Metagenomics-guided exploration of natural habitats for bacterial chitinases
Cretoiu, Mariana Silvia
Chapter VIII

Identification of novel chitinase genes by screening of a metagenomic library derived from chitin-amended soil

Mariana Silvia Cretoiu, Francesca Berini, Anna Maria Kielak, Flavia Marinelli, Jan Dirk van Elsas

Manuscript in preparation for publication.
Abstract

Metagenomes derived from the microbiota of soils are rich sources of novel genes coding for potentially highly active enzymes. Biasing soils towards particular functions, e.g. chitin degradation, has been recommended as a strategy to improve hit rates. Here, we report on the construction of a metagenomic library from a chitin-amended disease-suppressive soil, which was screened for novel chitin-active enzymes. The library constructed in fosmids in *Escherichia coli*, comprised 145,000 clones for a total of approximately 5.6 Gb of cloned soil DNA. We describe the isolation and characterization of several new putative bacterial chitinase genes from it, next to the genomic context of such genes. Briefly, library pools were used and repeated *chiA* gene PCR based screenings allowed the singling out of about five single fosmids, each containing 25-35 kb inserts, that produced promising amplicons. Subsequent sequencing of the whole fosmid inserts resulted in the identification of four putative chitinase genes and one chitin deacetylase gene. *De novo* annotation and comparative genomics of the whole fosmid regions revealed insights in the genetic environments of the respective putative *chiA* genes.
Introduction

Chitin and its derivatives are among the most promising natural biopolymers in use for biomedicine, agriculture and the pharmaceutical industry, particularly due to their biodegradability and non-toxicity. Enzymes active on chitin and chitin oligomers, here denoted as chitinases, are of great interest for use in large-scale applications. Two main areas have been described for chitinases, i.e. (1) the development of compounds that allow to antagonize chitin-containing phytopathogens and (2) the use as industrial biocatalysts in the production of chitin derivatives. Chitinases are glycoproteins with wide ranges of molecular weights (20-115 kDa), optimal temperature (18-90°C), pH (2-10.5), and pI (3.5-8) (http://www.brenda-enzymes.org/). Chitinases fall mainly in the glycoside hydrolase families 18 and 19. Remarkably, family-18 and -19 enzymes have different structures and modes of action. Moreover, based on the structure of the catalytic domain and on the position of the hydrolytic site on the polysaccharide chain, chitinases show either endo- or exoactivity (Henrissat & Davies, 2000; Van Scheltinga et al., 1994). Chitinases are prevalent in the microbiota of ecosystems, as it has recently been shown that the highest quantity of chitin is turned over by bacteria and fungi in marine and terrestrial ecosystems (Delpin & Goodman, 2009; Poulsen et al., 2008).

The taxonomic diversity of soil microorganisms involved in the degradation of chitin was previously investigated and the importance of diversity in the process in the N cycle was reported (Gooday, 1990a,b; Metcalfe et al., 2002a,b; Manucharova et al., 2007). Furthermore, chitin-degrading enzymes were shown to constitute important enzymes in bacteria-fungi competitive interactions for plant root exudates in the rhizosphere (Bonfante & Anca, 2009). Moreover, in the case of agricultural soils, the addition of chitin helps in enhancing suppressiveness against soil-borne pathogens by promoting active chitinolytic microbial communities (Cretoiu et al., 2013; Kielak et al., 2013; Korthals et al., 2010). Recent developments in metagenomic analysis of soil microbiota have enabled the access to novel genes and useful biomolecules (Ferrer et al., 2005; Nacke et al., 2011; Simon & Daniel, 2009). Furthermore, the use of a “substrate-enriched environment” method (also called ecological enhancement) was shown to increase the efficiency of mining for enzymes with improved features (Ekkers et al., 2012; Horn et al., 2012).

In the present study, we assessed the usefulness of a chitin-amended soil as a source for the recovery of novel genes encoding chitin-degrading enzymes. We report the construction of a large-insert metagenomic library in fosmids in an Escherichia coli host from an agricultural field soil amended with chitin. Subsequently, the library was subjected to PCR-based screenings for genes encoding relatives of the exo-chitinase chiA gene. Whole fosmid inserts were then sequenced using next generation sequencing (Illumina) technology. Five positive fosmid clones were retrieved and analyzed. One putative bacterial chitinase gene, with predicted novel features, was found and brought up for expression and protein characterization. Our results show that functional metagenomics analysis of soil enriched with chitin is an important approach for the discovery of novel enzymes from the
chitin metabolic pathway.

