

Identification of putative noncoding RNA genes in the *Burkholderia cenocepacia* J2315 genome

Tom Coenye¹, Pavel Drevinek², Eshwar Mahenthiralingam², Shiraz Ali Shah³, Ryan T. Gill⁴, Peter Vandamme¹ & David W. Ussery³

¹Laboratorium voor Microbiologie, Universiteit Gent, Gent, Belgium; ²Cardiff School of Biosciences, Cardiff University, Cardiff, UK; ³Center for Biological Sequence Analysis, Danish Technical University, Lyngby, Denmark; and ⁴Department of Chemical Engineering, University of Colorado, Boulder, CO, USA

Correspondence: Tom Coenye, Laboratorium voor Farmaceutische Microbiologie, Universiteit Gent, Harelbekestraat 72, 9000 Gent, Belgium. Tel.: +32 0 9 2648141; fax: +32 0 9 26458091; e-mail: tom.coenye@ugent.be

Received 7 June 2007; accepted 7 August 2007.
First published online October 2007.

DOI:10.1111/j.1574-6968.2007.00916.x

Editor: Craig Winstanley

Keywords

Burkholderia cenocepacia; noncoding RNA; microarray.

Introduction

Noncoding RNA (ncRNA) genes produce transcripts that function directly as structural or regulatory RNAs rather than expressing mRNAs that encode proteins (Eddy, 2001). Genes coding for these untranslated RNA molecules are present in the genomes of many different organisms, both prokaryotes and eukaryotes (including organelles such as the mitochondria and chloroplasts) (Eddy, 2001; Mattick, 2001, 2003; Hershberg *et al.*, 2003; McCutcheon & Eddy, 2003; Lung *et al.*, 2006). Most is known about ncRNAs in eukaryotes and it has been suggested that up to half or even three-quarters of the transcriptional output in higher organisms is ncRNA (Mattick, 2001). RNA-mediated gene expression is widespread in higher eukaryotes, but ncRNAs could also be involved in processes such as RNA interference, cosuppression, transgene silencing, imprinting and methylation. It has also been suggested that the basis of eukaryotic complexity and phenotypic variation may lie primarily in a control architecture composed of a system of RNAs that relay information required for the coordination and modulation of gene expression (Mattick, 2001, 2003).

Abstract

Noncoding RNA (ncRNA) genes are not involved in the production of mRNA and proteins, but produce transcripts that function directly as structural or regulatory RNAs. In the present study, the presence of ncRNA genes in the genome of *Burkholderia cenocepacia* J2315 was evaluated by combining comparative genomics (alignment-based) and predicted secondary structure approaches. Two hundred and thirteen putative ncRNA genes were identified in the *B. cenocepacia* J2315 genome and upregulated expression of four of these could be confirmed by microarray analysis. Most of the ncRNA gene transcripts have a marked predicted secondary structure that may facilitate interaction with other molecules. Several *B. cenocepacia* J2315 ncRNAs seem to be related to previously characterized ncRNAs involved in regulation of various cellular processes, while the function of many others remains unknown. The presence of a large number of ncRNA genes in this organism may help to explain its complexity, phenotypic variability and ability to survive in a remarkably wide range of environments.

Identifying ncRNAs in genomes is not straightforward as the DNA sequences coding for ncRNA lack many of the characteristic statistical biases often found in protein-coding regions (see e.g. Fickett, 1982; Staden, 1984). In addition, many ncRNAs are only expressed under specific environmental conditions, making confirmation of expression of candidate ncRNAs difficult (Eddy, 2001). Several computational approaches have been used to identify putative ncRNAs in prokaryotes. Using a machine learning approach with neural networks and support vector machines, several hundreds of putative ncRNAs were identified in the *Escherichia coli* K12 genome (Carter *et al.*, 2001). By looking for DNA regions that contain a sigma70 promoter within a short distance of a rho-independent terminator, 24 putative ncRNAs were identified in *E. coli*, of which the expression of 14 was confirmed using microarray analysis (Argaman *et al.*, 2001). A less restrictive search using the same approach identified 227 candidate ncRNAs (Chen *et al.*, 2002). Taking advantage of the high conservation of ncRNAs among closely related bacterial species, 59 putative ncRNAs were identified in the *E. coli* genome by comparing the intergenic regions of this genome with the genomes of

