25.47	12.86	14.66	0.11	2.55	3030
CA58	Chaklala	7.09	249	2.22	0.12	0.63	10.25	4.33	3.07	2.84	1.87	24.08	11.99	7.56	0.08	3.46	5440
CI59	Chaklala	7.15	254	3.84	0.14	1.45	8.65	1.83	3.27	1.99	1.85	42.53	104.4	31.47	0.72	2.64	5970
CR60	Chaklala	6.93	260	2.24	0.06	1.99	1.64	2.09	4.85	1.21	2.22	26.98	12.23	10.37	0.17	3.51	6390
DA43	Dina	7.05	293	0.22	0.02	3.57	7.12	26.95	64.76	191.9	95.46	32.78	14.99	9.78	0.07	3.72	6830
DI44	Dina	6.99	254	1.24	0.06	4.91	3.3	26.02	143.3	190.3	44.23	26.02	14.82	8.09	0.1	3.86	10500
DR45	Dina	7.1	302	2.38	0.05	1.63	11.34	42.71	102.52	193.23	98.57	17.69	14	31.5	0.13	1.48	11240
GA22	Gakhar	7.64	290	1.07	0.09	1.78	0.77	4.16	10.13	3.41	23.25	45.76	25.33	15.43	0.15	4.42	6070
GI23	Gakhar	7.57	311	2.41	0.12	1.77	5.91	24.38	182.65	15.52	62.02	25.04	27.94	10.15	0.16	5.15	2130
GKA52	Mandra	7.21	289	2.94	0.14	0.8	3.51	18.56	94.63	338.18	49.24	41.28	24.04	18.14	0.16	6.29	3450
GKI53	Mandra	7	318	2.17	0.14	1.24	3.97	14.96	89.33	563.68	42.09	20.53	31.39	11.51	0.21	2.18	6040
GKR54	Mandra	6.78	335	1.73	0.05	0.36	6.27	22.59	117.68	349.25	86.98	21.44	16.52	17.25	0.12	2.8	9360
GR24	Gakhar	7.43	277	1.45	0.07	2.83	4.92	18.97	111.89	7.771	44.72	43.41	20.14	15.15	0.15	3.12	9530
GTA28	Gujrat	6.61	286	1.94	0.05	3.7	2.24	14.32	192.34	51.71	20.46	32.96	93.98	10.2	0.11	5.43	15560
GTI29	Gujrat	6	234	4.76	0.07	0.22	1.02	2.12	9.35	21.29	43.2	23.92	1447.43	16.79	0.15	2.42	16160
GTR30	Gujrat	6.38	272	5.13	0.06	0.4	1.74	9.26	20.54	16.49	50.27	18.74	131.51	26.53	0.15	1.68	11070
GWA19	Gujranwala	7.03	380	1.55	0.14	98.18	780.64	1759.43	1871.15	816.24	161.35	34.24	21.72	34.16	0.25	4.18	9520
GWI20	Gujranwala	7.8	402	1.05	0.06	611.15	4999.29	16717.33	14908.44	9318.44	1903.23	20.12	24.06	27.04	0.27	4.45	14500
GWR21	Gujranwala	7.42	371	1.65	0.05	36.01	339.07	1052.38	1325.52	771.92	136.4	26.09	23.18	1226.02	0.5	6.79	14810
IA64	Islamabad	7.72	294	2.96	0.22	8.83	30.05	12.65	23.83	6.52	1.68	46.9	21.67	15.03	0.29	6.68	8940
II65	Islamabad	7.8	314	3.34	0.24	7.4	12.7	7.04	11.39	1.4	0.67	51.13	42.77	67.12	0.38	3.13	10840
IR66	Islamabad	7.91	276	3.21	0.25	6.33	43.16	20.31	22.19	0.2	0.44	48.81	22.38	16.51	0.4	7.99	8610
JA40	Jehlum	7.01	315	2.2	0.09	3.9	1.71	36.97	170.01	45.97	18.19	41.35	17.59	11.54	0.13	4.53	8700
JI41	Jehlum	7.06	346	1.93	0.12	4.58	4.12	31.14	192.78	66.3	19.08	25.64	30.93	16.79	0.26	1.92	7800
JR42	Jehlum	7.12	338	1.78	0.19	8.09	5.77	40.93	164.5	6.33	17.84	85.5	99.11	39.98	0.29	4.26	11370
KA16	Kamoki	7.5	367	1.14	0.08	1.13	7.06	16.15	13.04	15.87	6.41	3112.23	113.99	28.97	0.67	4.66	11500
KA37	Saraie	7.63	324	1.18	0.06	2.25	3.34	32.67	106.8	22.28	18.9	29.57	12.56	12.96	0.1	3.55	8190

