Bacterial human virulence genes across diverse habitats as assessed by In silico analysis of environmental metagenomes

Søborg, Ditte Andreasen; Hendriksen, Niels B.; Kilian, Mogens; Christensen, Jan H.; Kroer, Niels

Published in:
Frontiers in Microbiology

DOI:
10.3389/fmicb.2016.01712

Publication date:
2016

Document version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Bacterial Human Virulence Genes across Diverse Habitats As Assessed by In silico Analysis of Environmental Metagenomes

Ditte A. Søborg1,2, Niels B. Hendriksen1, Mogens Kilian3, Jan H. Christensen4 and Niels Kroer5*

1Department of Environmental Science, Aarhus University, Roskilde, Denmark, 2Department of Energy and Environment, VIA University College, Horsens, Denmark, 3Department of Biomedicine, Aarhus University, Aarhus, Denmark, 4Department of Plant and Environmental Sciences, University of Copenhagen, Frederiksberg, Denmark, 5Department of Biology, University of Copenhagen, Copenhagen, Denmark

The occurrence and distribution of clinically relevant bacterial virulence genes across natural (non-human) environments is not well understood. We aimed to investigate the occurrence of homologs to bacterial human virulence genes in a variety of ecological niches to better understand the role of natural environments in the evolution of bacterial virulence. Twenty four bacterial virulence genes were analyzed in 46 diverse environmental metagenomic datasets, representing various soils, seawater, freshwater, marine sediments, hot springs, the deep-sea, hypersaline mats, microbialites, gutless worms and glacial ice. Homologs to 16 bacterial human virulence genes, involved in urinary tract infections, gastrointestinal diseases, skin diseases, and wound and systemic infections, showed global ubiquity. A principal component analysis did not demonstrate clear trends across the metagenomes with respect to occurrence and frequency of observed gene homologs. Full-length (>95%) homologs of several virulence genes were identified, and translated sequences of the environmental and clinical genes were up to 50–100% identical. Furthermore, phylogenetic analyses indicated deep branching positions of some of the environmental gene homologs, suggesting that they represent ancient lineages in the phylogeny of the clinical genes. Fifteen virulence gene homologs were detected in metatranscriptomes, providing evidence of environmental expression. The ubiquitous presence and transcription of the virulence gene homologs in non-human environments point to an important ecological role of the genes for the activity and survival of environmental bacteria. Furthermore, the high degree of sequence conservation between several of the environmental and clinical genes suggests common ancestral origins.

Keywords: bacteria, environmental metagenomes, evolution, in silico, non-host, virulence
INTRODUCTION

Bacterial pathogens continue to cause problems for humans by the continuous evolution of known pathogens and emergence of new pathogens (Jackson et al., 2011). Approximately 160 new bacterial infectious diseases have been discovered between 1940 and 2010, and emerging infectious diseases constitute a significant burden on global economies and public health (Jones et al., 2008). While the mechanisms by which many pathogenic bacteria cause disease are relatively well-described, the evolutionary origin of virulence is not fully understood (Wilson et al., 2002; Persson et al., 2009). Bacterial virulence determinants causing disease in humans may evolve through complex host-pathogen interactions (Wilson and Salyers, 2003; Mikonranta et al., 2012) but their origin may also lie in non-human reservoirs amongst the environmental microbiota (Martínez, 2012).

The large majority of studies on virulence in the environment have focused on the detection and survival of clinically relevant bacteria in non-host environments (Garcia-Aljaro et al., 2009; Vezzulli et al., 2010; Davidson et al., 2015). Only few studies (e.g., Persson et al., 2009; Søborg et al., 2013) have targeted the environmental gene pool of pathogenic traits, hence evidence of the presence of virulence genes in environmental microbiomes is sparse.

By PCR and cultivation-based approaches we have previously investigated the presence of virulence factors in soil and freshwater environments (Søborg et al., 2013, 2014). However, to better understand (i) the significance of virulence gene homologs for bacterial survival in non-human environments and (ii) the role these environments play for the evolution of bacterial human pathogens, a much broader investigation across multiple environmental microbiomes is required. The increasing availability of environmental metagenomes provides an ideal opportunity for interrogating the occurrence of bacterial virulence genes in natural environments. Hence, we performed an in silico analysis targeting 24 bacterial human virulence genes in 46 environmental bacterial metagenomes. The analysis showed that not only were the bacterial virulence gene homologs widespread in natural (non-human) environments, evidence of bacterial human pathogenicity and a high degree of sequence conservation between some of the environmental and clinical genes was observed. The results provide documentation of an important role of virulence gene homologs in environmental microbiomes and suggest they may be precursors of virulence genes seen in clinically relevant pathogenic bacteria.