**Materials and methods**

**Soil samples**

Soil samples were collected from an experimental agricultural field amended with chitin located at the experimental farm “Vredepeel” in the south-east of the Netherlands. The field has been used since 1990 by Applied Plant Research (PPO) to test and monitor diverse agricultural practices. The soil was supplemented with 1.8% of shrimp waste chitin (20 tons/ha) calculated over the topsoil (20 cm). The soil, chitin-amended, was characterized as sandy, with pH 5.7±0.2 and 3.2% organic matter. Soil samples were homogenized by sieving (2 mm pore size mesh sieve) and 10 g of the homogenates were subsampled for DNA extraction.

**High molecular weight DNA extraction**

High molecular weight DNA extraction for the construction of the metagenomic library was performed using a modification of the protocol previously described by van Elsas et al (2008). Ten g of soil were suspended in 10 ml extraction buffer (100 mM Tris-HCl, 100 mM NaEDTA, 100 mM NaPO₄, 1.5% NaCl, 1% CTBA, pH 8.0), shortly vortexed and sonicated (water bath sonicator) for 15 minutes. After sonication, 100 µl of Proteinase K (10mg/ml) were added followed by 2 hours incubation at 37°C with gentle shaking (200 rpm). DNA extraction was performed with phenol / chloroform / isooamylnalcohol (25:24:1) at 60°C for 30 minutes (in a water bath). The metagenomic DNA was precipitated with 2-propanol and embedded in agarose plugs (1% low melting point agarose). Removal of phenolic contaminants and separation of 30-40 kb size DNA fractions was performed by Pulsed Field Gel Electrophoresis (PFGE) on 1% agarose gel supplemented in the upper part with 2% polyvinylpyrrolidone (PVP). Electrophoresis was run in 0.5X Tris-borate-EDTA (TBE) at 14°C using a PFGE DRIII System (BioRad) with the following parameters: gradient 6 V/cm, included angle 120°, initial switch time 0.5 s, final switch time 8.5 , linear ramping factor, 20 h. The 2 cm agarose fragments containing DNA in the range of 30-40 kb were cut out of gel, without staining and exposing to UV radiation. DNA was recovered using β-agarase (New England Biolabs) according to manufacturer’s protocol.

**Metagenomic library construction**

Construction of a metagenomic library was performed using the CopyControl Fosmid Library Production Kit (Epicentre, Madison, WI, USA). The metagenomic DNA was 5’- phosphorylated blunt-ended, subsequently ligated into pCC1Fos fosmid vector and transformed into Escherichia coli EPI300-T1R (Epicentre, Madison, WI, USA). The infected EPI300-T1R cells were plated on LB agar supplemented with 12.5 µg/ml chloramphenicol (positive selection of fosmid clones) and incubated overnight at 37°C.
The metagenomic library was stored as pools of “transformed EPI300 cells - amplified library” at -80°C according to cloning kit manufacturer recommendation’s.

**Chitinase A gene (chiA) PCR-based screening strategy**

The primers (GA1Fw and GA1Rev) and PCR conditions were previously described and optimized (Williamson et al., 2000). The PCR product, size between 450 and 600 bp according to a previous study of chitin-amended soil (Cretoiu et al., 2013), was determined by standard agarose gel electrophoresis. In order to confirm the presence of homologs of the chiA gene, the amplicons were extracted from gel using Wizard SV Gel and PCR CleanUp System (Promega, Madison, WI, USA) and directly sequenced using the reverse primer (LGC, Berlin, Germany). As a control, the EPI300-T1R and pCC1Fos DNAs were tested for the presence of chiA genes by PCR. Additionally, primer sequences were locally aligned (standalone BLAST v. 2.2.28+; Altschul et al., 1990) against the whole genome sequence of EPI300-T1R and pCC1 plasmid vector provided by the cloning kit manufacturer (Epicentre, Madison, WI, USA).

**Screening for positive clones**

The metagenomic clones were screened for their possession of chiA-like genes using a “pool-subpool-single” PCR strategy as described (Israel, 1993; Peterson et al. 2002). Clones, pooled, subpooled or single, were cultured overnight in LB broth supplemented with 12.5 µl/ml chloramphenicol in 96-well plates. The contents of 2 plates (192 clones) were combined for a single plasmid extraction using QIAprep Spin Miniprep kit (Qiagen, Venlo, The Netherlands). The resulting mixed DNA template was used for PCR reaction. The original clones were stored at 4°C and in case of a positive reaction subpools of rows of each plate were tested. Pooling after a single-plate positive result was reduced to a single clone level. At “single-row” and “single-clone” level the fosmid copy number was induced up to 50 copies by adding 0.4 µl of autoinduction solution (500X) per 200 µl LB broth (supplemented with 12.5 µl/ml chloramphenicol) according to the producer’s specifications (Epicentre, Madison, WI, USA).

All PCR products were checked for size and integrity by standard agarose gel electrophoresis. The presence of chiA-like sequences was confirmed by amplicon sequencing as described above. Amplicon sequences were assigned to chitinase genes by TBlast-X and aligned with a suite of 22 nucleic acid sequences of characterized chitinase (retrieved from GenBank and CAZy), by using Clustal-W (BioLinux7; Field et al., 2006). Phylogenetic reconstruction was based on Neighbor-joining tree method with bootstrapping (100 repetitions) and substitution model nucleic acid (MEGA 5.2; Tamura et al., 2011b).

