Abstract

The bacterial phylum Acidobacteria has a widespread distribution, and is known to be one of the most common and diverse phyla in soil habitats. However, members of this phylum have often been recalcitrant to cultivation methods, hampering the study of this presumably important bacterial group. In this study, we used a cultivation-independent metagenomic approach to recover genomic information from soil-borne members of this phylum. A metagenomic fosmid library, constructed from former agricultural soil, was screened by PCR targeting acidobacterial 16S rRNA genes, facilitating the recovery of 17 inserts of putative acidobacterial origin. Recovered inserts appeared to originate from a range of Acidobacteria subdivisions, with a high proportion of inserts affiliated with subdivision 6 (10 clones). Upon full-length sequencing of fosmid inserts, gene annotation identified a total of 435 ORFs, representing a broad range of functions. Remarkably, six inserts from subdivision 6 contained a region of gene synteny, containing genes involved in purine de novo biosynthesis, a tRNA synthetase gene and a gene encoding a conserved hypothetical protein. Similar genomic regions had previously been observed in several environmental clones recovered from soil and marine sediments, facilitating comparisons with respect to gene organization and evolution. Comparative analyses revealed a general dichotomy between marine and terrestrial genes both in phylogeny and G + C content. Although the significance of this homologous gene cluster across subdivision 6 members is not known, it appears to be a common feature within a large percentage of all acidobacterial genomes in marine and terrestrial environments. The metagenomic approach utilized proved suitable for the recovery of acidobacterial genomic DNA, providing a window into the genomes of members of this important, yet poorly characterised, bacterial phylum.
Introduction

The genetic structure and functioning of the soil-borne microorganisms is still poorly understood in part due to the huge complexity of these communities. Molecular methods targeting 16S rRNA genes, have indeed revealed that soil microbial communities may be comprised of many thousands of different species in a single gram of soil (Torsvik & Ovreas, 2002). Moreover, the fraction of bacterial populations in soil that can be accessed through standard cultivation techniques is very low and might be less than 1% (Amann et al., 1995; Torsvik & Ovreas, 2002; Curtis & Sloan, 2004), limiting our ability to obtain functional information about many of the microbial populations that dominate soil. Although novel methodologies for culturing soil microbes that better mimic natural niche conditions have been successful in culturing many soil bacteria (Janssen et al., 2002), limitations remain with respect to biases and failure to detect essential microbial interactions and processes (Stevenson et al., 2004).

As phylogenetic inventories of soil microbial communities have accrued, it has become clear that numerous bacterial lineages appear to be quite common in the environment, but are underrepresented in bacterial culture collections (Handelsman, 2004). One of the best examples of such a lineage is that of the Acidobacteria. This phylum is one of the most commonly detected phyla in soil habitats by molecular methods, suggesting that members of this phylum may be important drivers of key ecosystem processes in terrestrial ecosystems. However, physiological information on isolates is essentially lacking, and only three complete genome sequences of isolates of this phylum have been determined to date - Solibacter usitatus (subdivision 3), Korebacter versatilis Ellin345 and Acidobacterium capsulatum (both subdivision 1) (Ward et al., 2009).

The designation, Acidobacteria, was coined after the first described species of this phylum, Acidobacterium capsulatum, which was isolated in the late 1990s from an acidic mine drainage (Ludwig et al., 1997). The phylum name is somewhat misleading since Acidobacteria are not restricted to acidic environments, as evidenced by their common detection in other habitats including ocean water, sediments, hot springs, peat bogs, and soil (Barns et al., 1999; Lopez-Garcia et al., 2003; Polymenakou et al., 2005; Chan et al., 2006; Dedysh et al., 2006; Penn et al., 2006; Sogin et al., 2006; Gomez-Alvarez et al., 2007). Members of the Acidobacteria seem to be particularly dominant in the soil, representing up to 52% of 16S rRNA gene sequences from clone libraries (Sait et al., 2002) and typically accounting for 20 to 30% of all bacterial 16S rRNA sequences amplified by PCR from soil DNA (Handelsman, 2004; Janssen, 2006).