MATERIALS AND METHODS

Blunt Searches for Virulence Determinants in Metagenomic Datasets

Complete translated sequences of 24 bacterial virulence genes were used as queries in TBLASTN searches in 46 environmental metagenomes of the Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) database (Sun et al., 2011; note that the database has been transferred to iMicrobe (http://data.imicrobe.us). The genes covered several classes of virulence determinants including toxin genes (hlyA, hlyC, ply, stx1, stx2, vacA, hlgB), adhesin genes (papA, papH, papC, fimF, focG), secretion genes (invA, spIA, spIB, sipC, ssAD), inflammatory genes (wbdl, rfbE), regulatory genes (phoP, phoQ) and resistance genes (mecA, ybfI) originally described from Salmonella spp., Escherichia coli/Shigella, Helicobacter spp., Streptococcus pneumoniae, and Staphylococcus aureus (Table 1).

The metagenomes represented terrestrial, marine, limnic as well as several extreme environments (Supplementary Table 1). Environments potentially impacted by human or livestock discharges were excluded from the analysis as were metagenomes restricted to specific bacterial groups.

Default NCBI blastall parameters (expected threshold: 10; word size: 3; matrix: BLOSUM62; gap open cost: 11; gap extend cost: 1) were applied. Hits with an expect (E) value < 10^-6 and

TABLE 1 | Bacterial human virulence genes used for TBLASTN analysis of environmental metagenomic datasets.

<table>
<thead>
<tr>
<th>Virulence factor</th>
<th>Gene</th>
<th>Protein size (aa)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shiga-like toxin 1</td>
<td>stx1</td>
<td>315</td>
<td>ACU32680.1</td>
</tr>
<tr>
<td>Shiga-like toxin 2</td>
<td>stx2</td>
<td>319</td>
<td>YP_00324845.1</td>
</tr>
<tr>
<td>Vacuolating cytotoxin</td>
<td>vacA</td>
<td>1296</td>
<td>YP_002301516.1</td>
</tr>
<tr>
<td>Haemolysin</td>
<td>hlyA</td>
<td>1024</td>
<td>NP_755445.1</td>
</tr>
<tr>
<td>Haemolysin</td>
<td>hlyA</td>
<td>1066</td>
<td>NP_754851.1</td>
</tr>
<tr>
<td>Pneumolysin</td>
<td>plv</td>
<td>471</td>
<td>NP_346351.1</td>
</tr>
<tr>
<td>Gamma-haemolysin component C precursor</td>
<td>hlgB</td>
<td>325</td>
<td>YP_041861.1</td>
</tr>
<tr>
<td>P fimbiae</td>
<td>papC</td>
<td>839</td>
<td>NP_755465.1</td>
</tr>
<tr>
<td>P fimbiae</td>
<td>papA</td>
<td>199</td>
<td>NP_755467.1</td>
</tr>
<tr>
<td>P fimbiae</td>
<td>papH</td>
<td>195</td>
<td>NP_755466.1</td>
</tr>
<tr>
<td>Type 1 fimbiae</td>
<td>fimH</td>
<td>303</td>
<td>NP_757248.1</td>
</tr>
<tr>
<td>F1C fimbiae</td>
<td>focG</td>
<td>167</td>
<td>NP_753158.1</td>
</tr>
<tr>
<td>Type III, secretion apparatus</td>
<td>invA</td>
<td>685</td>
<td>YP_151923.1</td>
</tr>
<tr>
<td>Type III, secretion apparatus (SPP-2)</td>
<td>spIA</td>
<td>497</td>
<td>YP_150710.1</td>
</tr>
<tr>
<td>Type III, secretion apparatus (SPP-2)</td>
<td>ssaD</td>
<td>403</td>
<td>YP_150708.1</td>
</tr>
<tr>
<td>Type III, translocators and effectors (SPP-1)</td>
<td>sipB</td>
<td>593</td>
<td>YP_151912.1</td>
</tr>
<tr>
<td>Type III, translocators and effectors (SPP-1)</td>
<td>sipC</td>
<td>409</td>
<td>YP_151911.1</td>
</tr>
<tr>
<td>Regulator of virulence determinants</td>
<td>phoP</td>
<td>224</td>
<td>YP_150858.1</td>
</tr>
<tr>
<td>Regulator of virulence determinants</td>
<td>phoQ</td>
<td>487</td>
<td>YP_150859.1</td>
</tr>
<tr>
<td>Part of O-antigen 111</td>
<td>wbdl</td>
<td>149</td>
<td>AAC4881.1</td>
</tr>
<tr>
<td>Part of O-antigen 157</td>
<td>rbdE</td>
<td>364</td>
<td>NP_310888.1</td>
</tr>
<tr>
<td>L-Ara4N transferase</td>
<td>pgaB (lntT region)</td>
<td>548</td>
<td>YP_149876.1</td>
</tr>
<tr>
<td>L-Ara4N transferase</td>
<td>yfbI (lntT region)</td>
<td>550</td>
<td>NP_754865.1</td>
</tr>
<tr>
<td>Penicillin-binding protein</td>
<td>mecA</td>
<td>668</td>
<td>YP_039515.1</td>
</tr>
</tbody>
</table>