**Fosmid DNA extraction from selected chiA-like gene containing clones**

PCR positive clones were individually cultured in 2 ml LB broth supplemented with 12.5 µl/ml chloramphenicol and the fosmid copy numbers were induced by adding 4
µl of autoinduction solution (500X) before incubating overnight at 37°C. Fosmid DNA was extracted using the Gene Jet Plasmid Midi Preparation Kit (ThermoScientific, St. Leon-Rot, Germany). DNA size and integrity were verified by PFGE using the electrophoresis conditions described above. DNA concentration was measured by using a spectrophotometer (Nanodrop; ThermoFisher Scientific, St. Leon-Rot, Germany).

**Sequencing of full-length fosmid insert DNA**

Full-length fosmid inserts (average size 30-40 kb) of selected clones were sequenced using a second (next) generation of sequencing methodology on an Illumina platform (BaseClear, Leiden, The Netherlands). Prior to sequencing, the concentration and quality (the integrity number) of DNA were assessed by microfluidics-based electrophoresis (Bioanalyzer, Agilent Technologies, Waldbronn, Germany). Paired-end libraries were prepared for each individual fosmid DNA using the Paired-End DNA Sample Preparation Kit and specific adaptors (Illumina, Eindhoven, The Netherlands). Sequencing was optimized and carried out on HiScanSQ Illumina system.

**Assembly and sequence analysis**

Raw data processing was supported by the sequencing company (BaseClear, Leiden, The Netherlands) and implied the generation of FASTQ sequence reads, quality control and *de novo* assembly, yielding consolidated contigs. Briefly, FASTQ sequence reads were generated using the Illumina Casava pipeline version 1.8.2. Initial quality assessment was based on data passing the Illumina Chastity filtering. Subsequently, reads containing adapters and/or PhiX control signal were removed using an *in house* filtering protocol. The second quality assessment was based on the remaining reads using the FASTQC quality control tool version 0.10.0. The quality of the FASTQ sequences was enhanced by trimming off low-quality bases using “Trim sequences” option of the CLC Genomics Workbench v. 5.5.1 (CLC Bio, Aarhus, Denmark). The quality filtered sequence reads were further filtered by removing sequence reads that could align to the backbone vector sequences. Remaining sequences were used to generate the contiguous sequences using “*De novo assembly*” option of the CLC Genomic Workbench v. 5.5. The contigs were retained if the average coverage was higher than 500X. The coverage was calculated from read mapping against the contig sequences. The final contigs were considered to be the representative of the whole insert of each individual fosmid.

**De novo annotation of genome fragments**

The final contigs obtained after *de novo* assembly were confirmed for the presence of *chiA*-like gene sequences obtained during the screening. For each individual fosmid, the sequence of the amplicon obtained was aligned against the full-length contigs using the BioEdit v. 7.2.0 sequence alignment editor (Hall, 1999) and the alignment was manually checked for errors and gaps. Open reading frames (ORFs) were assigned and verified in
three ways. First assignment was using GLIMMER v.3.02 (Delcher et al., 1999; http://ceeb.umd.edu/software/glimmer/) on a BioLinux v.7 platform (Field et al., 2006). A second assignment of ORFs was by using MetaGene software (Noguchi et al., 2006). The third assignment of ORFs was performed during the automatic annotation of coding sequences using the Rapid Annotation Subsystems Technology (RAST server) provided by the National Microbial Pathogen Data Resource (NMPDR) (Aziz et al., 2008). The obtained ORFs were compared between the predictive tools.

The annotation of each ORF obtained from RAST (namely protein-encoding genes; further referred as coding sequences, CDS) was manually curated and completed by a similarity search against the non-redundant protein (http://www.ncbi.nlm.nih.gov) databases using BLASTP (http://blast.ncbi.nlm.nih.gov/Blastp). BLASTP algorithm parameters (Table 1A) were optimized according to BLAST Program Selection Guide (http://www.ncbi.nlm.nih.gov/guide/training-tutorials/ BLAST tutorials and guides/). The closest protein homolog was assessed based on the combined criteria previously described (Rost, 1999; Raghava & Barton, 2006): query coverage (%), maximum identity (%), alignment scores (maximum and total score) and e-value (Table 1B). RAST annotation also included a scan for tRNA genes and classification according to the “Cluster of Orthologous Groups” of Protein (COGs). ORFs shorter than 120 bp were discarded when the query coverage and maximum identity criteria were not in the established range. Spacers were subsequently searched against the non-redundant database (http://www.ncbi.nlm.nih.gov) using all BLAST options to ensure that no ORF was missed. Start and stop codons were identified for all annotated ORFs.

