Widespread occurrence of bacterial human virulence determinants in soil and freshwater environments

Ditte A. Søborga, Niels Bohse Hendriksenb, Mogens Kilianb, and Niels Kroera#

a Department of Environmental Science, Aarhus University, Frederiksborgvej 399, DK-4000, Roskilde, Denmark

b Department of Biomedicine, Aarhus University, Wilhelm Meyers Allé 4, DK-8000, Aarhus C, Denmark

Running title: Virulence determinants in soil and freshwater environments

#Corresponding author. Mailing address: Department of Environmental Science, Aarhus University, Frederiksborgvej 399, DK-4000 Roskilde, Denmark; E-mail nk@dmu.dk; Phone +45 87 15 87 01
ABSTRACT

The occurrence of 22 bacterial human virulence genes (encoding toxins, adhesins, secretion systems, regulators of virulence, inflammatory mediators, and bacterial resistance) was investigated by nested PCR in beech wood soil, roadside soil, organic agricultural soil, and freshwater biofilm. The presence of clinically relevant bacterial groups known to posses the virulence genes was tested by PCR of 16S and 23S rRNA genes. For each of the virulence genes detected in the environments, sequencing and NCBI BLAST confirmed the identity of the PCR products. The virulence genes showed widespread environmental occurrence as 17 different genes were observed. Sixteen genes were detected in the beech wood soil, 14 in the roadside and organic agricultural soils, while 11 were detected in the freshwater biofilm. All types of virulence traits were represented in all environments, however, the frequency by which they were detected was variable. A principal component analysis suggested that several factors related to the level of contamination by polycyclic aromatic hydrocarbons and pH. The occurrence of the virulence genes in the environments did generally not appear to be the result of the presence of clinically relevant bacteria, indicating an environmental origin of the virulence genes. The widespread occurrence of the virulences traits, and a high degree of sequence conservation between the environmental and clinical sequences, suggest that soil and freshwater environments may constitute reservoirs of virulence determinants normally associated with human disease.
INTRODUCTION

Bacterial pathogens continue to cause major threats to human health and welfare due to the emergence of new pathogens, re-emergence of previously well-controlled infectious diseases, and increasing prevalence of antimicrobial resistant bacteria (1-5). However, the origin of many emerging human pathogens and their virulence genes remains to be elucidated (2,6).

Traditionally, it has been believed that the complex interactions between pathogenic bacteria and their hosts are the primary driving forces for pathogens to develop mechanisms to overcome host defense strategies (4). For this reason, most of our knowledge regarding virulence factors and bacterial pathogenicity derives from studies of infections (6-8).

However, the origin of virulence genes may be environmental bacteria (8), and the acquisition of virulence genes through horizontal gene transfer of mobile genetic elements in outer non-host (non-human) environments may play an important role for the evolution of bacterial pathogens (4,9). Hence, studies on the occurrence and expression of virulence genes in non-host environments are important to obtain better knowledge of the origin and evolution of bacterial pathogens (6,8).

Virulence factors are likely to have similar physiological functions in the environment as in the host organism. This could be the case for e.g. iron acquisition systems or for structural elements such as capsules or polysaccharides for which the role in virulence is a consequence of the host’s mechanisms for detecting and eliminating pathogens (8,10). Other virulence factors may have different functions in e.g. bacterial infection and in fitness to specific environmental habitats (8,11). This dual role has been seen for an adhesion factor of Vibrio cholera which mediates attachment to human epithelial cells required for intestinal colonization as well as attachment to zooplankton enhancing its environmental survival (12).

Furthermore, it has been shown for E. coli, Pseudomonas aeruginosa, Legionella pneumophila,
and *Listeria monocytogenes* that virulence factors required for infecting humans are also
needed for infecting plants, invertebrates or insects (13-18).

If survival and adaptation of a bacterium in humans is similar to the survival and
adaptation of bacteria in the environment, then genes that typically are associated with
pathogenicity will be present in non-pathogens from the environment. For instance, Casas *et
al.* (19) have demonstrated that phage-encoded exotoxins are widespread in water, sediment
and soil. Additionally, sequence analysis revealed a high number of protein secretion systems
and virulence gene homologues in marine bacteria (6). These findings support the theory that
virulence genes are of environmental origin and that these traits may be used by naturally
occurring bacteria to survive and grow in the environment (20).

Here we report the widespread occurrence of several classes of human virulence
determinants, including toxins, adhesins, secretion systems, inflammatory mediators,
regulators of virulence, and bacterial resistance factors, in three different soils and a
freshwater biofilm. The study focuses on the environmental gene pool of pathogenic traits
rather than on the occurrence of specific pathogens in the environment. Previous studies have
either focused on the detection of clinically relevant bacteria in non-host environments (e.g.
(21,22) or on the environmental survival of known pathogens (e.g. (23). The widespread
occurrence of virulence genes in soil and freshwater suggests that these environments may
constitute reservoirs of virulence determinants normally associated with human disease.