*Salmonella Pathogenicity Island (SPI).*
a sequence similarity >30% were considered to be significant (Persson et al., 2009). We further applied the criterium of a minimal length of 35 amino acids (aa) to weed-out the shortest 454 pyrosequencing sequences. Longest TBLASTN hits were aligned against the query sequences, and query coverage and percent aa identity determined. As a final criterium to confirm that the environmental hits indeed matched known classes of virulence genes, NCBI BLASTP searches (default parameters) were done using environmental hits as queries.

**Blast Searches for Clinically Relevant Bacteria in Metagenomic Datasets**

Bacterial groups from which the virulence genes were originally described (Salmonella spp., E. coli/Shigella, Helicobacter spp., Streptococcus pneumoniae and Staphylococcus aureus) were targeted by BLASTN (default parameters, E-value < 10^-6) in the metagenomes. Partial 16S/23S rRNA gene sequences usually amplified with published PCR primers were used as queries (Supplementary Table 2). BLASTN hits were checked for presence of the primer sequences as well as for similarity to the 16S or 23S gene sequences of the clinical bacteria.

**Phylogenetic Analysis of Virulence Gene Homologs**

Phylogenetic trees showing the relationships between the translated environmental sequences and the corresponding sequences of the clinical genes, were prepared with MEGA Version 5.01 (Tamura et al., 2011) using the Minimum Evolution algorithm (pairwise deletion). In addition, the Maximum Parsimony method of MEGA Version 5.01 was employed to create phylogenetic trees (complete deletion; 500 replicates) to infer the evolutionary history of the homologs in comparison with the clinically relevant genes.

**Statistical Analysis**

To look for trends that potentially could explain differences between metagenomes with respect to presence of virulence gene homologs, a Principal Component Analysis (PCA) was carried out. The PCA was performed with the Matlab based chemometric software package Latentix 2.0 (Latentix Aps, Copenhagen N, Denmark, www.latentix.com) using the data set shown in Table 2. A PCA model, X, was built as the product of an I × K column-wise orthogonal score matrix T and the transpose of the column-wise orthogonal J × K loading matrix P, where I denotes the number of samples (metagenomes), J the number of variables (virulence gene homologs), and K the number of principal components. E is the matrix of model residuals of size I × J (Equation 1). Zeros were inserted in X if genes were not detected in a specific environment.

\[ X = TP^T + E \] (1)

The data set was autoscaled across metagenomes (i.e., the mean value of a variable across all metagenomes was subtracted from the data set (mean-centered) and the mean-centered data divided by the standard deviation) prior to PCA. Autoscaling was needed to allow each virulence gene to contribute equally to the PCA model, independent of the overall abundance of each gene. The model was validated by cross-validation (Martens and Martens, 2001).

Only metagenomes consisting of subsamples from similar environments were included in the analysis, hence, 10 metagenomes were excluded (MontereyBay transect, BotanyBay:Sanger, GuaymasBasin, MILOCO:454Shotgun, HOT:Sanger, AntarcticaAquatic:454, BotanyBay:454, AntarcticaAquatic:Sanger, MILOCO:454, GOS). Furthermore, one of the observed 16 virulence genes (papH) was not included in the PCA as it was not found in any of the 34 metagenomes included in the analysis.

**RESULTS AND DISCUSSION**

Homologs of 16 of the 24 investigated bacterial human virulence genes were identified in the metagenomic datasets (Table 2). The genes represented toxin, adhesin, secretion system, regulatory, inflammatory, and resistance virulence groups involved in urinary tract infections, gastrointestinal diseases, skin and foreign body infections, and systemic infections. To confirm that environmental homologs matched known virulence factors in the databases, we took the longest environmental homologs with the highest identity to the virulence genes and did a NCBI BLASTP search using the environmental sequences as queries. The environmental sequences matched relevant virulence groups in 95% of the cases (Supplementary Table 3).

A positive relationship between numbers of detected genes and metagenome size was observed (Figure 1). Highest numbers of observed genes were seen in some of the largest metagenomes, but sequencing of more than 3–4 Gb generally did not result in the detection of further genes. Several of the investigated metagenomes were obtained by 454 pyrosequencing with average sequence reads of 100–400 bp. Some metagenomes were based on Sanger sequencing or on assembled sequences with average reads of 600–1000 bp (22). Since short reads (≤400 bp) tend to miss a significant amount of the BLAST homologs detected by long reads (Wommack et al., 2008), this may explain the high number of homologs detected in some of the relatively small Sanger-based metagenomes such as the "HydrothermalVent" (50 Mb) (Table 2 and Supplementary Table 1).