	alamgir																
KI17	Kamoki	7.28	340	3.62	0.13	0.78	4.32	19.46	60.12	68.88	29.35	21.25	164.06	35.53	0.27	2.98	10570
KI38	Sarae alamgir	7.21	305	3.4	0.13	7.32	10.68	36	205.77	37.68	16.2	31.46	184.96	229.47	0.25	2.69	7720
KKA10	Kala Shah Kaku	7.34	313	0.96	0.08	1.32	10.18	28.01	20.22	12.89	1.14	331.05	28.78	17.19	0.21	11.41	6480
KKI11	Kala Shah Kaku	7.17	324	4.07	0.05	5.17	70.45	168.69	98.88	55.71	11.58	400.54	24.39	23.79	0.12	2.81	7270
KKR12	Kala Shah Kaku	6.82	312	1.1	0.06	65.79	1057.11	1661.68	210	57.55	20.11	16.24	54.57	46.58	1.45	4.08	11050
KR18	Kamoki	7.46	336	0.81	0.04	0.1	0.27	3.95	22.5	32.91	5.46	28.1	37.09	18.93	0.15	5	13450
KR39	Sarae alamgir	7.01	299	1.83	0.05	0.78	3.1	23.35	250.61	42.68	23.2	39.14	15.87	14.27	0.11	1.89	11650
LCA1	Lahore Cantt	7.18	297	3.47	0.34	17.83	100.33	387.86	261.65	151.96	21.19	46.05	25.88	15.73	0.28	7.48	7790
LCIA2	Lahore Cantt	7.03	327	0.97	0.07	3.91	24.97	58.55	75.02	42.32	10.25	31.59	25.48	16.77	0.25	8.65	12550
LCR3	Lahore Cantt	7.11	347	8.07	0.27	36.01	339.07	1052.38	1325.52	771.92	136.4	27.24	395.2	61.96	0.63	5.01	10690
LI35	Kharian	7.77	322	4.03	0.29	1.01	3.72	16.72	110.64	48.22	17.97	489.82	31.91	19.13	0.25	5.08	7140
LR36	Kharian	7.6	334	1.06	0.04	4.29	3.11	35.28	124.71	22.46	17.44	24.54	10.56	8.62	0.08	2.34	10200
LYA4	Lahore City	7.34	301	1.54	0.11	0.2	4.51	7.76	4.15	7.92	2.87	31.58	28.41	32.04	0.26	11.12	10900
LYI5	Lahore City	6.98	325	2.4	0.11	13.76	961.21	2722.45	2639.38	1555.71	127.3	193	117.95	530.82	0.45	5.03	15820
LYR6	Lahore City	6.76	316	2.03	0.08	11.67	104.96	387.26	275.24	192.17	81.26	414.58	50.95	32.08	0.31	4.89	15300
MA13	Muridke	7.21	285	5.03	0.24	0.2	4.51	7.76	4.15	7.92	2.87	39.15	21.43	23.18	0.17	10.66	9030
MDA49	Gujar Khan	7.03	321	1.61	0.03	2.15	2.03	28.76	247.79	135.81	44.97	41.45	16.34	8.77	0.11	5.55	2810
MDI50	Gujar Khan	6.56	349	3.01	0.18	1.37	3.03	13.12	104.17	189.1	19.76	41.23	28.47	18.46	0.23	7.77	4090
MDR51	Gujar Khan	7.14	311	2.88	0.12	0.46	2.33	29.23	166.31	176.24	67.44	29.6	23.54	12.81	0.34	7.2	8890
MI14	Muridke	6.44	299	4.63	0.13	1.18	8.72	17.61	24.3	45.98	25.94	75.97	2161.91	138.71	0.96	9.49	11680
MR15	Muridke	7.04	321	2.13	0.16	2.16	11.78	36.55	113.83	208.89	72.38	33.56	143.87	18.73	0.15	10.17	10760