**MATERIALS AND METHODS**

**Sampling.** Triplicate surface soil samples (0-10 cm depth) of approximately 1 kg were
collected on November 16th 2009 at three different localities: 1) an organic agricultural field
(Snubbekorsgård, Taastrup, Denmark), 2) a beech wood (Boserup Skov, Roskilde, Denmark),
and 3) an urban roadside (Jarmers Plads, Copenhagen, Denmark). Sampling was done using a sterile shovel and samples were transferred to sterile plastic bags. Soil samples were passed through a sterile 4 mm mesh sieve, homogenized by careful mixing with a sterile spoon, and stored at 4°C until use. In addition to the soil samples, triplicate ~1 kg samples of pebbles (diameter 2-4 cm) and 1 L of water were collected on December 1st 2009 from a lake (Lake Lyngby, Lyngby, Denmark) at 50 cm depth using sterile gloves, plastic bags and glass bottles. To dislodge biofilm, pebbles and water were sonicated for 30 min at room temperature using a Branson 5210 ultrasonic bath (Branson, Germany). After removal of the pebbles, the water was centrifuged at 3 000 g for 15 min at 20°C and the pelleted biofilm stored at 0°C until further analysis. Sterile gloves were used during all handling of the samples.

**Chemical and biological analyses.** Soils were analyzed by Eurofins-Steins A/S (Odense, Denmark) for pH, texture (grain sizes) and content of inorganic nutrients (phosphorus, total nitrogen). Determination of water holding capacity and loss of ignition (carbon content) of the soils were measured by standard procedures: 105°C overnight and 550°C for 2 h, respectively (24) while the concentration of polycyclic aromatic hydrocarbons (PAHs) was measured according to Hollender et al. (25). pH and concentration of dissolved organic carbon (26) were determined in the water samples.

Total numbers of bacteria in the soils were determined by acridine orange (AO) direct counting (27). Briefly, a 1:10 000 dilution of 1 g soil in 9 ml sterile Milli-Q water was filtered onto 0.2 µm polycarbonate membranes (25 mm in diameter, Nucleopore®) and at least 400 bacteria per sample were counted.

**Detection of virulence determinants.** DNA was extracted using the PowerSoil™ DNA Isolation Kit (MoBio Laboratories Inc., USA) according to the manufacturer’s instructions. Soil DNA was extracted from triplicate 250 mg subsamples from each of the triplicate soil samples.
per sampling locality, while biofilm DNA was extracted from triplicate 100 mg subsamples of
the triplicate biofilm samples. Thus, nine DNA extractions were prepared from each
environment. DNA was stored at -20°C until use. DNA concentration and purity were
evaluated spectrophotometrically by UV absorption at 260 nm (A260) and A260/A280 ratio,
respectively, using a NanoDrop™ 2000C Spectrophotometer (Thermo Scientific, USA).

Detection of 22 different virulence genes (Table 1) was carried out by PCR. Primer sets
(Supplemental Table S1) were derived directly from the literature or in degenerated versions,
or they were designed de novo from available nucleotide sequences using Primer3 Version
0.4.0 (28). The specificity of all primer sets was verified by NCBI BLAST (29,30).

PCR was performed in 25 µl reaction mixtures containing 1.25 U AmpliTaq Gold® DNA
polymerase, 1 × PCR Gold buffer, 3.0 mM MgCl₂ (all from Applied Biosystems, USA), 200 µM
dNTP (Thermo Scientific, USA), 0.4 µM of each primer and 1 µl DNA template (1-5 ng). The
PCR was performed with a Hybaid PCR Express Thermal Cycling (AH Diagnostics, Denmark)
and consisted of 1 cycle of denaturation at 95°C for 5 min, 35 cycles of denaturation at 95°C
for 45 s, annealing at the appropriate temperature for each primer set for 60 s, and extension
at 72°C for 90 s followed by a final extension at 72°C for 5 min after which the PCR products
were maintained at 4°C. Nested or semi-nested PCR was performed in a similar manner with
0.5 µl product of the first PCR as template. Two PCRs were run for each sample (i.e. 18 PCRs
per environment per primer set). PCR products were analyzed by 1.5% or 2.5% (nested PCR)
agarose gel electrophoresis. The presence of a virulence gene (i.e. an internal amplicon with
homology to a fragment of a targeted virulence gene) was only scored as positive if it was
documented in at least two replicate PCRs. For both conventional and nested PCR, 4-10
negative controls (no added DNA) were included for every 72 sample PCRs. In total this
summed up to more than 200 negative controls. In no case did any of these negative controls
give rise to false positives.

Bacteria known to possess the respective virulence genes (Table 1) were used as positive
controls. Generally, bacteria were cultured on Luria-Bertani agar and incubated overnight at
30°C. For *H. pylori* NCTC 11638, Columbia agar supplemented with 4% horse blood was used.
During the incubation period, a liquid film of brain heart infusion was maintained on the agar
surface. Incubation was performed under microaerobic conditions (10% [vol/vol] CO₂, 85%
[vol/vol] N₂, and 5% [vol/vol] O₂) for 7 days. DNA was extracted by boiling (31). *E. coli* O157
and *B. anthracis* DNA was donated by Statens Serum Institute (Denmark).