The most frequently observed virulence gene homologs were the inflammatory O-antigen gene, rfbE, the regulator gene, phoP, and the Type III secretion apparatus gene, invA, which were detected in ≥85% of the metagenomes (Table 2). These genes also showed the highest relative abundance within the individual metagenomes (Table 2). Homologs to the regulator gene phoQ, the toxin gene hlyA, the fimbrial gene papC, the Type III secretion apparatus gene spiA, the inflammatory O-antigen gene wbbI, and the resistance genes pqaB, yfbI and mecA were also found widely distributed (28–72% of the metagenomes). On the other hand, homologs to the fimbrial genes papA, papH, fimH, and focG and the toxin gene hylA were only found in 2–10% of the metagenomes (Table 2).

To assess if the origin of the virulence gene homologs was truly environmental or due to the presence of the
<table>
<thead>
<tr>
<th>Metagenome</th>
<th>Toxin</th>
<th>Adhesin</th>
<th>Secretion</th>
<th>Regulatory</th>
<th>Inflammatory</th>
<th>Resistance</th>
<th>Genes detected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>hlyA</td>
<td>hylA</td>
<td>papA</td>
<td>papH</td>
<td>psGC</td>
<td>fimH</td>
<td></td>
</tr>
<tr>
<td>Microbialites (Marine microbial mat)</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>FLAS (Saltern)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Linelsland (Seawater)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>PeruMarginSediment (Seawater)</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Mountain lake (Freshwater)</td>
<td>37</td>
<td>74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>GeneExpression (Seawater)</td>
<td>81*</td>
<td>162*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Contaminated soil</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>AmazonRiverPlume (Estuarine)</td>
<td>58</td>
<td>43</td>
<td>1</td>
<td>41*</td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>SargassomaSea (Seawater)</td>
<td>186</td>
<td>127</td>
<td>17</td>
<td>214</td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>DeepMed (Seawater)</td>
<td>498</td>
<td>134</td>
<td>197*</td>
<td>8*</td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Yellowstone (Hot spring)</td>
<td>355</td>
<td>135</td>
<td>779</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>AcidoMine (Biofilm)</td>
<td>212</td>
<td>734</td>
<td>28</td>
<td>295</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>PacificOcean (Seawater)</td>
<td>106</td>
<td>85</td>
<td>1</td>
<td>51*</td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>PBSM (Marine beach sand)</td>
<td>167</td>
<td>834</td>
<td>500</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>MontereyBay transect (Seawater)</td>
<td>33</td>
<td>101</td>
<td>2</td>
<td>512</td>
<td>2</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>HOT:454 (Seawater)</td>
<td>37</td>
<td>177</td>
<td>4</td>
<td>956</td>
<td>5</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>N. Rec. Line67 (Seawater)</td>
<td>15</td>
<td>40</td>
<td>29</td>
<td>2</td>
<td>1</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>BotanyBay:Sanger (Seawater)</td>
<td>13</td>
<td>3</td>
<td>764</td>
<td>81</td>
<td>527*</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>AlvinellaPompejana (Hydrothermal vent)</td>
<td>10</td>
<td>4</td>
<td>286</td>
<td>1078</td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>HydrothermalVent</td>
<td>101</td>
<td>182</td>
<td>706</td>
<td>40</td>
<td>948</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>TermiteGut</td>
<td>16</td>
<td>246</td>
<td>410</td>
<td>66</td>
<td>689</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>SalternMetagenome</td>
<td>5*</td>
<td>7*</td>
<td>35*</td>
<td>14*</td>
<td>2*</td>
<td>5*</td>
<td>14*</td>
</tr>
<tr>
<td>MontereyBay (Seawater)</td>
<td>123</td>
<td>337</td>
<td>7</td>
<td>88</td>
<td>4</td>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td>GuaymasBasin (Hydrothermal Vent)</td>
<td>49</td>
<td>122</td>
<td>27</td>
<td>729</td>
<td>8</td>
<td></td>
<td>34</td>
</tr>
<tr>
<td>Bacterioplankton (Seawater)</td>
<td>3</td>
<td>110</td>
<td>3</td>
<td>250</td>
<td>3</td>
<td>503</td>
<td>3</td>
</tr>
<tr>
<td>BATS (Seawater)</td>
<td>1</td>
<td>1</td>
<td>51</td>
<td>151</td>
<td>3</td>
<td>445</td>
<td>2</td>
</tr>
<tr>
<td>BermudaOceanic (Seawater)</td>
<td>4</td>
<td>92</td>
<td>454</td>
<td>33</td>
<td>1141</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>IceMetagenome (Freshwater ice)</td>
<td>4</td>
<td>4</td>
<td>138</td>
<td>960</td>
<td>13</td>
<td>605</td>
<td>13</td>
</tr>
<tr>
<td>HypersalineMat (Seawater)</td>
<td>83*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CellCapture (Marine sediment)</td>
<td>63</td>
<td>333</td>
<td>10</td>
<td>1553</td>
<td>135</td>
<td>1053</td>
<td>156</td>
</tr>
<tr>
<td>Drifting-ESP (Seawater)</td>
<td>0.45</td>
<td>0.45</td>
<td>0.23</td>
<td>21</td>
<td>58</td>
<td>1</td>
<td>152</td>
</tr>
</tbody>
</table>