Prediction of putative chitinase genes

Prediction of putative chitinase /glycosyl hydrolase family 18 functions was performed using the InterProScan (EMBL) integrative tool for search of similarities within all available functional annotated protein databases (sequences of proteins, protein superfamilies and hidden Markov models) (Quevillon et al., 2005). Furthermore, the secondary and tertiary structure of protein was predicted on ITASSER server (Roy et al. 2010, http://zhanglab.ccmb.med.umich.edu/I-TASSER/) using default parameters. Bacterial gene promoters were predicted using BProm (SoftBerry, http://linux1.softberry.com/berry). Ribosomal binding sites (RBS) were identified using the web version of RBS Calculator (https://salis.psu.edu/software/; Salis et al., 2009) and manually checked according to accepted models (Shultzaberger et al., 2001, Stewart et al., 1998). The taxonomical affiliation of genes annotated as chitinase and glycosyl hydrolases family 18 was confirmed by comparison with the complete Carbohydrate Sequence Database available (CAZy) using Mothra.ornl (CAZymes Analysis Toolkit, Park et al., 2010).
Table 1A: Functional annotation criteria. BLASTP settings according to protein size (amino acid residues).

<table>
<thead>
<tr>
<th>Blast settings</th>
<th>Query size (amino acids residues)</th>
<th>&gt;85 residues</th>
<th>50-85 residues</th>
<th>35-50 residues</th>
<th>&lt;35 residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight matrix</td>
<td>BLOSUM*62</td>
<td>BLOSUM 80</td>
<td>PAM*70</td>
<td>PAM30</td>
<td></td>
</tr>
<tr>
<td>Gap cost</td>
<td>11.1</td>
<td>10.1</td>
<td>10.1</td>
<td>9.1</td>
<td></td>
</tr>
<tr>
<td>Low complexity filter</td>
<td>on</td>
<td>off</td>
<td>off</td>
<td>off</td>
<td></td>
</tr>
<tr>
<td>E value</td>
<td>10</td>
<td>10</td>
<td>1000</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>Word size</td>
<td>3</td>
<td>3</td>
<td>3 or 2</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

*BLOSUM= Bloks Substitution Matrix. Calculates sequence comparison with less than 62% (default) or 80% identity.

*PAM=Point Accepted Mutation. Calculates sequence comparison based closely related proteins.

Table 1B: Functional annotation criteria. BLASTP criteria used to validate the affiliation of ORFs to proteins.

<table>
<thead>
<tr>
<th>Query coverage (%)</th>
<th>Maximum identity (%)</th>
<th>Alignment scores (total &amp; maximum)</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>20-100%</td>
<td>best hit</td>
<td>&lt;10^-15</td>
</tr>
<tr>
<td>50%</td>
<td>48-100%</td>
<td>best hit</td>
<td>&lt;10^-10</td>
</tr>
<tr>
<td>30%</td>
<td>75-100%</td>
<td>best hit</td>
<td>&lt;10^-5</td>
</tr>
</tbody>
</table>

Phylogenetic analysis of chitinase-like genes

All recovered chitinases and glycosyl hydrolases (family 18) were aligned using Clustal-W (BioLinux v. 7; Field et al., 2006) along with 65 sequences of characterized chitinases extracted from CAZy database. Phylogenetic reconstruction was performed by Maximum Likelihood (ML) analysis using the amino acid type substitution model of Jones-Taylor-Thorton (JTT) with uniform rates, partial deletion and site coverage cut-off of 95%. The tree was bootstrapped using 100 replicates. A characterized cellulose of Escherichia coli P12b was used as an outgroup sequence.

Prediction of the origin of fosmid inserts, comparative genomics and identification of potential horizontal gene transfer

The origin of the inserts was predicted based on the phylogenetic affiliation of more than 75% of genes identified. We interpreted these with the cautionary note in mind that de novo annotation is relative at the level of bacterial genomes in soil (Raes et al., 2007) and that the size of insert represented less than 1% of an average annotated genome.

Synteny of the recovered fosmid inserts with regions of existing genomes and
inter-gene regions similarities was determined using the multiple genome alignment progressive Mauve software (Darling et al., 2010). Searches for G+C-rich islands were performed using CpGFinder (SoftBerry; http://linux1.softberry.com/berry). Nucleotide frequency analysis of nucleotide sequences was per-formed for screening of potential horizontal gene transfer regions (Scater Plot Viewer; http://www.jcvi.org/cms/research/).