**Sequencing of PCR products.** To confirm the identity of the PCR products, one PCR
product of each of the detected virulence gene fragments was sequenced. Products was run on
a 2% agarose gel, the specific bands cut out under UV light, and DNA extracted using the
QIAEX II Gel Extraction Kit (QIAGEN, Germany). Extracted DNA was amplified by PCR in 10 μl
reaction mixtures containing 1 μl BigDye® Terminator v1.1 Ready Reaction Mix, 0.75 X
Sequencing Buffer (Applied Biosystems, USA), forward or reverse primers at a concentration
of 2.5 μM, and 1 μl of the extracted DNA. The PCR conditions were 1 cycle of denaturation at
96°C for 5 min, 25 cycles of denaturation at 96°C for 30 s, annealing at 59°C for 15 s, and
extension at 60°C for 4 min after which the PCR products were maintained at 4°C. DNA was
precipitated with 75% isopropanol, and following centrifugation, the pellet was dried for 5
min in a vacuum oven and stored at -20°C. DNA was resuspended in Hi-Di™ Formamide
(Applied Biosystems, USA) and sequenced in one direction by Macrogen Inc. (Seoul, Korea)
using an ABI 3130xl Genetic Analyser (Applied Biosystems, USA).

**Assessment of detection limit.** To determine the detection limit of the nested PCR
procedure, bacteria known to possess the respective virulence associated genes (Table 1)
were added at densities of approximately 0, 10, 100, 1 000, and 10 000 cells per 250 mg soil or 100 mg biofilm prior to DNA extraction. *E. coli* O157 and *B. anthracis* were not added as live cells but as DNA. DNA extraction and PCR conditions were as described above.

**Detection of clinically relevant bacteria in environmental samples.** To assess whether the origin of the observed virulence genes was environmental or due to the presence of clinically relevant bacteria, bacterial groups from which the 22 virulence genes were originally identified (Salmonella spp., *E. coli*, Helicobacter spp., Agrobacterium tumefaciens, and *B. cereus* group bacteria; Supplemental Table S2), were targeted by nested PCR of 16S and 23S rRNA genes. PCR was performed as described for the detection of the virulence genes, except that annealing temperatures were adjusted appropriately and 40 cycles were run for *E. coli* and *Helicobacter* spp. in the conventional PCR. The nine DNA extractions from each environment, prepared for the detection of the virulence genes, were used as template DNA and duplicate PCR was run for each extract. Three negative controls containing sterile Milli-Q water instead of template DNA were set up for each of the primer sets.

**Phylogenetic analysis.** Alignments of positive PCR products were performed with the Clustal software and sequence variations determined on basis of pairwise distances using MEGA Version 5.01 (32). Sequences were manually edited and a few sequences of poor quality omitted. Phylogenetic associations of the PCR products of each of the genes *cytK2*, *hblA*, *invA*, *stx-1*, and *wbdI* were determined by including sequences from NCBI BLAST searches (29,30) in the phylogenetic analyses. The analyses were performed with MEGA Version 5.01 using the Minimum Evolution algorithm (complete deletion) (Fig. 3 and Supplemental Fig. S1, A-D).

**Statistical analyses.** A principal component analysis (PCA), based on the frequency of which the virulence genes were detected in the replicate environmental samples, was
performed using the software LatentiX Version 2.00 (Latent5 Aps, Denmark). One-way ANOVA was applied to explore differences along the axes. Further, one-way ANOVA was applied to determine differences between environmental characteristics of the soil environments.

**Accession numbers.** Sequences longer than 200 bp were deposited in GenBank under the following accession numbers: KC699845 to KC699851 for \textit{fimH} BWS, \textit{focG} BWS, \textit{wbdI} BIO 1A.3, \textit{wbdI} BWS 1B.2, \textit{wbdI} BWS 3C.3, \textit{wbdI} OAS 1C.2, and \textit{wbdI} RS 1A.3, respectively. All sequences, including those shorter than 200 bp, can be found in Supplemental Table S3.

**RESULTS**

**Physico-chemical characteristics of environments.** The physico-chemical characteristics of the individual soil samples differed in a number of respects (Table 2). Most significant was the difference in PAH content, with the roadside soil having a concentration > 25× higher than the beech wood soil and > 100× higher than the organic agricultural soil (P < 0.0001 in both cases). No significant differences in PAH content between the beech wood and organic agricultural soils were observed (P > 0.5). Other characteristics that differed between the soils were pH and phosphorus concentration. All other measured parameters were basically similar, including the bacterial density.