RA55	Rawat	7.15	314	2.22	0.17	1.25	3.52	19.14	49.05	465.8	37.57	45.55	26.35	15.65	0.17	7.34	4360
RI56	Rawat	7.23	353	4.09	0.23	3.91	7.42	27.58	104.17	528.98	70.06	37.7	21.77	21.32	0.24	6.46	10190
RPA61	Rawalpindi	6.84	341	1.22	0.11	24.5	9.95	18.23	37.71	1.23	1.43	47.83	21.85	13.25	0.16	7.72	8270
RPI62	Rawalpindi	6.79	307	3.52	0.26	107.29	39.56	1041.02	408.8	1.06	4.29	29.21	20.94	10.98	0.19	3.54	27770
RPR63	Rawalpindi	6.66	352	0.48	0.05	10.96	116.8	775.91	321.58	18.21	8.17	46.09	21.02	12.91	0.11	7.6	31590
RR57	Rawat	7.16	304	5.44	0.26	2.38	3.96	35.73	62.14	412.07	24.09	198.55	27.02	18.9	0.37	3.96	11500
SA46	Sohawa	7.04	323	2.31	0.1	5.5	3.04	54.12	80.36	196.49	91.74	20.89	10.8	7.45	0.08	3.47	38100
SAA31	Lalamusa	7.73	344	0.84	0.05	0.4	1.43	3.22	13.84	58.65	51.04	31.1	16.02	10.26	0.08	5.79	72220
SAI32	Lalamusa	7.34	312	2.09	0.1	0.83	3.17	8.07	8.93	22.7	25.03	32.97	46.27	30.28	0.45	2.35	90900
SAR33	Lalamusa	7.4	297	1.75	0.06	0.7	0	22.18	5.28	16.94	18.83	444.96	30.04	26.04	0.28	2.31	13410
SDA7	Shahdara	7.39	352	5.21	0.35	137.12	2219.58	6081.02	3117.21	1163.82	216.82	29.68	23.59	24.9	0.25	6.78	11080
SDI8	Shahdara	7.09	310	1.44	0.1	223.42	1131.15	4789.92	16701	7736.8	908.29	42.17	112.27	146.99	1.5	6.43	14180
SDR9	Shahdara	7.3	302	1.65	0.14	58	379.84	1387.31	1660.17	708.59	121.47	21.56	32.74	41.03	0.39	5.81	12430
SI47	Sohawa	6.82	358	3.48	0.14	0.99	3.17	33.93	70.66	116.51	87.91	64.36	17.98	198.95	0.62	4.72	5680
SR48	Sohawa	6.91	310	2.26	0.04	0.62	3.1	10.48	149.43	184.84	51.41	24.55	9.73	9.6	0.15	2.25	10050
WA25	Wazirabad	7.06	340	1.51	0.08	2.43	2.87	24.59	57	15.6	91.24	459.37	29.45	14.65	0.15	4.95	10290
WI26	Wazirabad	6.87	328	1.96	0.1	1.81	3.25	18.73	197.23	53.02	10.85	30.61	31.36	16.05	0.17	4.18	7950
WR27	Wazirabad	6.71	303	1.84	0.05	5.51	4.51	29.73	122.39	12.13	47.09	20.01	25.31	11.43	0.12	2.23	13110

Table A.2. Overview of the Pakistani samples with their land use and spatial factor.

Sample-id	Cities	Land use	Latitude	Longitude
LA34	Kharian	Agriculture	0.249749	0.212649
CA58	Chaklala	Agriculture	1.325178	-1.54088
CI59	Chaklala	Industrial	1.297894	-1.58159
CR60	Chaklala	Roadside	1.280071	-1.54729
DA43	Dina	Agriculture	0.543412	-0.43988
DI44	Dina	Industrial	0.539903	-0.43519
DR45	Dina	Roadside	0.538162	-0.4284
GA22	Gakhar	Agriculture	-0.48365	0.776716
GI23	Gakhar	Industrial	-0.40432	0.756432