closely related species using BLAST (Wassarman *et al.*, 2001). The detection of putative ncRNAs using comparative sequence analysis is based on the fact that, contrary to conserved coding regions that are characterized by patterns of synonymous substitutions, conserved structural RNA genes are characterized by patterns of compensatory mutations consistent with some base-paired secondary structure (Rivas & Eddy, 2001; Rivas *et al.*, 2001). Using this approach, 275 candidate ncRNAs were identified in *E. coli*. Experimental confirmation of predicted ncRNAs is crucial to elucidate their function (Huttenhofer & Vogel, 2006). Wassarman *et al.* (2001) confirmed the expression of several of the predicted 59 ncRNAs in the *E. coli* genome by microarray analysis. Similarly, during a transcriptome analysis using high-density oligonucleotide probe arrays, 1102 transcripts were identified from intergenic regions in the *E. coli* genomes (Tjaden *et al.*, 2002), indicating the possible presence of expressed ncRNA genes. Using a shotgun cloning approach to generate cDNA libraries of small RNAs, the expression of 62 *E. coli* ncRNAs was experimentally confirmed (Vogel *et al.*, 2003). Hu *et al.* (2006) used an antibody-based microarray to detect very low amounts of *E. coli* ncRNA (down to 0.25 fmol). A detailed survey of 55 experimentally confirmed ncRNA genes in *E. coli* (Hershberg *et al.*, 2003) and revealed that they have a preference for the left replicore but no preference for the leading or lagging strand. The majority (71%) of ncRNA genes were found in intergenic regions ranging in size from 300 to 900 nt, while no ncRNA genes were found in intergenic regions < 50 nt and only very few in intergenic regions > 900 nt. In comparison with the G+C content of housekeeping genes (51.9%), the average G+C content of the ncRNA genes was lower (48.2%), but higher than the average G+C content of intergenic regions (42.4%). Eighty-five percent of the ncRNAs detected was between 50 and 250 nt and there appeared to be a weak correlation between the size of the ncRNA and the size of the intergenic region in which they were found. Using the approaches described above, putative ncRNAs have also been identified and experimentally confirmed in other organisms, including *Bacillus subtilis*, *Deinococcus radiodurans*, *Pseudomonas aeruginosa*, *Vibrio cholerae*, various extremophilic Gram-positive eubacteria (including *Bacillus halodurans*) and several Archaea (Schattner, 2002; Lenz *et al.*, 2005; Livny *et al.*, 2005, 2006; Hu *et al.*, 2006; Puerta-Fernandez *et al.*, 2006; Silvaggi *et al.*, 2006).

Members of the *Burkholderia cepacia* complex (a complex of at least nine closely related bacterial species, Coenye *et al.*, 2001) are commonly found in very diverse ecological niches, ranging from water, contaminated soils and the rhizosphere of various plants, to the respiratory tract of humans (Coenye & Vandamme, 2003). Persons with cystic fibrosis (CF) (the most common lethal inherited disorder in Caucasians) are particularly at risk for respiratory infections with these

organisms. The *B. cepacia* complex species most frequently recovered from respiratory secretions of CF patients is *Burkholderia cenocepacia*, and infection with this species is associated with increased morbidity and mortality (Coenye & LiPuma, 2003; Coenye & Vandamme, 2003). However, this species is also frequently isolated from various environmental sources including the rhizosphere and phyllosphere of plants (Coenye & Vandamme, 2003). The molecular and physiological backgrounds of the ecological diversity of this organism are at present unknown.

The exact function of many ncRNAs remains a mystery, although it is anticipated that they affect a large variety of cellular processes and play an important role in the regulation of gene expression. For instance, many of the anecdotally discovered ncRNAs in *E. coli* seem to be 'riboregulators,' which, using base complementarity, specifically interact with translational start sites and repress or activate translation (Eddy, 2001). The goal of the present study was to identify ncRNA genes in the genome of *B. cenocepacia* J2315. This genome is one of the largest prokaryotic genomes sequenced so far, with a size of nearly 8 Mb, and consists of three replicons.

Materials and methods

Genome sequences, gene prediction and construction of a genome atlas

The genome sequence of *B. cenocepacia* J2315 was produced by the Pathogen Sequencing Unit at the Sanger Institute and was obtained from http://www.sanger.ac.uk/Projects/B_cenocepacia/. Other genome sequences, including the genome sequence of *Ralstonia solanacearum* GMI1000 (Salanoubat *et al.*, 2002), were downloaded from the GenBank database. EasyGene (Larsen & Krogh, 2003) was used to predict the coding regions in the *B. cenocepacia* J2315 genome (with R -value=2). The location of tRNA-genes and rRNA genes were determined using tRNA-scan-SE (Lowe & Eddy, 1997) and BLASTN (Altschul *et al.*, 1997), respectively. Different atlases were made for each of the three chromosomes of the *B. cenocepacia* J2315 genome using the GENEWIZ software (Pedersen *et al.*, 2000).