**Environmental distribution of virulence genes.** The virulence genes showed widespread environmental distribution as 77% of the investigated genes (17 out of 22) were observed in the soil and freshwater habitats (Table 3). The highest number of genes [16] was found in the beech wood soil, followed by the roadside and organic agricultural soils [14], and the freshwater biofilm [11]. All the detected genes were present in the beech wood soil except...
for the *H. pylori* vacuolating cytotoxin gene, *vacA*, which was only found in the organic
agricultural soil and the freshwater biofilm (Table 3).

Virulence genes of all groups of virulence determinants were found in at least one of the
investigated environments (Table 3). Among the group of toxins, five toxin genes were
observed. In the organic agricultural soil, all five genes were present, while four toxin genes
were detected in the other environments. The three undetected toxin genes were the *E. coli*
hemolysin A gene, *hlyA*, the *B. cereus* cytotoxin K1 gene, *crtK1*, and the *B. anthracis* edema
factor, *cva* (Table 3). All the adhesion group genes were found in the beech wood soil, while
the P fimbrial gene, *papC*, was found in all four environments. Two *Salmonella* genes involved
in the Type III secretion system, *invA* and *spiA*, were detected in all environments while a
third *Salmonella* secretion system gene complex, *sipB-sipC*, remained undetected. A fourth
gene, *virD4*, of *A. tumefaciens* was seen in two of the soil environments. Among the other
groups of virulence determinants (regulators of virulence, inflammatory mediators, and
bacterial resistance), the PhoQ sensor, *phoQ*, the *wbdi* gene of the *E. coli* O-antigen gene
cluster, the capsular biosynthesis protein, *capA*, and the polymyxin resistance gene, *arnT*,
were observed in all environments (Table 3).

**Relative frequency of virulence genes.** A semi-quantitative nested PCR approach (Fig. 1),
based on the number of positive to total number of PCRs, was applied to assess the relative
frequency of the investigated virulence determinants. As shown in Table 3, *hblA, invA, phoQ,
and wbdi* were relatively frequently observed in organic agricultural soil, beech wood soil,
roadside soil and freshwater biofilm, respectively, whereas e.g. *crtK2* and *spiA* were relatively
infrequently observed in all the environments. The detection limit was approximately 400
gene copies (cells) g⁻¹ soil (Fig. 1) and 1000 gene copies (cells) g⁻¹ biofilm (not shown).
Differences between the four environments with respect to occurrence and relative abundance of the virulence determinants were demonstrated by PCA (Fig. 2). Along the PC1 axis, the organic agricultural soil clustered separately from the roadside soil and the freshwater biofilm (One-way ANOVA, \( P < 0.05 \)). The distribution of the soil environments along the PC1 axis, from the lowest concentration of PAHs (organic agricultural soil, beech wood soil) to the highest concentration (roadside soil), suggested that PC1 explained variation arising from differences in PAH concentration. Along the PC2 axis, the beech wood soil was significantly different from all the other environments (One-way ANOVA, \( P < 0.05 \)). A possible explanation for the separation of beech wood soil from the other environments along this axis could be differences in pH, as the beech wood soil had a low pH (5.3) compared to the other environments (Table 2).

**Sequence homology and phylogenetic analysis.** The sequence homology between the 17 detected virulence gene fragments and the bacterial strains known to carry the corresponding target genes was 90 - 100%. Furthermore, NCBI BLAST analyses demonstrated that the environmental sequences were homologous to relevant sequences in the databases (highest E-score \( \leq 10^{-14} \) for best hit, \( stx-2 \)). Alignment of sequences of all PCR products for \( cytK2, hblA, invA, wbdI, \) and \( stx-1 \) showed a sequence variation of 2% for \( cytK2 \), 0 - 19% for \( hblA \), 0 - 10% for \( invA \), 0 - 1% for \( wbdI \), and 0% for \( stx1 \). Despite variations of up to 19%, as seen between two \( hblA \) sequences, all sequences of the five genes were homologous to the appropriate virulence associated genes (highest E-score \( \leq 10^{-16} \) for best hit, \( stx-1 \)).

Phylogenetic analysis of the partial \( cytK2, hblA, invA, wbdI, \) and \( stx-1 \) gene sequences confirmed their close relatedness to genes of clinically relevant bacteria (Fig. 3 and Supplemental Fig. S1 A-D). Thus for \( invA \), 28 sequences representing all four environments clustered closely together with \( invA \) of various serovars of *Salmonella enterica* (Fig. 3). Three
roadside soil sequences and one freshwater biofilm sequence, however, formed a separate cluster. Compared to invA, the phylogenetic distances between the environmental and database hblA sequences were relative large, and some of the environmental sequences formed their own cluster or sub-cluster (Supplemental Fig. S1, A). For wbdI, the majority of the environmental sequences fell within a cluster defined by several clinical E. coli serotype O111; however, two sequences formed separate clusters (Supplemental Fig. S1, B). The ten stx-1 and four cytK2 environmental sequences basically clustered together with the corresponding clinical sequences (Supplemental Fig. S1, C and D).