By means of phylogenetic analyses, Acidobacteria were originally described as having 4 to 5 subdivisions (Kuske et al., 1997; Ludwig et al., 1997), with subsequent expansion to between 8 and 11 subdivisions (Hugenholtz et al., 1998; Zimmermann et al., 2005). With larger sequencing efforts, it has recently been proposed that the Acidobacteria may be composed of 26 or more subdivisions (Barns et al., 2007), the majority of which still lack culture representatives. Given their prevalence in soil environments, and the lack
of cultures for most subdivisions, the Acidobacteria have been proposed as ideal targets for study via large-insert metagenomic approaches (Kowalchuk et al., 2007).

The first description of acidobacterial genome fragments obtained from a metagenome library was provided by (Liles et al., 2003). By screening 24,400 bacterial artificial chromosome (BAC) clones from a soil metagenome library for 16S rRNA genes, these authors detected 12 clones putatively derived from members of the phylum Acidobacteria (9 clones from subdivision 6, 2 from subdivision 4, and 1 from 5). A single BAC insert (from subdivision 5) was selected for full-length sequencing, giving the first glimpse into the genome of an organism from this mysterious bacterial phylum. In the first study specifically targeting the Acidobacteria, Quaiser et al. (2003) recovered 6 acidobacterial fosmid clone fragments (4 out of 6 related to subdivision 6) from a sandy ecosystem, followed by the recovery of an additional 11 clones from deep sea environments (Quaiser et al., 2008). This allowed for a preliminary comparison of genomic organization in the chromosomal regions adjacent to rRNA operons, revealing a number of clones, affiliated with subdivision 6, that contained a co-linear region encoding genes involved in purine biosynthesis.

In this study, we sought to recover additional genomic information from Acidobacteria within soil environments with the goal of providing greater comparative insight into the function and evolution of this phylum. To this end, a large-insert fosmid library (28,800 clones) was constructed from high-molecular-weight DNA isolated from the rhizosphere of Festuca ovina in a former agricultural soil, and screened for inserts of acidobacterial origin. The resulting 17 positive clones were subsequently subjected to full-length DNA sequence analysis, followed by identification and annotation of open reading frames (ORFs). General genomic properties were determined for all sequenced inserts and compared across the Acidobacteria subdivisions represented in the library. A number of clones contained regions of synteny with previously described acidobacterial gene fragments, facilitating the phylogenetic analysis of genes contained within multiple clones and comparisons across soil versus marine habitats.

Materials and Methods
Field site and sampling
Soil samples were collected from a former arable field site located near Ede (52°04’ N, 5°45’ E), The Netherlands. An ecosystem restoration experiment was initiated on the field after the 1995 harvest season within the European project entitled “Changing Land Usage: Enhancement of biodiversity and ecosystem development” (Van der Putten et al., 2000; Bezemer & van der Putten, 2007).

Rhizosphere soil samples were collected in August 2005 from Festuca ovina L. plants from within plots sown with a monoculture of this species, adjacent to the main plant diversity treatment experiment (Kielak et al., 2008). For purposes of this study, rhizosphere was defined as combined sample of rhizosphere and root surface (rhizoplane), where pieces
of roots and adhering soil after shaking were collected. Soil samples were frozen at −20°C for approximately 2 weeks prior to DNA extraction.

**Metagenome library construction and screening**

For large-insert metagenomic library construction, high-molecular-weight DNA was extracted from 10 g of soil (wet weight) as previously described (van Elsas et al., 2008; Kielak et al., 2009; van Elsas et al., submitted). The construction of the metagenome library utilized the CopyControl Fosmid Library Production Kit (Epicentre Medison, WI, USA) and was preformed according to the manufacturer’s protocol. This protocol yielded 5′-phosphorylated blunt-end DNA, which was subsequently ligated into pCC1FOS vector and transformed into *E. coli* EPI300 (Epicentre Medison, WI, USA). The final metagenome library contained 28,800 chloramphenicol resistance clones selected on LB medium plates with addition of 12.5 µg/ml of chloramphenicol.