(Continued)
### TABLE 2 | Continued

<table>
<thead>
<tr>
<th>Metagenome</th>
<th>Toxin</th>
<th>Adhesin</th>
<th>Secretion</th>
<th>Regulatory</th>
<th>Inflammatory</th>
<th>Resistance</th>
<th>Genes detected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>hlyA</td>
<td>hyIA</td>
<td>papA</td>
<td>papH</td>
<td>papC</td>
<td>focG</td>
<td>fimH</td>
</tr>
<tr>
<td>GutlessWorm</td>
<td>16</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOT: Sanger (Seawater)</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WesternChannelOMM (Seawater)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MILOCO:454 Shotgun (Seawater)</td>
<td>8</td>
<td>61</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Washington Lake (Freshwater)</td>
<td>5</td>
<td>71</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BotanyBay:454 (Seawater)</td>
<td>2*</td>
<td>71</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AntarcticaAquatice:Shotgun (Seawater)</td>
<td>1</td>
<td>41</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YLAKE (Hot spring)</td>
<td>0.25*</td>
<td>6</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AntarcticaAquatice:454 (Seawater)</td>
<td>0.44</td>
<td>0.08</td>
<td>19*</td>
<td></td>
<td></td>
<td>136*</td>
<td>1*</td>
</tr>
<tr>
<td>BisonMetagenome (Hot spring)</td>
<td>11*</td>
<td>2*</td>
<td>8*</td>
<td></td>
<td></td>
<td>144*</td>
<td>1*</td>
</tr>
<tr>
<td>MILOCO:454 (Seawater)</td>
<td>0.08*</td>
<td>1*</td>
<td>18*</td>
<td>2*</td>
<td>4*</td>
<td>19*</td>
<td>2</td>
</tr>
<tr>
<td>GOS (Seawater)</td>
<td>4*</td>
<td>0.08</td>
<td>4*</td>
<td>19*</td>
<td>0.16*</td>
<td>16*</td>
<td>92</td>
</tr>
<tr>
<td>DayNight (Seawater)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SapeloIsland (Seawater)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*The prevalence of the individual genes in the different metagenomes were normalized to the metagenomic library sizes, i.e., reported values are frequencies per billion bases sequenced. Virulence genes not detected in any of the metagenomic datasets were phyA, vacA, stx1, stx2, NrgB, sipB, sipC, and ssaD. Please refer to Supplementary Table 1 for a description of the metagenomes.

*Detection (BLASTN; E-value < 10^-6) of partial 16S/23S rRNA genes of clinical strains known to carry the virulence genes.
clinically relevant bacteria from which the virulence factors were originally described, partial 16S or 23S rRNA gene sequences of *Salmonella* spp., *E. coli*, *Helicobacter* spp., *Streptococcus pneumoniae* and *Staphylococcus aureus* (Supplementary Table 2) were targeted by BLASTN in the metagenomes. Although we specifically avoided metagenomes influenced by livestock and human discharges, relevant bacterial groups and their respective virulence genes co-occurred 66 out of the 313 times a virulence gene homolog was observed (Table 2). Only *Salmonella* spp., *E. coli*, and *Helicobacter* spp. were observed. The presence of the bacterial groups may be explained by contamination during sample handling, survival and spreading from farms and wastewater treatment plants, or by the natural occurrence of environmental strains of the bacteria. We argue that the presence of the bacterial groups most likely is due to their natural occurrence. For instance, environmental survival of *E. coli* and *Salmonella* is well documented (Mauger et al., 2004; Miller et al., 2006) and may be occurring in e.g., coastal environments. However, *E. coli* and *Salmonella* were also identified in oceanic, hot spring, and hypersaline environments (Table 2), and albeit contamination by these bacteria cannot be ruled out, it does suggest that they naturally inhabit “unexpected” environments. If, on the other hand, it is assumed that the observation of clinically relevant bacteria is caused either by contamination or by livestock and human discharges, and all cases of dual presence of virulence genes and bacteria are excluded from the analysis, 80% of the detections of virulence gene homologs would still be unbiased by the presence of clinically relevant bacteria. Since 16S/23S rRNA genes contrary to virulence genes often are found in multiple copies per cell (Vos et al., 2012), the risk that 16S/23S rRNA genes of clinically relevant bacteria were overlooked in the metagenomic datasets is miniscule. Thus, the large majority of our data supports the existence of a natural environmental reservoir of genes normally associated with human pathogenic bacteria.