**Occurrence of clinical relevant bacteria.** The occurrence of virulence genes in the environments did generally not appear to be the result of the presence of clinically relevant bacteria. *Salmonella* spp. and *Helicobacter* spp. were not detected in any of the environments. Similarly, *E. coli/Shigella* were not detected in the beech wood and roadside soils, *B. cereus* group bacteria were not detected in the freshwater biofilm, and *A. tumefaciens* not detected in the roadside soil and the freshwater biofilm. *B. cereus* group bacteria, however, were relatively frequent in the beech wood soil (17 out of 18 replicate PCR tubes; 94%), the organic agricultural soil (94%), and the roadside soil (89%). Similarly, *E. coli/Shigella* were detected in 55% of the replicate nested PCRs of the agricultural soil and in 89% of the replicate nested PCRs of the freshwater biofilm, while *A. tumefaciens* was observed in the beech wood and agricultural soils (11% and 22%, respectively).

**DISCUSSION**

Numerous bacterial virulence genes were found in three different soil environments and a freshwater biofilm by use of a nested PCR approach, with DNA extracted directly from the natural environments. The nested PCR approach theoretically made it possible to detect as
few as one bacterial cell carrying a specific gene in the reaction tube. However, to minimize
the influence of PCR inhibitors such as humic substances (33), it was necessary to dilute the
soil and biofilm extracts, resulting in a detection limit of approximately 400 genes copies g⁻¹
soil (Fig. 1) or 1 000 gene copies g⁻¹ biofilm. This is comparable to the detection limit of 100 -
1 000 gene copies g⁻¹ soil found by Agersø et al. (34) who used nested PCR to study the
presence of the tetracycline resistant gene, \( \text{tet}(M) \), in farmland soil. Higher detection limits
have been reported in studies applying qPCR, e.g. \( 1.3 \times 10^5 \) \( \text{amoA} \) gene copies g⁻¹ dry soil (35),
and \( 10^4 \) \( \text{narG} \) gene copies g⁻¹ soil (36). Thus, compared to qPCR, our approach was suitable for
the specific and sensitive detection of virulence genes present in low numbers.

Nested PCR has the limitation that only the presence or absence of a specific gene can be
assessed. However, we applied a semi-quantitative approach based on the fraction of positive
to total number of replicate PCRs. This was possible because the virulence genes were present
in numbers very close to the detection limit, hence, of stochastic reasons, not all the replicate
18 PCRs resulted in a positive amplification (Table 3). Since the number of positive to
negative PCR tubes thus followed the Poisson distribution, it was possible to determine the
relative abundance of the investigated virulence determinants by calculating the fraction of
positive reactions.

A common problem with sensitive methods such as nested PCR is the risk of obtaining false
positive results (37). To minimize the risk of detecting false positive virulence determinants,
the primers targeting the 22 virulence genes were subjected to homology search using NCBI
(29,30) and found to match the genes of interest. Annealing temperatures and primer lengths,
which are the main factors influencing the specificity of the amplification (38), were used as
described in the literature or carefully selected with respect to melting temperatures.
Nevertheless, several cases of unspecific PCR amplifications were observed, although only
when the specific gene of interest was absent (e.g. *fimH*, Fig. 1). In other words, in all positive
detections only a single band was observed after agarose gel electrophoresis. As a further
measure to eliminate false positives, sequencing of positive PCR products confirmed the
identity of the bands. Finally, numerous negative controls were included in all PCR
experiments. In no case was a negative control observed to be positive.

The study was based on DNA extracted directly from the natural environments. As a
consequence, it was difficult to assess if the 17 detected genes were present in naturally
occurring environmental bacteria, in bacteria spread to the environments from human or
animal sources (39-41), contained in free phage particles, or existed as free DNA. Clinically
relevant bacteria may survive for varying periods of time in the environment depending on
their ability to cope with the environmental conditions (42). Environmental survival of
potential human pathogens has for instance been seen for *Vibrio* spp., *Aeromonas* spp., *E. coli*,
*Campylobacter* spp., *Arcobacter* spp. (43), *Salmonella* spp., *Clostridium perfringens*,
*Pleissmonas shigelloides* (44), and *H. pylori* (45). However, since we did not detect *Salmonella*
ssp. or *Helicobacter* ssp. in any of the environments and only found *E. coli*, *B. cereus* group
bacteria and *A. tumefaciens* in some of the environments, it is likely that the majority of
virulence genes were present amongst the natural bacterial (or phage) communities. This is
supported by the cases of separate clustering of sequences (Fig. 3 and Supplemental Fig. S1,
A-D), which suggests independent evolution. The presence of *E. coli* in the agricultural soil and
freshwater biofilm may be related to the fertilization of this soil with manure and discharge of
wastewater, respectively, and we cannot exclude that the virulence genes detected in these
environments were related to clinical strains. Bacilli and *Agrobacterium* are natural
components of soil microbial communities and it is, therefore, not surprising that these
bacteria were found in soil.
Differences in the occurrence of the virulence genes were observed between the four environments with the highest number being present in the beech wood soil [16], followed by the roadside and organic agricultural soils [14 each], and the freshwater biofilm [11]. It was unexpected that the organic agricultural soil did not have the highest number of virulence determinants as fertilization with manure would be expected to have positively influenced the number of bacteria possessing these genes (46). Poor survival of the bacteria is probably the reason why that was not the case. The lowest number of virulence genes was detected in the freshwater biofilm. The presence of *E. coli* in this environment suggested the freshwater lake to be influenced by human discharges and, as for the organic agricultural soil, a large number of virulence genes were to be expected. It is difficult to assess if fewer pathogenic traits were actually present in this environment, but the 2.5 × higher detection limit compared to the soil environments is probably part of the reason for the relatively few observed genes.