Detection of ncRNAs

The QRNA software package (Rivas & Eddy, 2001) was used to identify putative ncRNA genes in the *B. cenocepacia* J2315 genome sequence. This program runs on a BLAST output (resulting from the comparison of two genomes) and scores all alignments according to different models, including a model that searches for motifs with secondary structure. QRNA combines the search for structural motifs using statistical models with comparative sequence analysis and as such may be a powerful tool for ncRNA prediction.

To reduce the probability of false-positive and false-negative predictions, the optimal phylogenetic distance between two genomes was determined for a QRNA search. When the phylogenetic distance between both genomes is small (i.e. when both organisms are closely related), a high number of false-positive predictions can be expected. In contrast, when the phylogenetic distance between both genomes is large (i.e. when both organisms are only distantly related), a high number of false-negative predictions can be expected. For this the genome of *E. coli* K12 MG1655 was used as query genome and the genomes of *E. coli* K12 W3110, O157 RIMD0509952, O157 EDL93 and CFT073, *Shigella flexneri* 2a301, *Salmonella typhi* CT18, *Salmonella enterica* Ty2, *Salmonella typhimurium* LT2, *Yersinia enterocolitica* 8081, *Yersinia pestis* CO92 and KIM and *Wigglesworthia brevipalpis* as test genomes, as well as a set of 50 experimentally confirmed ncRNA genes from the *E. coli* K12 MG1655 genome (Hershberg *et al.*, 2003).

Intergenic regions of the *B. cenocepacia* and *R. solanacearum* genomes were extracted and were used as query and subject sequences in the WU-BLAST. To reduce the number of reported alignments, alignments that were smaller than 50 nt, shared < 50% identity and/or had an *E* value higher than 0.1 were filtered. Each alignment was then scored by QRNA using the BLOSUM62 scoring matrix (Thompson *et al.*, 1994). The window size was set to 150 nt, the slide size to 50 nt and the cut-off for the posterior log-odds score of the RNA model to 5 bits. Subsequently, a number of Perl scripts was used to parse the QRNA output, resulting in a list of isolated regions that correspond to putative ncRNAs genes. Following the removal of tRNA and rRNA genes, the putative ncRNA genes that were located in regions with on average more repeats (global direct repeats, global inverted repeats, simple repeats, direct repeats, local inverted repeats, mirror repeats and everted repeats) than the rest of the genome were selected, by comparing them with a Repeat Atlas constructed for the *B. cenocepacia* J2315 genome. These repeat-rich regions are likely to have more secondary structure than the genomic average, and as biologically interesting RNA molecules have a more stable structure than expected by chance (Seffens & Digby, 1999; Le *et al.*, 2002; Bonnet *et al.*, 2004); this can be used as an additional selection criterion.

Characterization of putative ncRNAs

For all the predicted ncRNA genes that were identified as described above, the secondary structure and the minimal free energy were calculated. The minimal free energy, ΔG° , which is a measurement of the stability of the secondary structure, was calculated with *mfold* (Zuker, 2003) using the free energy data from Mathews *et al.* (1999). The conditions for folding were the standard conditions (37 °C, 1 M NaCl,

no divalent ions), which are equivalent to physiological conditions (Zuker, 2003). *mfold* was also used to calculate and visualize the secondary structures of all ncRNAs. To infer the function of the putative ncRNAs, sequences were compared with sequences present in the noncoding regulatory RNAs database (<http://biobases.ibch.poznan.pl/ncRNA/>) (Szymanski *et al.*, 2003), the Rfam database (<http://www.sanger.ac.uk/Software/Rfam/help/index.shtml>) (Griffiths-Jones *et al.*, 2003) and the tmRNA database (<http://www.indiana.edu/~tmrna/>) (Williams, 2002). The Rfam database contains 503 families of ncRNAs, while the tmRNA database holds the sequences of almost 400 ncRNA involved in a *trans*-translation process, in which a C-terminal peptide tag is added to an unfinished polypeptide chain stalled on a ribosome, targeting the unfinished protein for proteolysis. Conservation of the putative *B. cenocepacia* J2315 ncRNA genes in the genomes of other taxa was determined by BLAST (with *E* < 10). The genomes included are listed in Table 2. ncRNA genes were considered to be conserved in a group if they were present in all members of that group.