**Distribution of Virulence Gene Homologs across Environmental Metagenomes**

The bacterial virulence gene homologs were present in virtually all the metagenomes representing soil, seawater, freshwater, marine sediment, hot spring, the deep-sea, hypersaline mats, microbialites, gutless worms, and glacial ice environments (Table 2). Hence, environments were characterized by highly contrasting environmental conditions in terms of pH, pressure, temperature, salinity, and nutrient availability. Moreover, the virulence gene homologs were globally distributed on six continents and in four oceans (Figure 2), i.e., all continents and oceans for which relevant metagenomes were available. The maximal number of different genes detected within one metagenome was 16 (Table 2). Only in two metagenomes were no genes detected (Table 2). These metagenomes originated from surface water off Hawaii, and coastal sea water off the Sapelo Island (USA) (Table 2 and Supplementary Table 1). The fact that more than half of the sequences of the Hawaiian metagenome were cyanobacterial may explain the lack of virulence gene homologs in this metagenome. With respect to the Sapelo Island metagenome, homologs to virulence genes were observed in another and larger metagenome from Sapelo Island (the Sapelo2008 metagenome; Table 2). Thus, the lack of virulence gene homologs in the SapeloIsland metagenome was most likely due to its small size.

Metagenomes representing comparable environments (e.g., seawater) did not share a set of typical or common virulence gene homologs. To discern relative differences between the metagenomes with respect to occurrence and frequency of the gene homologs, a principal component analysis (PCA) was performed. The variation explained by the first two principal components was relatively low (43%) yet useful for describing correlations between metagenomes and virulence gene homologs (Figure 3). No general trends were observed as the bulk of the metagenomes clustered relatively closely together with low, and in some cases, negative scores for both PC1 and PC2. However, the CellCapture and Bison metagenomes had relatively high positive PC1 and PC2 scores, while a high positive PC1 and a large negative PC2 score were observed for Washington Lake. The DeepMed metagenome, on the other hand, had a low PC1 score and a very large negative PC2 score. The high positive PC1 scores of the adhesion genes *papC* and *focG*, the resistance genes *pqaB* and *yjbI*, and the secretion gene *spiA*, indicated that these genes were relatively over-represented in BisonMetagenome and CellCapture as compared to the other metagenomes (Figure 3 and Table 2). The Washington Lake metagenome had above average gene frequencies of the regulatory gene *phoP*, the resistance gene *mecA* and the inflammatory gene *rbfE*, whereas the DeepMed metagenome had a relative over-representation of the toxin gene *hlyA* in comparison to the other metagenomes (Figure 3 and Table 2).
Homologs to eight of the investigated virulence genes (vacA, stx1, stx2, sipB, sipC, hlgB, ply, ssaD) were not identified in any of the metagenomes. According to the databases, these genes have narrow phylogenetic host ranges, limited to only one or two bacterial classes/families (Søborg et al., 2014; unpublished data). This narrow phylogenetic host range, however, is unlikely to be the reason why homologs were not identified in the metagenomes since e.g., phoQ and yfbI also have rather limited host ranges (Gammaproteobacteria), yet were frequently observed (Table 2). By PCR-based techniques we have previously observed vacA,
stxl1, and stx2 in various soils and a freshwater biofilm (Søborg et al., 2013) demonstrating that these genes are present amongst non-human environmental microbiomes.

**Potential Ecological Functions of Virulence Gene Homologs**
The ecological function of the virulence gene homologs in environmental bacteria, and why certain homologs seem to be overrepresented in some microbiomes, is uncertain. The ubiquitous presence of the genes suggests they carry important functions, and it has been proposed that they are housekeeping genes essential for survival and proliferation in outer environments (Søborg et al., 2013) for instance by enabling the bacteria to cope with their physico-chemical environment, avoid predation and/or interact with host organisms. Bacterial DNA isolation procedures usually involve a size fractionation step of the environmental sample prior to DNA extraction. Thus, with the exception of the termite gut, the gutless worm and the Alvinella pompejana metagenomes it is not likely that DNA of bacteria associated with large multicellular organisms is represented in the metagenomes. This could suggest that the primary function of the virulence gene homologs is unrelated to interactions with multicellular host organisms. However, since free-living and host-associated subpopulations may coexist, one cannot rule out the possibility that the identified virulence gene homologs also play a role in host-pathogen interactions in the outer environment.