A total of 28,800 clones in the metagenome library were subjected to a PCR-based screening strategy using a semi-nested clone pooling strategy (Kielak et al., 2008; Kielak et al., 2009) and the primers pA/1378R (Edwards et al., 1989) and Ac31F/1378R (Barns et al., 1999). Briefly, clones were cultured overnight in 96-well plates, with the contents of 4 plates (384 clones) combined in a single plasmid extraction (QIAprep Spin Miniprep Kit; Qiagen, Venlo, The Netherlands). The resulting mixed template was used for PCR with the pAF/1378R followed by amplification with primers Ac31F/1378R as described above. In the cases where PCR product was detected, clone pooling was reduced to a single plate. Pooling was further reduced to a single row of a culture plate and eventually a single clone for subsequent reactions that produced positive product. After partial 16S rDNA sequencing and preliminary phylogenetic analysis, 17 clones affiliated to Acidobacteria phylum were selected for full-length sequencing. Sequencing was carried out at Macrogen (Macrogen, Seoul, Korea). Contig assembly was carried out using Lasergene's SeqMan (DNAstar). Remaining sequence gaps were closed by primer walking with sequence-derived oligonucleotides. A few gaps could not be closed due to the technical difficulties (see Table 1).
Table 1. General characteristics of acidobacterial fosmid clones.

<table>
<thead>
<tr>
<th>Description</th>
<th>246</th>
<th>162</th>
<th>282</th>
<th>89</th>
<th>148</th>
<th>164</th>
<th>253</th>
<th>59</th>
<th>66</th>
<th>70</th>
<th>92</th>
<th>98</th>
<th>126</th>
<th>213</th>
<th>259</th>
<th>270</th>
<th>293</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subdivision</td>
<td>*</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Length (kb)</td>
<td>32.61</td>
<td>36.44</td>
<td>29.31</td>
<td>31.62</td>
<td>30.78</td>
<td>33.66</td>
<td>32.88</td>
<td>37.26</td>
<td>29.90</td>
<td>33.76</td>
<td>33.28</td>
<td>30.79</td>
<td>28.09</td>
<td>35.44</td>
<td>31.19</td>
<td>38.49</td>
<td>36.48</td>
</tr>
<tr>
<td>Nr of gaps</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>GC%</td>
<td>57.7</td>
<td>58.7</td>
<td>55.5</td>
<td>59.3</td>
<td>70.3</td>
<td>54.4</td>
<td>53.8</td>
<td>64.5</td>
<td>63.4</td>
<td>62.7</td>
<td>62.7</td>
<td>65.5</td>
<td>66.3</td>
<td>65.2</td>
<td>65.9</td>
<td>65.0</td>
<td>69.2</td>
</tr>
<tr>
<td>GC% in ORF</td>
<td>57.4</td>
<td>60.8</td>
<td>59.9</td>
<td>71.1</td>
<td>55.2</td>
<td>53.6</td>
<td>65.9</td>
<td>64.8</td>
<td>63.9</td>
<td>63.8</td>
<td>67.6</td>
<td>67.9</td>
<td>67.1</td>
<td>68.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nr of ORF</td>
<td>30</td>
<td>31</td>
<td>26</td>
<td>25</td>
<td>29</td>
<td>22</td>
<td>25</td>
<td>20</td>
<td>26</td>
<td>22</td>
<td>25</td>
<td>18</td>
<td>21</td>
<td>21</td>
<td>27</td>
<td>30</td>
<td>33</td>
</tr>
<tr>
<td>Nr of non-</td>
<td>13</td>
<td>20</td>
<td>17</td>
<td>14</td>
<td>10</td>
<td>14</td>
<td>13</td>
<td>13</td>
<td>9</td>
<td>12</td>
<td>5</td>
<td>20</td>
<td>5</td>
<td>10</td>
<td>13</td>
<td>15</td>
<td>24</td>
</tr>
<tr>
<td>Nr of tRNAs</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

* Subdivision designation unclear

Length of recovered sequence without gap length estimation
Annotation and sequence properties