Results

Construction of a metagenomic fosmid library from chitin-treated soil

Using 10 g of chitin-amended soil, we produced 0.250 µg HMW total microbial community DNA per g soil, with an average fragment size of 40 kb. The DNA was found to be pure enough to serve for direct cloning into the fosmid system. Following the cloning and plating steps, a total of 145,000 Escherichia coli fosmid clones was generated. These were pooled in pools that each contained 1,500 individual fosmids. The estimated size of the library was 5.8 GB, which is comparable to the large soil metagenomic libraries that were previously reported (Nacke et al., 2011). The percentage of insert-carrying clones was estimated at 100%.

Screening for chiA-related genes

All fosmid pools were successfully screened by chiA based PCR. In total, 18 of these gave positive PCR results twice. The resulting amplicons, of 450-600 bp in size, were then cloned and subjected to sequence analysis, after which the sequences were compared to existing chiA sequences by database comparisons. The analysis showed that 13 sequences were very remote from the canonical chiA sequences, having < 35% homology. Thus, we focused on the remaining five predicted/detectable chiA sequences, with homologies > 35%. The thus identified fosmid pools were then subjected to several cycles of splitting up in subpools and PCR detection of the chiA gene, until final fosmids that generated the signal could be isolated. The chiA amplicon of individual clone was confirmed as chitinase by sequencing (Figure 1). These (five) fosmids were subjected to further characterization.

De novo annotation and general characteristics of genetic fragments recovered from fosmids of the chitin-amended soil metagenomic library

Five clones - denoted 14A, 22G3, 28C5, 53D1 and 101F8- were retrieved from the chitin-amended soil metagenomic library based on positive amplification of the chiA gene. All clones were confirmed to contain a chiA type gene by amplicon sequencing Phylogenetic analysis of chiA amplicon sequencing showed a strong support for the affiliation of genes to bacterial GH18/chitinase (Figure 1). Furthermore, the sequences were sufficiently different from known chitinases to warrant further investigations. Consequently, the inserts of all five fosmids were subjected to full-length sequencing using Illumina paired-end technology. The number of paired reads yielded total sequence information per fosmid.
amounting to 6.4 to 9.6 Mb, and coverage was, thus, higher than 100 (Table 2A). The sizes of the assembled inserts were found to vary from 21.2 to 39.7 kb. The G+C contents of the fosmids was, on average, 58.8±6.4%. This was specific per fosmid, and, for each fosmid, consistent across the full length of insert (Table 2B).

Figure 1. Neighbor-joining tree of chiA amplicon sequences. Reference sequences represents nucleic acid sequences of characterize chitinases retrieved from GenBank and CAZy. Number near nodes indicate bootstrap values (only values ≥25).
Table 2. General characteristics of genetic fragments recovered from chitin-amended soil fosmid library (A) Sequence information (B) Fosmids length, gaps, GC%, ORFs.

<table>
<thead>
<tr>
<th>A</th>
<th>Consensus length (bp)</th>
<th>Total read count</th>
<th>Single reads</th>
<th>Reads in pairs</th>
<th>Average coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>14A</td>
<td>22632</td>
<td>7547953</td>
<td>342043</td>
<td>7205910</td>
<td>5645.37</td>
</tr>
<tr>
<td>22G3</td>
<td>21255</td>
<td>8725007</td>
<td>436681</td>
<td>8288326</td>
<td>6428</td>
</tr>
<tr>
<td>28C5</td>
<td>31983</td>
<td>6426283</td>
<td>296701</td>
<td>6129582</td>
<td>47985.18</td>
</tr>
<tr>
<td>53D1</td>
<td>35473</td>
<td>9639420</td>
<td>29260</td>
<td>9610160</td>
<td>13743</td>
</tr>
<tr>
<td>101F8</td>
<td>37907</td>
<td>7863506</td>
<td>14887</td>
<td>7535920</td>
<td>1367.57</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B</th>
<th>Length (bp)</th>
<th>Gaps</th>
<th>GC%</th>
<th>ORFs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Non-hypothetical</td>
<td>Hypothetical</td>
<td>Unknown</td>
</tr>
<tr>
<td>14A</td>
<td>22632</td>
<td>3</td>
<td>52.7</td>
<td>18</td>
</tr>
<tr>
<td>22G3</td>
<td>21255</td>
<td>1</td>
<td>58.8</td>
<td>19</td>
</tr>
<tr>
<td>28C5</td>
<td>31983</td>
<td>0</td>
<td>65.2</td>
<td>35</td>
</tr>
<tr>
<td>53D1</td>
<td>35473</td>
<td>1</td>
<td>54.6</td>
<td>34</td>
</tr>
<tr>
<td>101F8</td>
<td>37907</td>
<td>0</td>
<td>59.5</td>
<td>43</td>
</tr>
</tbody>
</table>

Tetrnucleotide counts varied from 188 (fosmid 22G3) to 355 (fosmid 101F8). Comparisons of nucleotide frequencies indicated potential regions of horizontal gene transfer and similarities in sequence composition between fosmids 22G3, 28C5 and 53D1 (Figure 2).