To be important for the survival of the bacteria in the environment the virulence gene homologs must be both functional and expressed in the non-host environments. Full-length or almost full-length (>95%) sequences of seven genes (papA, papC, focG, invA, phoP, rfbE, and wbdI) were observed by single TBLASTN hits. In the case of the wbdI gene in the Washington Lake metagenome, 98% of the sequence, including the GDP-mannose binding site, was covered by a single TBLASTN hit (Supplementary Figure 1A). In some cases (e.g., phoP and wbdI) complete or almost complete sequences were found in several metagenomes. For the genes papH, spiA, phoQ, pqaB, yfbI and mecA, 65–80% of the complete gene sequence were covered by single TBLASTN hits in at least one metagenome (exemplified by pqaB; Supplementary Figure 1B). The fact that full-length genes were frequently observed suggests they are potentially functional and do not just occur as fragmented virulence genes within the microbiomes. As further support we have previously demonstrated *in vitro* expression of fimH in natural soil bacterial isolates and the expression of a virulent phenotype in soil bacterial isolates carrying homologs to the hemolysin gene, hblA, (Søborg et al., 2014). Some of the metagenomes (DayNight, Contaminated soil, MILOCO454, PacificOcean, and WesternChannel) were metatranscriptomes. Hence, the 15 different virulence gene homologs detected in these metatranscriptomes provide evidence of environmental transcription.

The BisonMetagenome represented samples collected along a thermal gradient (93–56°C) of a hot spring (Bison Pool) in Yellowstone National Park (Supplementary Table 1). Theoretically, the bacterial community in Bison Pool may have been contaminated by bacteria from livestock as bison frequenting the pool during winter months may potentially have been in contact with livestock outside the borders of the park. However, since samples were taken at water temperature between 93 and 56°C, the risk of contamination is miniscule. Hot spring environments are often characterized by dense biofilms, hence, a relative overrepresentation of adhesin genes (papC and focG) is intuitively understandable, as fimbriae on the outer surface of the bacteria may play an important ecological role in biofilm formation (Lasaro et al., 2009). The resistance genes pqaB and yfbI also appeared to be relatively overrepresented in BisonMetagenome. They encode L-Ara4N transferases that add L-Ara4N moieties to lipid A of the cell membrane of Gram-negative bacteria. The modification of lipid A interferes with the recognition by the Toll-like receptor TLR-4 on macrophages/neutrophils and promote resistance toward cationic antimicrobial peptides in phagosomes (Baker et al., 1999; Trent et al., 2001; Miller et al., 2005). Thus, an over-representation of these genes may be explained as an adaptation to resist phagocytosis by thermophilic amebae which in some cases have been shown to be abundant in hot water environments (Tyndall et al., 1989; Reeder et al., 2015). In a similar manner, spiA may play a role for the bacteria in their interaction with amebae. In *Salmonella*, spiA encodes a Type III secretion system outer membrane protein, SpiA, involved in the transport of virulence proteins into host cells (Cirillo et al., 1998). Therefore, SpiA, in combination with the L-Ara4N transferases encoded by pqaB and yfbI, may be involved in the infection (internalization) of amebae.

Homologs to papC, focG, pqaB, yfbI, and spiA were also overrepresented in the CellCapture metagenome based on samples of Pacific deep-sea sediment (Supplementary Table 1). A plausible explanation for an over-representation of adhesion genes in deep-sea sediment is not straightforward but the resistance genes and the Type III secretion gene may help the bacteria avoid predation by protozoa the same way as in hot springs.

The Washington Lake metagenome had above average frequencies of homologs to the regulatory gene phoP, the inflammatory gene rfbE and the resistance gene mecA. The two-component regulatory system phoP/Q, encoded by phoP and phoQ, governs adaptation to low Mg²⁺ and Ca²⁺ environments and to intramacrophage survival (Groisman, 2001; Tu et al., 2006), while rfbE encodes a perosamine synthetase involved in the synthesis of the O antigen of lipopolysaccharide in the outer membrane of Gram-negative bacteria (Iguchi et al., 2011). The relative over-representation of phoP and rfbE in the WashingtonLake metagenome suggests they provide specific selective advantages for the bacteria in these two environments. On the other hand, since they encode general house-keeping functions, their presence may not be exceptional. mecA encodes a peptidoglycan transpeptidase (with low affinity to β-lactam antibiotics) involved in cell wall synthesis (Wielders et al., 2002). Although mecA may be considered a basic house-keeping gene, the finding of environmental homologs is noteworthy. mecA is highly conserved among clinical methicillin-resistant strains of...
Staphylococcus aureus (MRSA) (>90% sequence identity between strains) and encodes a peptidoglycan transpeptidase unlike any of the peptidoglycan transpeptidases normally produced by S. aureus (<21% sequence identity) (Lim and Strynadka, 2002). It consists of two regions with homology to a penicillinase gene and a penicillin-binding protein gene, and has been hypothesized to have evolved by homologous recombination and transferred to MRSA by horizontal gene transfer from an unidentified species (Hiramatsu, 1995; Lim and Strynadka, 2002). Our data suggest that this unidentified bacterial species may have been of environmental origin.