Open Reading Frames (ORFs) were assigned using the GLIMMER (Delcher et al., 1999) and FGENESB (http://linux1.softberry.com) software tools. Annotation of the identified ORFs was accomplished based on a similarity search against non-redundant protein databases using BLASTX and BLASTP algorithms with using an e-value threshold < 10^{-15}. ORFs shorter than 150 bp were discarded. tRNA genes were identified using the tRNAscan-SE program version 1.21 (Lowe & Eddy, 1997). ORFs were classified into functional categories as clusters of orthologous groups (COGs) by using COGNITOR (Tatusov et al., 2001). Searches for GC islands were performed using CpGFinder (SoftBerry, http://linux1.softberry.com). For identification of potential horizontal gene transfer regions, the method described by Tamanes & Moya (2008) was applied to examine tetranucleotide frequencies. For comparative analysis, full-length fosmid inserts sequences were compared using PolyDot (http://emboss.bioinformatics.nl/cgi-bin/emboss/polydot). This was followed by reciprocal BLASTN and TBLASTX searches between selected fosmids using the BLASTv2.2.18 package (ftp://ftp.ncbi.nih.gov; Altschul et al., 1990). Genomic fragment comparisons were visualized using the Artemis Comparison Tool ACTv.7 (Carver et al., 2005).

Phylogenetic analyses

All obtained 16S rRNA gene sequences were aligned using the ARB software package (Ludwig et al., 2004) and manually edited considering the secondary structure of the rRNA molecule. Gaps and ambiguously aligned positions were excluded from analysis. Two different methods were used for phylogenetic tree construction based upon SSU rRNA gene sequences; neighbor joining (NJ) and maximum likelihood (ML). MrModeltest2.2 (Nylander, 2004) was used to determine the most appropriate nucleotide substitution model. The likelihood ratio test, as well as the Akaike information criterion, identified the general time reversible model (GTR) with invariable sites and Γ-shaped distribution of substitution rates as the best fitting model. Both trees were constructed using the PHYLIP 3.67 package (Felsenstein, 2005) with the model parameters calculated by MrModeltest. The NJ tree was bootstrapped using 1,000 replicates, the ML tree using 100 replicates. For protein phylogeny reconstruction, the closest orthologues were identified with BLASTP from NCBI (http://www.ncbi.nlm.nih.gov/BLAST; Altschul et al., 1997). Phylogenetic reconstructions were performed with the PHYLIP 3.67 package (Felsenstein, 2005), using NJ and ML analysis.

Testing for potential sites of recombination

To help identify potential sites of recombination, a multiple nucleotide sequence alignment was made for the overlapping regions of fosmids 270, 59 and 98 and sequence accessions AY281356, AY281352, EF597689, EU686586, EF597694, EU686638, EU686603, EU686608 and EU795230 (Quaiser et al., 2003, Quaiser et al., 2008). This
alignment was examined using several different statistical algorithms for predicting recombination probability, namely RDP, GENECONV, MaxChi, Chimera and SiScan. Analyses were carried out in the RDP software package version 3.34 (Martin et al., 2005).

Results

General characteristics and annotation of acidobacterial genomic fragments recovered from a rhizosphere metagenomic library

In total, 17 putatively acidobacterial genomic fragments were detected in the library (28,800 fosmid clones) based upon PCR detection of 16S rRNA genes affiliated with this phylum. Insert sizes ranged from approximately 28 – 38 kb. The majority of clones appeared to be derived from members of Acidobacteria subdivision 6 (10 out of 17), and fragments from subdivisions 4, 1 and 3 were also recovered (3, 2 and 1 fosmid clones, respectively) (Table 1). The G + C content of inserts ranged from 54 – 70 %, and G + C contents were generally rather consistent within each subdivision. Subdivision 6 sequences had the highest average overall G + C content, 65.2 %, ranging between 62.7% in clones 70 and 92 to 69.2 in clone 293. The G + C contents observed in this study were comparable to those obtained previously from soil fosmids by Quaiser et al. (2003), where four subdivision 6 genome fragments averaged 66.4 % G + C and two subdivision 4 fragments 52.7 %. G + C contents of acidobacterial inserts, however, differed sharply from the values observed in fosmids recovered from aquatic environments (Quaiser et al., 2008), where nine subdivision 6 inserts averaged only 54.1 % G + C, and a fragment affiliated subdivision 3 had only 39.2 % G + C. Clone 148 of the current study displayed exceptionally high G + C content. At 70.3 % G + C, it was over 15% higher than the other two recovered subdivision 4 clones. As with all other inserts, there was no evidence for a genomic island of high G + C content within this insert, as G + C content was consistent across the full length of the insert, excluding the rRNA genes (data not shown). Analysis of tetranucleotide usage frequencies, anomalies in which are indicative of regions of potential horizontal gene transfer (Tamames & Moya, 2008), also displayed uniformity across the full length of clone 148, as well as other clone sequences examined (data not shown).