A priori, we hypothesized that bacteria in outer non-host environments are exposed to pressures selecting for survival traits that may confer virulence in host environments, and that differences in the occurrence of virulence genes can be related to different environmental pressures. The widespread distribution of the virulence determinants in the soil and freshwater environments provided circumstantial evidence that environmental pressures indeed select for traits that potentially confer virulence in host environments. Further proof of the hypothesis was the high degree of homology (90 – 100%) between the environmental and clinical virulence gene sequences and the fact that the large majority of the environmental sequences clustered together with the clinical sequences (Fig. 3 and Supplemental Fig. S1, A-D). This high degree of sequence conservation suggests that the virulence factors are housekeeping genes, essential for the survival and proliferation of the bacterial cells not only in human host environments but also in outer environments. With respect to the second part
of the hypothesis, that differences in the occurrence of virulence genes can be related to different environmental pressures, the principal component analysis indicated that about half of the variation (46%) in the distribution of the virulence traits could be explained by differences in PAH concentration (PC1) and pH (PC2) (Fig. 2). pH is known to affect microbial communities by influencing the carbon and nutrient availability and by controlling the biomass composition of fungi and bacteria (47). The mechanism behind the apparent effect of pH on the presence of the virulence genes is, however, not evident. The rationale for including contaminated roadside soil in the study was that "accessory" genes involved in e.g. pathogenicity, antibiotic resistance, or degradation of xenobiotics are often clustered together on mobile genetic elements such as plasmids and pathogenicity islands (PAIs) (48). Therefore, we expected the added stresses in the PAH-contaminated roadside soil would have led to selective conditions that promoted the existence of mobile genetic elements containing virulence factors. Our data, however, did not provide evidence that this was the case. Firstly, the high PAH concentration in the roadside soil did not result in more virulence traits compared to the other soils and secondly, although plasmids were present in the roadside soil their numbers did not seem to be higher than in the beech wood and organic agricultural soils (unpublished data). Thus, similar to pH, the PAH concentration appeared to influence the distribution of the virulence determinants but the potential underlying mechanism remains elusive.

The variation explained by the 2 principal components was relatively low (Fig. 2). This suggests that factors other than pH and PAH concentration affected the occurrence of virulence genes in the four environments. Bacteria have developed different defence mechanisms against predation including secretion of toxins, avoidance of lysozymal killing, and intracellular multiplication. Virulence factors of *Legionella*, *Mycobacterium*, *Pseudomonas*,...
and *Vibrio* have for instance been found to be important for the interaction with invertebrates and mammalian hosts including humans (49). Therefore, another possible factor controlling the presence of the virulence factors could be the interaction with bacterial predators such as free-living protozoa and nematodes.

In conclusion, we demonstrated the occurrence of numerous bacterial human virulence genes in soil and freshwater environments. A high degree of sequence conservation between the clinical and environmental genes sequences was observed. Altogether, the data suggest that outer environments constitute reservoirs of several groups of bacterial virulence determinants that possibly play an essential role for the survival and adaptation of the bacteria. Further studies are needed to investigate to which extent the virulence genes are expressed by the environmental bacterial communities.
ACKNOWLEDGMENTS

The work was supported by a grant (645-08-0110) from The Danish Agency for Science, Technology and Innovation to The Postgraduate School of Environmental Chemistry and Toxicology (RECETO). Jan H. Christensen is thanked for valuable advice on the principal component analysis.
REFERENCES