GKA52	Mandra	Agriculture	0.983409	-1.2292
GKI53	Mandra	Industrial	0.982741	-1.2184
GKR54	Mandra	Roadside	0.987658	-1.21827
GR24	Gakhar	Roadside	-0.45271	0.770398
GTA28	Gujrat	Agriculture	0.084987	0.733224
GTI29	Gujrat	Industrial	-0.10918	0.638397
GTR30	Gujrat	Roadside	-0.18878	0.64073
GWA19	Gujranwala	Agriculture	-0.51829	0.691812
GWI20	Gujranwala	Industrial	-0.59008	0.824074
GWR21	Gujranwala	Roadside	-0.57426	0.824401
IA64	Islamabad	Agriculture	1.373948	-1.21697
II65	Islamabad	Industrial	1.383749	-1.60985
IR66	Islamabad	Roadside	1.445152	-1.55096
JA40	Jehlum	Agriculture	0.399638	-0.18862
JI41	Jehlum	Industrial	0.485388	-0.26119
JR42	Jehlum	Roadside	0.441993	-0.19949
KA16	Kamoki	Agriculture	-0.89636	0.986889
KA37	Sara e alamgir	Agriculture	0.362921	-0.0764
KI17	Kamoki	Industrial	-0.89844	0.878016
KI38	Sara e alamgir	Industrial	0.358423	-0.08301
KKA10	Kala Shah Kaku	Agriculture	-1.24548	1.017805
KKI11	Kala Shah Kaku	Industrial	-1.23541	1.028856
KKR12	Kala Shah Kaku	Roadside	-1.25345	1.029987
KR18	Kamoki	Roadside	-0.90426	0.930807
KR39	Sara e alamgir	Roadside	0.352663	-0.06135
LCA1	Lahore Cantt	Agriculture	-1.60637	1.562054
LCIA2	Lahore Cantt	Industrial	-1.73486	1.24411
LCR3	Lahore Cantt	Roadside	-1.59122	1.34401
LI35	Kharian	Industrial	0.233867	0.17316
LR36	Kharian	Roadside	0.238514	0.158157
LYA4	Lahore City	Agriculture	-1.93599	0.912697
LYI5	Lahore City	Industrial	-1.87412	0.826035
LYR6	Lahore City	Roadside	-1.42092	1.091819
MA13	Muridke	Agriculture	-1.15989	0.984974
MDA49	Gujar Khan	Agriculture	0.844978	-1.08306
MDI50	Gujar Khan	Industrial	0.875741	-1.10608
MDR51	Gujar Khan	Roadside	0.844066	-1.07679
MI14	Muridke	Industrial	-1.12984	0.998268

MR15	Muridke	Roadside	-1.14065	0.999704
RA55	Rawat	Agriculture	1.156671	-1.31601
RI56	Rawat	Industrial	1.194376	-1.2119
RPA61	Rawalpindi	Agriculture	1.382822	-1.56696
RPI62	Rawalpindi	Industrial	1.248217	-1.60724
RPR63	Rawalpindi	Roadside	1.350762	-1.67399
RR57	Rawat	Roadside	1.180995	-1.34376
SA46	Sohawa	Agriculture	0.678194	-0.84333
SAA31	Lalamusa	Agriculture	0.099314	0.386062
SAI32	Lalamusa	Industrial	0.099164	0.343447
SAR33	Lalamusa	Roadside	0.098578	0.34253
SDA7	Shahdara	Agriculture	-1.39643	1.035208
SDI8	Shahdara	Industrial	-1.4039	1.098929
SDR9	Shahdara	Roadside	-1.3843	1.072552
SI47	Sohawa	Industrial	0.678934	-0.83919
SR48	Sohawa	Roadside	0.677938	-0.82933
WA25	Wazirabad	Agriculture	-0.25636	0.664758
WI26	Wazirabad	Industrial	-0.26509	0.714939
WR27	Wazirabad	Roadside	-0.24558	0.689981

Use of Generative Artificial Intelligence (GenAI)

Student name: Kristina Yefimak
Student number: r0775935

Please indicate with "X" whether it relates to a course assignment, to the BIG-project or to the master's thesis:

X This form is related to my master's thesis.
Title master's thesis: Deciphering the role of IS*6100* in microbial community adaptation to anthropogenic environmental stressors
Promoter: Prof. Dirk Springael