Full-length fosmid insert sequences were subjected to annotation revealing a range of predicted gene functions (Appendix Table A1-A17). A large fraction of predicted ORFs encoded housekeeping functions, and a large number was also assigned to hypothetical protein groups. The majority of best BLAST hits were with genes from one of the three available Acidobacteria genome sequences (Ward et al., 2009). However, a sizeable proportion (46 %) of ORFs showed best BLAST hits to genomes other than Acidobacteria. Of these, the greatest number had best scores with members of the Proteobacteria, especially the delta-subclass of this phylum (Figure 1).
Gene order and regions of synteny between fosmid clones

A number of fosmid inserts exhibited regions of synteny in gene order, and high levels of homology within the genes encoded in these regions. Six clones from subdivision 6 (clones 59, 70, 92, 98, 259, and 270) shared a syntenous region of variable length that included genes involved in purine de novo biosynthesis, a tRNA synthetase gene and a gene encoding a conserved hypothetical protein (Figure 2). Interestingly, similar regions of conserved gene order were observed for two soil fosmid clones (Quaiser et al., 2003) as well as seven fosmids recovered from aquatic environments (Quaiser et al., 2003). The overall gene organization and conservation was more extensive in the sequences obtained from the aquatic environment. Soil clones were more diverse in term of genomic fragment organization with single gene indel events mostly involving genes of unknown functions (hypothetical proteins). The degree of similarity of genome fragment organization generally followed the phylogenetic distances determined based on 16S rRNA gene sequence analysis (Figure 2).
Comparative analysis of acidobacterial genomic fragments affiliated with subdivision 6 (fosmid clones EF597689, EF597694, EU686603, EU686608, EU795230; Quaiser et al., 2008). A neighbor-joining tree of 16S rRNA gene sequences is shown on the left as reference. Sequences are aligned according to the ribosomal operon localization. 16S rRNA and 23S rRNA genes are indicated by grey arrows. Conserved regions between genomic fragments are indicated by grey shaded areas, grey intensity being a function of sequence similarity by tBLASTx over lengths greater than 50 bp. Particular ORFs mentioned in the text are highlighted in color.
A conserved syntenic region was also found between a clone affiliated with subdivision 4 (clone 253) and a clone from this subdivision previously recovered by Quaiser et al. (2003) (designated clone 46h5; AY281357S1&2). This region contained seven shared genes (putative 2-methylthioadenine synthetase, putative phosphatidylglycerophosphatase, magnesium and cobalt transport protein CorA, putative D-3-phosphoglycerate dehydrogenase and putative penicillin-resistance DD-carboxypeptidase) (Figure 3). No similarities were observed with the other two subdivision 4 genomic fragments (clones 148 and 164).

**Phylogenetic analysis of members of subdivision 6 based on 16S rRNA and genes shared between fosmid clones**

Given the large proportion of subdivision 6 sequences recovered within this and previous studies, we focused more detailed phylogenetic analyses on this subdivision (Figure 4). Phylogenetic analysis of the 16S rRNA gene sequences belonging to subdivision 6 revealed that sequences recovered from aquatic environments seemed to cluster rather tightly, with sequence similarities ranging from 0.911 to 0.999 (Appendix Table A18). Soil clones, on the other hand, were spread across the breadth of the subdivision, with sequence similarities ranging from 0.822 to 0.999. Thus, soil-borne Acidobacteria appear to be more diverse, with a total 16S rRNA gene diversity that is comparable to that observed within many phyla (Schloss & Handelsman, 2005).