The ecological role of the above average frequency of homologs to the toxin gene hlyA in the DeepMed seawater metagenome is somewhat perplexing. hlyA encodes a hemolysin (HlyA) involved in the lysis of (red blood) cells. However, marine opportunistic fish and shellfish pathogens such as Vibrios may produce hemolysins, which share common features with the HlyA hemolysin (Zhang and Austin, 2005; Ruwandeepika et al., 2010). Hence, the bacteria carrying homologs to hlyA may possibly be involved in the pathogenesis of marine animals.

Another potential role of hemolysin could be as an antipredator defense strategy to subvert phagocytosis similar to the group B streptococcal β-hemolysin/cytolysin encoded by the cylE gene (Liu et al., 2004).

**Phylogenetic Relationships between Virulence Genes and Environmental Homologs**

Similarities >50% between clinical and environmental sequences were observed for eleven genes (hlyA, papC, fimH, invA, phoP, phoQ, pqaB, rfbE, wbdI, yfbI and mecA) and for 6% of all hits. In four cases (fimH, phoP, phoQ and yfbI) were an identity of more than 90% observed between the environmental and clinical sequences (Supplementary Table 3). The results concur with a previous study in which sequence similarities of 90–100% between environmental and clinical virulence genes were found (Søborg et al., 2013), and taken together emphasize that sequence conservation between environmental and clinical genes may be high.

**FIGURE 4 |** Minimum evolution distance tree (A) and maximum parsimony tree (B) of environmental sequences homologous to the protein sequence of wbdI. Only environmental sequences with a coverage >95% (>141 aa) are included. The clinical sequence of wbdI is shown in red boldface type. Bootstrap values above 50% are displayed.
Phylogenetic analyses underlined the close relationship between clinical genes and environmental homologs, and demonstrated relatively large sequence variations between environmental homologs with no clear relationship between environment and phylogeny (Figures 4A, 5A and Supplementary Figure 2). For instance, the phylogenetic distance between the clinical sequence of WbdI and its most closely related environmental homolog (HypersalineMat metagenome), and the distance between YfbI and its most closely related environmental neighbor (GOS Sargasso Station 13 metagenome), were both 0.7 (Figures 4A, 5A). In contrast, phylogenetic distances up to 1.8 and 2.2 were observed between the environmental homologs to WbdI and YfbI, respectively. Furthermore, distances between the environmental sequences were larger than the largest distances between the clinical sequences and their environmental homologs (Figures 4A, 5A and Supplementary Figure 2).

Examination of the evolutionary history of the virulence genes indicated that the environmental homologs in some cases represented ancient lineages in the phylogeny of the clinical genes (Figures 4B, 5B and Supplementary Figure 2). In the case of YfbI, all environmental sequences were evolutionary basal to the clinical sequence (Figure 5B), while several of the environmental homologs to WbdI, PapA and HlyA showed deep branching positions relative to the clinical sequences (Figure 4B and Supplementary Figure 2). This suggests that the environmental genes in some cases evolved relatively early in the phylogeny of the clinical genes and that the origin of virulence genes in clinically relevant bacteria may be environmental.

CONCLUSION

Our findings highlight the widespread occurrence and transcription of genes often associated with bacterial human pathogens across diverse natural (non-human) environments. Environmental bacteria carrying homologs to virulence genes are probably not opportunistic human pathogens. Most likely, the homologs to the virulence factors are maintained because of the advantages they provide for the survival of the bacteria outside the host environment (Brown et al., 2012). Pathogenesis may, therefore, evolve by “coincidental” selection in non-host environments because traits that confer virulence in a host environment may also be beneficial for bacteria in non-host environments (Mikonranta et al., 2012). The high degree of sequence conservation between some of the environmental and clinical genes, and the phylogenetically deep branching positions of the environmental sequences, suggest that the environmental genes are precursors of virulence genes found in clinically relevant bacteria. Hence, our results suggest that coincidental selection in outer (non-host) environments may be a source of virulence traits for opportunistic pathogenic bacteria.

AUTHOR CONTRIBUTIONS

DS, NH, MK, and NK conceived and designed the research project and contributed with the interpretation of results.
ACKNOWLEDGMENTS

The work was supported by a grant [645-08-0110] from The Danish Agency for Science, Technology and Innovation to the Postgraduate Research School of Environmental Chemistry and Toxicology (RECETO). We wish to thank the staff of the CAMERA Portal for their assistance during the data analysis. In particular, we would like to express our gratitude to M. Chiu for his speedy response to all our many queries. G. Levin is thanked for his help with the preparation of the global richness map, and L. H. Hansen for his valuable comments on the manuscript. Finally, we wish to thank the reviewers for their constructive criticism that substantially improved the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2016.01712/full#supplementary-material


Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Søborg, Hendriksen, Kilian, Christensen and Kroer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.