O This form is related to a BIG-project.
Title BIG-project: …
Promoter: …

O This form is related to a course assignment.
Course name: …
Course code: …

Please indicate with "X":

- o I did not use GenAI tools.
- X I did use GenAI tools. In this case specify which one (e.g. ChatGPT/GPT4/...): ChatgGPT/ QuillBot

Please indicate with "X" (possibly multiple times) in which way you were using it:

- X As a language assistant for reviewing or improving texts you wrote yourself, provided that the model does not add new content. In this case, the use of GenAI is similar to the spelling and grammar check tools we already have today, so you do not need to explicitly mention using GenAI for this).
- X As a search engine to get initial information on a topic or to make an initial search for existing research on the topic. (This way of gathering information is similar to using an ordinary search engine when working on anassignment. As a student, you are responsible for checking and verifying the absence and correctness of references. Therefore, after this initial search, look for scientific sources and conduct your own analysis of the source documents. Interpret, analyse and process the information you obtained; don't just copy-paste it. If you then write your own text based on this information, you do not have to mention you used GenAI.)
- o To generate text blocks. (If you do copy-paste text blocks of GenAI output, you have to cite your GenAI sources and quote them, i.e. you clearly state that the item was created via GenAI by citation/reference.)
- o To generate graphs or Fig.s. (If you do copy-paste graphs/Fig.s of GenAI output, you have to cite the GenAI sources and quote them, i.e. you clearly state that the item was created via GenAI by citation/reference.)
- o To generate some code as part of a larger assignment. (Watch out, this can only be done if the teacher/promotor explicitly allows it.)

- Other (Contact the teacher of the course or the supervisor of the thesis or BIG project. Explain how you comply with article 84 of the examination regulations. Explain the usefulness or added value of using GenAI.)

Further important guidelines and remarks:

The faculty follows the KU Leuven policy regarding responsible use of GenAI. This form is an aid towards transparency about the use of GenAI by the student which is essential. Irresponsible and non-transparent use of GenAI can be considered an irregularity and can be sanctioned. Students who consider to use GenAI should inform themselves through the university website concerning the additional guidelines (How to correctly quote and refer to GenAI? What is (not) allowed? Tips and points of attention for responsible use):
https://www.kuleuven.be/english/education/student/educational-tools/generative-artificial-intelligence

Popularized summary

In our modern world pollution poses a challenge as cities expand, industries thrive, and agricultural activities intensify, leading to the increasing presence of xenobiotics, substances foreign to natural ecosystems. As these xenobiotics alter soil characteristics and soil bacterial communities, bacteria have evolved to adapt to changes in their environment to cope with these stressors. Among these mechanisms, insertion sequences (IS) can pick up and carry genes by forming composite transposons, including genes involved in degradation of xenobiotics. Despite their importance, research on the amount and functional role of IS in bacterial communities within their natural environment remains limited. This study aims to understand the role of IS*6100* in genetic adaptation within soil bacterial communities. The first part of the thesis was conducted on samples from diverse regions of Pakistan to see how factors like type of land, where the soil is taken from and present pollutants, affected bacterial community composition and the distribution of IS*6100*. The study revealed that while soil factors such as copper, total nitrogen, arsenic, cadmium, and total carbon contributed to explain variability in bacterial composition, the type of land has a bigger impact. Different types of land and sampling location had a significant effect on the variety of soil bacteria, with agricultural soil showing the highest bacterial diversity. However, 86% of the bacterial community variation remained unexplained, suggesting that there are other factors, that were not measured, driving the variation. IS*6100* was present in all Pakistani soil samples, although there was no difference in abundance of IS*6100* based on the type of land and the sampling location. Interestingly, a negative correlation existed between the amount of IS*6100* and toxic contaminants such as DiPCB and TriPCB. The second part focused on Belgian samples, where soil samples were contaminated with different pollutants and the amount of IS*6100* and cargo genes were determined. IS*6100* was also present in all samples, although the amount differed depending on the type of pollution. Soil samples that had been regularly exposed to pesticides showed higher levels of IS*6100* compared to those with less obvious pollutants. By employing advanced molecular techniques including long-range PCR, and nanopore sequencing, a first step was taken in identifying the cargo genes carried by IS*6100*. It was confirmed that the genes were amplified successfully.