The genes held in common between the different subdivision 6 clones containing regions of synteny could also be used for phylogenetic analysis, affording a comparison with the 16S rRNA gene-based phylogeny. The phylogenetic analysis of amino acid sequences of TyrS, whose gene was present in the greatest number of fosmid clones, revealed two highly distinct TyrS subclasses (Figure 5A). All the sequences recovered from deep-sea environments (designated with a “W” in Figure 5A) clustered tightly and were affiliated with the sequences derived from two soil clones, as well as the three fully Acidobacteria genomes. The remaining soil clones belonged to the other TyrS subclass, and clustered together, showing greatest similarity to the sequence available for Desulfococcus oleovorans (YP001528496).

In contrast, phylogenetic analysis of derived amino acid sequences of PurM (Figure 5B) and PurN (Figure 5C) were essentially congruent with the 16S rRNA gene-based phylogeny. In these cases, all sequences of acidobacterial origin were grouped together and distinct from the protein sequences obtained from other organisms. For PurM, a clear dichotomy was observed between sequences recovered from soil versus aquatic clones. The same was true for PurN, except for a single aquatic clone (designated KM4-9-H7; Quaiser et al., 2008), which grouped with the soil clones (Figure 5C).
**Analysis of potential recombination**

For detection of potential hotspots for recombination within the overlapping region of subdivision 6 clones, a composite alignment (12 clones, 8750 bp and using clone 270 as reference) was subjected to several tests designed to detect sequence regions for which recombination events may explain sequence patterns (Martin *et al.*, 2005). All predicted events with statistical significance of p<0.05 detected by one or more of the five algorithms used (RDP, GENECONV, MaxChi, Chimera and SiScan) are listed in the Appendix Table A19. A total of 56 potential sequence exchange tracks were detected within 63 breakpoints in the sequence alignment (Figure 6). Putative sequence exchange tracks were rather short, ranging in size from 9 to 294 bp. Most of the intervals (34) were detected only in one homologue whereas the most frequent one was found in 6 homologues (conserved hypothetical protein) (Appendix Table A19). Breakpoints were not evenly distributed across the examined region, with the highest number of breakpoints detected within PurM (22) followed by TyrS, PurN and two conserved hypothetical proteins (16, 12, 10 and 3, respectively; Figure 6). The various methods used yielded different results, but generally revealed the same patterns of hot and cold spots for potential recombination across the analyzed region, and only a single site was identified as significant by all algorithms used.

**Discussion**

**General features of recovered acidobacterial genomic fragments**

The metagenomic strategy employed in this study proved successful in the recovery of genomic fragments of acidobacterial origin, yielding approximately 563 kb of sequence information. The frequency of recovery of positive clones is within the range that would be expected given the proportion of *Acidobacteria* among the total community (18 – 26 %), number of rRNA operons per cell (1 - 2), and acidobacterial genome size (4,127 kb *A. capsulatum*; 9,966 kb *S. usitatus*). The majority of clones showed affiliation with subdivision 6. This prevalence of subdivision 6 is in agreement with previous studies targeting soil habitats (*Kuske et al.*, 1997; *Barns et al.*, 1999; *Janssen*, 2006; *Eichorst et al.*, 2007; *Hansel et al.*, 2008). It should be noted that the screening strategy used here to screen for target inserts, utilized an acidobacteria-specific primer, which has recently been demonstrated to miss some lineages within this phylum (*Barns et al.*, 2007; *Kielak et al.*, 2009). However, based upon general bacterial clone analyses of the soil used in this study (*Kielak et al.*, 2007), these lineages probably represent only a very small fraction of the total acidobacterial community.

Previous studies have suggested an evolutionary link between the *Acidobacteria* and the *Proteobacteria*, with the suggestion that the *Acidobacteria* may be a sister group of the *Deltaproteobacteria* (*Ciccarelli*, 2006). This link is understandable given the results of the best BLAST hits from the current study that did not match with available *Acidobacteria* genome sequences (Figure 1), but did match with *Gammaproteobacteria*. However, it is
still not clear if this is more due to the prevalence of available proteobacterial genome sequences in available databases as compared to the *Acidobacteria* than to a real link. This may be resolved as additional *Acidobacteria* genome sequences become available.