<table>
<thead>
<tr>
<th>Virulence group</th>
<th>Virulence factor</th>
<th>Gene</th>
<th>Bacterial strain known to carry target gene</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toxin</td>
<td>B component of hemolysin BL</td>
<td>hblA</td>
<td>Bacillus thuringiensis serovar tenebrionis</td>
<td>DSM 5526</td>
</tr>
<tr>
<td>Shiga-like toxin 1</td>
<td>cytK2</td>
<td>cytK2</td>
<td>Bacillus thuringiensis serovar kurstaki</td>
<td></td>
</tr>
<tr>
<td>Shiga-like toxin 2</td>
<td>stx-2</td>
<td>stx-2</td>
<td>Escherichia coli O157</td>
<td>SSI² C45-89</td>
</tr>
<tr>
<td>Vacuolating cytotoxin</td>
<td>vacA</td>
<td>vacA</td>
<td>Helicobacter pylori NCTC 11638</td>
<td>DSM 10242</td>
</tr>
<tr>
<td>Hemolysin A</td>
<td>hylA</td>
<td>hylA</td>
<td>Escherichia coli CFT073</td>
<td>ATCC 700928</td>
</tr>
<tr>
<td>Cytotoxin K1</td>
<td>cytK1</td>
<td>cytK1</td>
<td>Bacillus cereus NVH 391-98</td>
<td>DSM 22905⁵</td>
</tr>
<tr>
<td>Edema factor</td>
<td>cya</td>
<td>cya</td>
<td>Bacillus anthracis 2160</td>
<td>SSI² 2160</td>
</tr>
<tr>
<td>Adhesin</td>
<td>papC</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Type 1 fimbriae</td>
<td>fimH</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>F1C fimbriae</td>
<td>focG</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Secretion system</td>
<td>invA</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Regulation of virulence determinants</td>
<td>spaA</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Type IV (conjugative DNA transfer)</td>
<td>virD4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Type III, translocators and effectors</td>
<td>sipB-sipC</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Regulator of virulence mediators</td>
<td>phoQ</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Inflammatory mediator</td>
<td>wblF</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Bacterial resistance</td>
<td>arnT</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>L-Ara4N transferase</td>
<td>yfbI</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Capsule biosynthesis protein</td>
<td>capA</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

⁵SSI-strains derived by Statens Serum Institute, [http://www.ssi.dk/Bestil/SSI%20Diagnostica/Produkter%20fra%20SSI%20Diagnostica/Bakteriestammer.aspx](http://www.ssi.dk/Bestil/SSI%20Diagnostica/Produkter%20fra%20SSI%20Diagnostica/Bakteriestammer.aspx)
<table>
<thead>
<tr>
<th></th>
<th>Beech wood soil</th>
<th>Roadside soil</th>
<th>Organic agricultural soil</th>
<th>Freshwater biofilm</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>7.3</td>
<td>7.3</td>
<td>7.0</td>
<td>7.4±</td>
</tr>
<tr>
<td>Texture:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 2 µm</td>
<td>14.2</td>
<td>13.8</td>
<td>16.4</td>
<td>-</td>
</tr>
<tr>
<td>2 - 50 µm</td>
<td>25.9</td>
<td>21.2</td>
<td>23.2</td>
<td>-</td>
</tr>
<tr>
<td>50 - 250 µm</td>
<td>38.2</td>
<td>37.2</td>
<td>31.3</td>
<td>-</td>
</tr>
<tr>
<td>250 - 2000 µm</td>
<td>20.7</td>
<td>19.1</td>
<td>17.5</td>
<td>-</td>
</tr>
<tr>
<td>&gt; 2000 µm</td>
<td>0.9</td>
<td>7.5</td>
<td>11.7</td>
<td>-</td>
</tr>
<tr>
<td>Total nitrogen</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%</td>
<td>0.27</td>
<td>0.36</td>
<td>0.23</td>
<td>-</td>
</tr>
<tr>
<td>Phosphorus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>µg g⁻¹</td>
<td>0.99</td>
<td>4.8</td>
<td>2.3</td>
<td>-</td>
</tr>
<tr>
<td>DOC (water⁻)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ppm</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>11.1 ± 0.6</td>
</tr>
<tr>
<td>Loss on ignition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%</td>
<td>8.0 ± 0.3</td>
<td>8.1 ± 0.3</td>
<td>6.1 ± 1.7</td>
<td>-</td>
</tr>
<tr>
<td>PAH, total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>µg kg⁻¹ DW</td>
<td>94.1 ± 5.4</td>
<td>2537.4 ± 555.2</td>
<td>22.5 ± 4.3</td>
<td>-</td>
</tr>
<tr>
<td>Water content</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%</td>
<td>26.4 ± 3.9</td>
<td>24.1 ± 1.7</td>
<td>21.1 ± 0.9</td>
<td>-</td>
</tr>
<tr>
<td>WHC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%</td>
<td>33.2 ± 2.6</td>
<td>32.4 ± 3.9</td>
<td>28.7 ± 1.4</td>
<td>-</td>
</tr>
<tr>
<td>Number of bacteria</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>× 10⁸ g⁻¹</td>
<td>2.1 ± 0.3</td>
<td>2.2 ± 0.2</td>
<td>2.2 ± 0.2</td>
<td>-</td>
</tr>
</tbody>
</table>

* Polycyclic aromatic hydrocarbons (PAHs), water holding capacity (WHC), dissolved organic carbon (DOC), dry weight (DW).