ORFs such as N-acetyl-glucosamine transport, sugar ABC transporter, molecular chaperone, transcriptional regulator and carbohydrate metabolism were consistently predicted by all three annotation methods used (Figure 3).
Figure 2. Nucleotide frequency matrix along the whole insert fosmid sequences. Similarity (%) of overlap dotpoints indicated in regions identified as significant similar.

Figure 3. ORF orientation and position of selected genes.
All inserts were confirmed to have a bacterial origin by using BLAST analyses. Furthermore, no tRNA or rRNA genes were identified. Across the fosmids, annotation of the putative protein-encoding sequences revealed a range of predicted gene functions (Supplementary Tables), being a large fraction of predicted genes involved in housekeeping and cell replication functions. The fraction of genes denoted as hypothetical proteins varied from 55 to 76%. All predicted protein types and their distributions among the fosmids are presented in Table 3. The overall characterization of the fosmid inserts with respect to genes predicted to be chitin-active was based on several criteria. Thus genes related to chitin degradation, general carbohydrate metabolism, transmembrane ABC transporters of N-acetyl-glucosamine and other sugar-like molecules, transcriptional regulators and chaperonines were taken into account. The number of putative genes per fosmid belonging to the class “chitinases and general carbohydrate transport/capture and metabolism” ranged from 1 (fosmids 22G3, 28C5 and 53D1) and 2 (fosmid 14A) to 5 (fosmid 101F8).

Table 3. Distribution of selected proteins and groups of cellular functions among fosmids.

<table>
<thead>
<tr>
<th>Fosmid</th>
<th>Chitinases</th>
<th>Chitin deacetylases</th>
<th>Carbohydrate metabolism</th>
<th>ABC sugar transporters</th>
<th>Transcriptional regulator</th>
<th>DNA repair</th>
<th>Non-enzymatic peptidases</th>
<th>Non-ribosomal peptides</th>
<th>Anaerobic metabolism</th>
<th>Antibiotic resistance</th>
<th>Degradation of organic compounds</th>
<th>Fosmid-partitioning protein</th>
<th>Transposase/integrase</th>
<th>Hypothetical</th>
</tr>
</thead>
<tbody>
<tr>
<td>14A</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>22G3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>28C5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>53D1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>101F8</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>11</td>
</tr>
</tbody>
</table>

*a= other than chitinases & chitinase-like; b= other than carbohydrate hydrolases; c= N-acetyl-D-glucosamine ABC transport system & ABC-type sugar transport component

**Fosmid annotation and closest homolog prediction**

**Fosmid 14A**

Eighteen ORFs were predicted to exist in the insert in fosmid 14A (Table S14A). The G+C content was 52.7% and three regions with no ORF were identified. The majority of ORFs presented a positive transcription frame (Figure 3-14A). The gene length varied from 188 bp (CDS8, hypothetical protein) to 2,393 bp (CDS7, closest hit beta-D-galactosidase CAZy glycoside hydrolase family 2). One CDS (CDS 10, 1,697 bp) was annotated as a putative endochitinase gene with best BLAST hit (99% similarity, 99% coverage) to a recently described *Kitasatospora setae* chitinase (Ichikawa et al., 2010). The number of genes affiliated to carbohydrate metabolism and sugar ABC transporters was low (Table 3). Surprisingly, 33% of CDSs (so 6 genes) were assigned to proteins involved in plasmid partitioning next to a phage-type integrase. Another 33% of the putative CDSs remained hypothetical. Half of CDS were affiliated to sequences from a *Burkholderia*-like source organism, at a protein level of similarity between 27 (CDS6, ABC transporter) and
82% (CDS14, hypothetical protein) (Appendix table fosmid 14A).

**Fosmid 22G3**

Twenty putative ORFs and one gap were identified in fosmid 22G3. The G+C content was 58.8%. Only one ORF had a negative transcription frame (Figure 3-22G3; Appendix table fosmid 22G3). The size of CDSs was, on average, large, with 63% of these having more than 500 bp. One typical chitinase gene was retrieved, which revealed 100% identity and coverage with chitinase A of *Acidobacterium capsulatum* ATCC51196. One CDS, for N-acetyl-glucosamine transport (similarity 51%, coverage 86%) was also affiliated to a homolog from *A. capsulatum* ATCC51196, whereas one transcriptional regulator (similarity 33%, coverage 78%) similar to a region from *Granullicella tundricola* MP5ACTX9 was found downstream of the chitinase gene. No chaperonins and sugar ABC transporter genes were found (Table 3). The majority (55%) of putative CDSs were affiliated to an *Acidobacterium*-like organism.