Although the strategy employed served to recover the largest amount of functional gene information yet obtained from *Acidobacteria* in the environment, the majority of genes recovered encoded mostly general housekeeping functions or hypothetical proteins that lack further characterization. Thus, without supporting physiological studies, which do not yet exist for *Acidobacteria*, it is difficult to make ecological inferences based upon the recovered metagenomic information alone. Clearly, future studies that couple bacterial identity and function, as well as culturing and physiological studies of *Acidobacteria*, will need to be expanded to define the niches of *Acidobacteria* in soil environments.

Figure 3. Comparative genomic organization of fosmids 253 (subdivision 4) (on top) in relation to previously described fosmid 46h5 (Quaiser *et al.*, 2003) (on the bottom). Conserved genomic regions between fosmid clones are indicated by grey shaded areas, grey intensity being a function of sequence similarity by tBLASTx over lengths greater than 150 bp. Particular ORFs mentioned in the text are highlighted by colors.
Figure 4. Neighbor-joining tree of subdivision 6 phylum of the Acidobacteria, highlighting the sequences recovered in this and two other metagenomic studies (Quaiser et al., 2003; Quaiser et al., 2008). Names in italics indicate sequences retrieved from GenBank. Green- this study and soil metagenome library (Quaiser et al., 2003); blue- deep sea metagenomic library (Quaiser et al., 2008). Numbers near nodes indicate bootstrap values. Sequences from subdivision 4 were used as outgroup.
Figure 5. Maximum likelihood phylogenetic analysis of A. Tyrosyl-tRNA synthetase (TyrS); B. phosphoribosylformylglycinamidine cycligase (PurM); C. trifunctional purine biosynthesis protein (PurN). Amino acid sequences obtained in this study are underlined. “W” behind the name indicates sequences obtained from deep-sea clones. “S” behind the name indicates the sequences obtained from soil clones. The analysis was performed respectively on 220, 262, 125 unambiguously aligned amino acid positions. Numbers near nodes indicate bootstrap values (only bootstrap \(\geq 60\)).

**Phylogeny of subdivision 6**

Given that the majority of metagenomic clones were affiliated with subdivision 6, we chose to focus special attention on the phylogenetic analysis of this subdivision. There is a large degree of sequence variation within this subdivision, based upon 16S rRNA gene sequences, and the diversity of sequences derived from soils appears to be much greater than that found for aquatic environments (Figure 4).

The recovery of regions of synteny (see below) between several clones within this and previous studies (Quaiser et al., 2003; Quaiser et al., 2008), also allowed for a more complete analysis of several protein-encoding genes for comparison with 16S rRNA gene-based phylogeny. The clustering of soil and aquatic sequences within the SSU gene-based tree is generally mirrored in the phylogenies of the protein-encoding genes that we examined, as well as comparisons across the full length of the inserts that exhibit regions of synteny (Figure 2). The phylogenetic reconstructions derived from PurM and PurN show the Acidobacteria as a monophyletic group, with a clear separation of environmental sequences based upon habitat (Figure 5B & 5C). Interestingly, TyrS displays a different pattern with two distinct groups of sequences, affiliated with different bacterial groups (Figure 5A). It remains to be investigated why these genes display such contrasting phylogenies.

**Comparison of environmental clones with regions of gene synteny**

Among the ten clones that were affiliated with subdivision 6, six contained a homologous region of variable length. Remarkably, this region of co-linearity is also present in two previously published soil metagenomic clones (Quaiser et al., 2003; Quaiser et al., 2008), as well as several metagenomic clones from deep sea samples (Quaiser et al., 2008). Thus, it appears that a large fraction of the total acidobacterial communities in diverse habitats share this feature. The region of synteny contains a Tyrosyl-tRNA transferase, a hypothetical protein, as well as genes involved in de novo purine synthesis. It is not known whether there is some special feature of these genes that has led to their conservation adjacent to rRNA operons within this phylum. It also remains to be seen if these Acidobacteria populations also share syntenous regions elsewhere on the chromosome.

A region of co-linearity was also identified for two fosmid inserts affiliated with subdivision 4, clone 253 from this study and clone 46h5 from Quaiser et al. (2003) (Figure 3). However, there are only five genomic fragments that have been recovered to date from
members of this subdivision (three from this study and two from Quaiser et al. (2003). Thus, it remains to be seen if observation represents a lucky coincidence, or a more widespread phenomenon, as seen for the syntenous region within subdivision 6 clones.