* Measured in water samples collected at the same time and location as the biofilm.
### Table 3: Distribution and relative frequency (number of positive PCR reactions to the total number of PCR reactions) of 22 selected virulence genes in three soils and a freshwater biofilm

<table>
<thead>
<tr>
<th>Virulence group</th>
<th>Virulence factor</th>
<th>Gene</th>
<th>BWS</th>
<th>RS</th>
<th>OAS</th>
<th>BIO</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Toxin</strong></td>
<td>B component of hemolysin BL</td>
<td>hblA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shiga-like toxin 1</td>
<td>stx1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cytotoxin K2</td>
<td>cytK2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shiga-like toxin 2</td>
<td>stx2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vacuolating cytotoxin</td>
<td>vacA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hemolysin A</td>
<td>hlyA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cytotoxin K1</td>
<td>cytK1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Edema factor</td>
<td>cya</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Adhesin</strong></td>
<td>P fimbriae</td>
<td>papC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>P fimbriae</td>
<td>papA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type I fimbriae</td>
<td>fimH</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F1C fimbriae</td>
<td>focG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Secretion system</strong></td>
<td>Type III, secretion apparatus</td>
<td>invA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type III, secretion apparatus</td>
<td>spiA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type IV (conjugative DNA transfer)</td>
<td>virD4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type III, translocators and effectors</td>
<td>sipB-sipC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Regulator of virulence</strong></td>
<td>Regulator of virulence determinants</td>
<td>phoQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Inflammatory mediator</strong></td>
<td>Part of O-antigen 111</td>
<td>wbdI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Part of O-antigen 157</td>
<td>rbfE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Bacterial resistance</strong></td>
<td>L-Ara4N transferase</td>
<td>arnT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>L-Ara4N transferase</td>
<td>ybfI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Capsule biosynthesis protein</td>
<td>capA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Shading indicates the frequency of positive identifications in 18 separate PCR runs for each environment. Black: 50 < % ≤ 75, dark grey: 25 < % ≤ 50, light grey: 0 < % ≤ 25, white: 0 % (not detected).*  

BWS = beech wood soil; RS = roadside soil; OAS = organic agricultural soil; BIO = freshwater biofilm.
FIGURE LEGENDS

FIG 1 Detection of the E. coli adhesion gene fimH in beech wood soil (BWS) (A) and organic agricultural soil (OAS) (B), and estimation of the detection limit of fimH. fimH was targeted by two parallel nested PCRs of triplicate DNA extractions of triplicate soil samples (i.e. 18 PCRs in total per soil type). A: fimH was detected in 2 of 3 replicate beech wood soil samples and in 4 out of the total of 18 PCRs. B: fimH was not detected in any of the replicate soils samples of the organic agricultural soil. The detection limit of fimH was estimated by adding a known number of E. coli CFT073 carrying fimH to soil subsamples before DNA extraction. The detection limit was estimated to be 400 bacteria g⁻¹ soil. M is the molecular weight marker V (pBR322-Hae III).

FIG 2 Principal component analysis based on the occurrence and relative abundance of the virulence determinants in beech wood soil (BWS), roadside soil (RS), organic agricultural soil (OAS), and freshwater biofilm (BIO). Percentages in parentheses on the axes indicate the variation explained by the 2 principal components. Small sized data points show the scores of replicate samples. Large data points represent the average of the scores of the replicate samples of each environment. Error bars indicate standard errors of the mean.

FIG 3 Minimum evolution tree (unrooted) showing the phylogeny of the environmental invA sequences (black triangles). Bootstrap values of 6 - 98. Numbers in parentheses indicate numbers of identical invA sequences. The bacterial strain used as positive control is shown in bold.
FIG 1 Detection of the *E. coli* adhesion gene *fimH* in beech wood soil (BWS) (A) and organic agricultural soil (OAS) (B), and estimation of the detection limit of *fimH*. *fimH* was targeted by two parallel nested PCRs of triplicate DNA extractions of triplicate soil samples (i.e. 18 PCRs in total per soil type). A: *fimH* was detected in 2 of 3 replicate beech wood soil samples and in 4 out of the total of 18 PCRs. B: *fimH* was not detected in any of the replicate soils samples of the organic agricultural soil. The detection limit of *fimH* was estimated by adding a known number of *E. coli* CFT073 carrying *fimH* to soil subsamples before DNA extraction. The detection limit was estimated to be 400 bacteria g\(^{-1}\) soil. M is the molecular weight marker V (pBR322-*Hae* III).

238 bp

![DNA gel image](image_url)
FIG 2 Principal component analysis based on the occurrence and relative abundance of the virulence determinants in beech wood soil (BWS), roadside soil (RS), organic agricultural soil (OAS), and freshwater biofilm (BIO). Percentages in parentheses on the axes indicate the variation explained by the 2 principal components. Small sized data points show the scores of replicate samples. Large data points represent the average of the scores of the replicate samples of each environment. Error bars indicate standard errors of the mean.
**FIG 3** Minimum evolution tree (unrooted) showing the phylogeny of the environmental *invA* sequences (black triangles). Bootstrap values of 6 - 98. Numbers in parentheses indicate numbers of identical *invA* sequences. The bacterial strain used as positive control is shown in bold.