**Fosmid 28C5**

A contiguous sequence of 35 ORFs with positive transcription frame was assigned for fosmid 28C5 (Figure 3-28C5; Appendix table fosmid 28C5). The fosmid insert G+C content was 65.5%, which was the highest among the inserts. The sizes of the CDSs was between 143 (CDS9-hypothetical protein) and 2,309 bp (CDS34- transcriptional regulator). One gene (CDS24, 1,190 bp) was annotated as a chitinase with best BLAST hit to a gene from *Stenotrophomonas maltophilia* AU12-09 (45% similarity, 88 % coverage). One N-acetyl-glucosamine ABC transporter with a best hit (25% similarity, 85% coverage) to *Streptomyces bingchenggensis* BCW-1 was identified downstream of the chitinase gene. Other putative CDSs, such as transcriptional regulators and chaperonin GroEL, were affiliated at moderate similarity level (average 38%) and high coverage (98%) to sequences from *Chloroflexi*-like organisms. One duplicate gene was assigned to an organophosphate pesticide hydrolase (similarity 44%, coverage 76%) of *Pseudomonas* sp. Ag1. Overall, the analysis suggested a broad range of putative source organisms such as *Chloroflexi, Actinobacteria, Firmicutes* and *Cyanobacteria*.

**Fosmid 53D1**

A total of 34 ORFs with positive and negative transcription frames and one gap were identified in the fosmid 53D1 insert sequence (Figure 3-53D1, Appendix table fosmid 53D1). The overall G+C content was 54.6%. The minimum gene size was 143 bp (CDS6, transposase IS66) and the maximum 2,309 bp (CDS31, putative protein kinase-transcriptional regulator). Hypothetical protein represented 32% of 53D1 sequence. One chitinase gene (CDS20, 1,190 bp) was identified. The best BLAST hit of this gene was with “uncultured bacterium” (48% similarity, 94% coverage), followed by *Stenotrophomonas maltophilia* AU12-09 (45% similarity, 87% coverage) and *Ktedonobacter racemifer* DSM
44963 (41% similarity, 93% coverage). Moreover, N-acetyl-glucosamine transporter and sugar ABC transporter genes were assigned to Kedonobacter racemifer DSM 44963 downstream of the chitinase gene. Similarly, for the CDSs corresponding to transcriptional regulators and to hypothetical proteins flanking the chitinase gene, a Chloroflexi-like source organism was predicted (Table S53D1). Overall, of the CDS annotated as Chloroflexi-associated genes, 35% had as close homologs genes from the recently described Nitrolancetus holandicus (Sorokin et al., 2012).

**Fosmid 101F8**

Fosmid 101F8 yielded the longest contiguous insert sequence of 37,907 bp. In total 43 ORFs were found, with positive and negative transcription frames (Figure 3-101F8; Appendix table fosmid 101F8). The gene sizes varied from 230 (CDS4, hypothetical protein) to 1,580 bp (CDS33, putative sensory transduction protein). The G+C content was 59.6%. Two different putative chitinase genes and one putative chitin deacetylase gene were found. CDS3 (1,121 bp), with best BLAST hit (100% identity and coverage) to a gene region from the Niastella koreensis GR20-10 genome and CDS25 (404 bp) best BLAST hit (77% similarity, 82% coverage) to Streptomyces avermitilis MA-4680 were annotated as chitinase genes. CDS5 (869 bp), with best BLAST hit (100% identity and coverage) to a region from Thermodesulfatator indicus DSM15286 was assigned as belonging to a polysaccharide deacetylase protein family. The fosmid 101F8 sequence also contained the functional groups considered for inter-fosmid comparison (chitinase, transcriptional regulator, N-acetyl-glucosamine and sugar ABC transporters and carbohydrate metabolism). With the exception of the chitinases, putative deacetylase and one antiporter protein (CDS26), all CDSs were affiliated, with high similarity and coverage value, to Aeromonas genes. At the level of species, 72% of CDS were similar with Aeromonas veronii.

**Genes and regions of similarity between fosmids**

Although the genomic organization was unique for each fosmid insert, a number of common features were identified between them. Fosmids 22G3 and 101F8 revealed the presence of a chitinase gene close to the 5’ end of the insert. All fosmids contained transcriptional regulators (of the LuxR, LitR, LysR types) and sugar ABC transporter genes. Fosmids 22G3, 28C5 and 53D1 presented, downstream to the Gh18/chitinase, one N-acetyl-glucosamine transporter gene. Nucleotide frequency analyses indicated the presence of overlapping tetranucleotides between fosmids 28C5 and 53D1 (40.3% similarity), 28C5 and 22G3 (12.07% similarity) and 14A and 22G3 (7.26%). Progressive Mauve alignment of the ORF nucleic acid sequences showed 13 regions of significant similarity between fosmids 28C5 and 53D1 (Figure 4). Fosmids 14A and 101F8 presented lower similarity when compared within the group of fosmids and they were considered to have unique sequences.
Putative chitinase genes and selection of candidate genes for expression analyses

All fosmid inserts identified primarily via amplicon chiA sequencing were confirmed as being part of chitinase-like gene complexes after complete annotation. Phylogenetic analysis of the putative chitinases (predicted protein sequences) showed high-similarity clustering with reference chitinases of the ChiA class and distant from the outgroup sequence of *E. coli* P12b cellulase (Figure 5).