In comparison of gene order and the phylogeny of 16S rRNA and functional gene markers, it appears that there is clear separation between the sequences recovered from soil versus deep sea environments (Figure 2, 4, 5A, B & C). This distinction is also apparent at the level of general sequence properties such as G + C content. In comparing the G + C content of subdivision 6 clones across terrestrial versus marine habitats, a sharp dichotomy was apparent. On average, clones from soil habitats had a G + C content that was 14.6 % higher compared to clones recovered from marine environments. It has previously been suggested that environment factors can influence G + C content and potentially the specific amino-acid composition of proteins (Foerstner et al., 2005). Similar contrasts in average G + C have been observed in the comparison of metagenomic data obtained from shot-gun cloning and sequencing approaches. For instance, the G + C content of soil versus surface water metagenomic sequences differs by almost 30 % (61 % versus 34 %, respectively) (Venter et al., 2004; Foerstner et al., 2005; Tringe et al., 2005). Interestingly, this dichotomy across habitats appears to hold for organisms that are phylogenetically rather closely related. However, the reasons and mechanisms by which this occurs are not yet known. The effects of genome size (Hu et al., 2007) and responses to shifts in optimal growth temperature have been suggested as potential factors involved in shaping G + C content (Musto et al., 2004; Musto et al., 2006), although some evidence also speaks

Figure 6. Detection of potential recombination tracts within the identified syntenic region in an alignment from fosmid clones 270, 59, 98 (this study), AY281356, AY281352 (Quaiser et al., 2003), EF597689, EU686586, EF597694, EU686638, EU686603, EU686608 and EU795230 (Quaiser et al., 2008). Detected events are shown in relation to fosmid clone 270 within the corresponding 8750 bp region, containing 7 ORFs. Colours indicate the number of methods which detected the recombination events (see Appendix Table A19).
Comparative analysis of acidobacterial genomic fragments

against this latter explanation (Hurst & Merchant, 2001; Wang et al., 2006; Garcia et al., 2008).

Highly similar DNA stretches from closely related strains can undergo successful sequence exchange events, and, as a result, these changes can be vertically inherited (Majewski and Cohan, 1999). We therefore performed analyses to examine the potential course of recombination across homologous genes present in the identified syntenic regions. These analyses were consistent with the conclusion that homologous recombination has been involved in shaping the currently observed sequence patterns. Sequences exchanges were not exclusive to specific protein domains, although there were clearly hotspots across the length of the syntenous regions. While the analyses preformed here clearly do not compare the sequences that have directly recombined, they do suggest that the sequence patterns observed in these genomic regions are partially the result of past recombination events.

Conclusions

The metagenomic screening method employed in this study proved to be successful for the recovery of genomic information from diverse Acidobacteria directly from the environment. However, although a large amount of genomic information could be obtained, insights into ecology are still limited due to the lack of niche-defining genes and physiological knowledge for members of this phylum. Despite this limitation, the information recovered did allow for the most extensive phylogenetic and evolutionary study of Acidobacteria subdivision 6 to date. Also, the information gathered via the rRNA-based PCR screening method required a large number of clones to be screened and only provided insight into the genomic regions directly adjacent to rRNA operons. Additional genome sequencing efforts for a broad range of strains Acidobacteria will no doubt facilitate metagenomic efforts, as this should allow the identification of metagenomics clones of acidobacterial origin by end sequencing and binning. Also, novel high-throughput screening methods (Ingham et al., 2007) and single-cell sequencing approaches (Lasken, 2007; Ishoey et al., 2008) hold great promise in expanding our understanding of the genomic diversity of this phylum.

It is amazing that the observed region of gene synteny within subdivision 6 fosmid clones was found so broadly and frequently. The significance of this finding and the potential for other regions of homology across the chromosome of subdivision 6 members remains to be investigated. Given the widespread nature of strains exhibiting this feature, and the large proportion of Acidobacteria subdivision 6 in diverse ecosystems, this represents a clear research priority.
Acknowledgements

This work was funded by the BSIK program “Ecogenomics” (http://www.ecogenomics.nl/). Eiko E. Kuramae is acknowledged for her support and valuable discussions regarding the work presented in this manuscript.