In the light of the annotation of the chitinase of fosmid 53D1 (as affiliated with a sequence from an uncultured bacterium, followed by best hits to regions of *Ketodonobacter racemifer* DSM 44963 and *Stenotrophomonas maltophilia* AU12-09), we focused more detailed analyses on this sequence. The 53D1 chitinase ORF was retrieved together with 200 nucleotides located upstream to the identified codon start. The essential genetic elements (promotor, RNA polymerase interaction site, Shine-Dalgarno sequence, start and stop codons) necessary for expression in a heterologous recombination system were identified (Figure 6). The sequence of chitinase 53D1 falls into the 14% of bacteria with an unusual promotor region and codon start (Tikole & Sankararamakrishnan, 2006; Nakamoto, 2009). The “–35…–10” region revealed the atypical sequence “ATGACT…CGGGAT”, while the Shine-Dalgarno sequence was the universal AGGA. The start codon was GTG (encoding Val). Overall, the rare reported transcriptional elements suggested that chitinase 53D1 had a weak promotor.
Figure 5. Maximum likelihood phylogenetic analysis of chitinase protein sequences obtained in this study (marked) and 60 sequences of representative chitinases retrieved from CAZy. Substitution model Jone-Taylor-Thorton, uniform rates, partial deletion and site coverage cutoff 95%. Bootstrap values >25 indicated.
Five protein models were used to predict structure, based on both global and local structural comparisons with all available proteins from the Protein Data Bank (PDB) by using the ITASSER platform. Based on the comparison with proteins of similar folds (from the PDB) and multiple alignments, an iterative protein template is generated. This protein template is compared with 3D models available from characterized proteins and confidence and similarity scores are assigned for each comparison. The best prediction for the fosmid 53D1 putative chitinase was with chitinase A1 of *Bacillus circulans* WL-12, with a confidence score (C-score) of 0.79 and a structural similarity score (TM-score) of 0.82±0.08 (protein alignment presented in Figure 7). Considering the accepted scores for active enzymes (C-score: -5 - 2; TM: >0.5) the model obtained for the 53D1 sequence predicted an active protein. Moreover, identification of the family-18 glycosyde hydrolase consensus sequence DGIDIDWE confirmed the existence of a putative conserved active site within catalytic domain (Figure 6). This sequence was further offered for gene expression and protein characterization studies (work in progress).
Figure 7. Protein alignment of putative chitinase 53D1 and *Bacillus circulans* chitinase A1, as generated with ITASSER global and local alignment tool after screening of RCBS Protein Data Bank.

**Concluding remarks**

Assuming the prevalence of prokaryotic organisms in soil and their average genome size estimated to be about 5Mb (Hardeman & Sjoling, 2007), the metagenomic library produced from the chitin-amended soil represented approximately 1,200 prokaryotic genomes. Genetic screening of this metagenome from the chitin-amended soil had as main objective the identification of genes for new amino acid sequences affiliated to the functional group of chitin active enzymes. The screening strategy applied proved successful in recovering genomic fragments containing putative active chitinases. Five sequences of putative novel chitinases were found, next to a considerable number of sequences related to the metabolic pathway of carbohydrate degradation, cellular transport and excretion systems and regulation of transcription. Given the estimated proportion of chiA genes in soil bacteria (roughly 1-5%), the frequency of recovery of chiA positive clones was consistent with that in other reports (Wellington *et al.*, in preparation).

When the source organisms of the recovered genes for chitin-active proteins was considered, the diversity of the fosmid inserts was relatively high. All five identified fosmids had different predicted origins, being two of them rather close. The majority of the putative genes found revealed homologies to gene regions found in the genomes of organisms like *Burkholderia*, *Actinobacteria*, *Stenotrophomonas*, *Acidobacterium* and *Aeromonas*. Particularly fosmids 28C5 and 53D1 comprised genes which were similar to those recently described in *Nitrolancetus hollandicus* (Sorokin *et al.*, 2012) and in *Ktedonobacter racemifer*. Both organisms are members of the phylum *Chloroflexi*. The identification of such a putative chitinase gene in a chitin-enriched habitat can be hypothesized as another asset of the remarkable physiology of *Nitrolancetus hollandicus* like organisms. In fact, *N. hollandicus* has been described as the only known nitrite oxidizer that is not affiliated to the *Proteobacteria*. 
Acknowledgements
This project was funded by METAEXPLORE FP-7 awarded to J.D. van Elsas. We acknowledge Liz Wellington and Sara Sjoling for help in data analysis and selection of candidate genes for further applications.