Microarray analysis

The expression of putative ncRNA genes was tested using a two-colour Agilent SurePrint custom microarray (Agilent Technologies UK Ltd., Stockport, UK) comprising probes for all annotated genes as well as for 1489 intergenic regions of the *B. cenocepacia* strain J2315 (Leiske *et al.*, 2006). Using this array, the transcriptional activity of J2315 incubated in 10% (w/v) CF sputum was compared with the growth in a minimal growth medium containing glucose and casein amino acids. Three different sputum samples in duplicate were each used to carry out the growth experiment. Total RNA was extracted from cultures reaching a difference at OD_{600 nm} of 0.6, converted to cDNA, fluorescently labelled and hybridized to the microarray. The hybridized arrays were scanned with the Agilent scanner and subsequently analysed using GENESPRING GX 7.3 software. Twenty-one features representing intergenic regions were found to be significantly up-regulated in at least two samples out of three upon the definition of minimal twofold change in gene expression and of the one-sample *t*-test *P*-value < 0.01. Details of the entire microarray experiment and procedure will be described elsewhere.

Semi-quantitative PCR

Microarray results were validated by semi-quantitative PCR targeting of one of ncRNAs found to be overexpressed in sputum, designated Bc4. The primers used to detect this transcript were designed manually: forward primer 5'-CCGCCGTACAGGCGTATAG-3' and reverse primer 5'-CTTGGACGGTGACGACAGAG-3'. The same cDNA

that had been analysed previously in the microarray experiments was used as the template for PCR. All cDNA aliquots for each growth condition (sputum or minimal growth medium growth) were pooled and added in equal concentration to a PCR reaction mix containing 2.5 mM MgCl₂, 0.2 mM (each) dNTPs, 0.5 μM (each) primer and 0.75 U of Taq polymerase (Promega, Madison). The PCR was run on an MJ Research PTC-200 thermal cycler with initial denaturation 5 min at 94 °C and a subsequent run of 40, 45 or 50 cycles, each comprising 30 s at 94 °C, 30 s at 64 °C and 30 s at 72 °C. The predicted PCR product of 101 bp in length was visualized on a 2.5% agarose gel after electrophoresis at 4 V cm⁻¹.

Results and discussion

Optimization of qRNA parameters

Using a number of enterobacterial genomes and a set of 50 experimentally verified ncRNAs from *E. coli*, it was shown that, as the phylogenetic distance between the two genomes used increased, the number of false-positive predictions reported by qRNA decreased, while the number of false-negative predictions increased drastically with decreasing phylogenetic distance (Fig. 1). To obtain a manageable number of predictions, the query and the test genomes should not be too closely related, but if they are too distant from each other, many ncRNA genes will go undetected. The present analyses indicate that using any other *Burkholderia* genome as the test genome would most likely result in a large number of false-positive predictions. Thus, the genome of *R. solanacearum* (the sequence similarity of its 16S rRNA gene to that of *B. cenocepacia* J2315 is 92%) was used to reduce the number of false-positive predictions. Of course, this also implies that the sensitivity is reduced, i.e. that some ncRNA genes will be missed. However, considering the large genome sizes and the large number of ncRNA predictions that can be expected in these genomes, missing some ncRNA genes was preferred over making a high number of false-positive predictions.

Identification of putative ncRNA genes in the *B. cenocepacia* J2315 genome

Using qRNA and the *R. solanacearum* GMI1000 genome as a reference, 3441 putative ncRNAs genes were identified in the *B. cenocepacia* J2315 genome (1781 on chromosome 1, 1333 on chromosome 2 and 327 on chromosome 3). Following the removal of tRNA and rRNA genes, the putative ncRNA genes that were located in regions with a more pronounced secondary structure than the rest of the genome were selected, by comparing them with a Repeat Atlas constructed for the *B. cenocepacia* J2315 genome (data not shown). This reduced the number of putative ncRNA genes to 213 (Table 1).

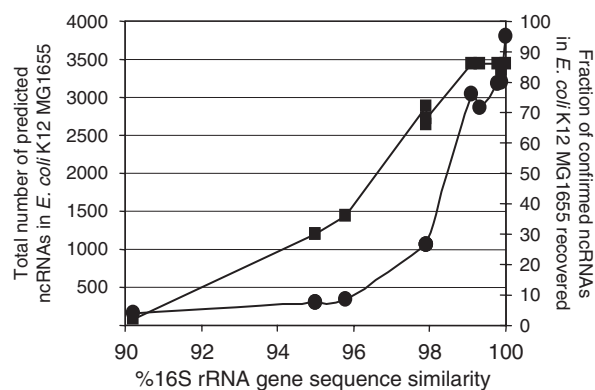


Fig. 1. Total number of predictions (squares, left axis) and fraction of experimentally confirmed ncRNAs recovered (circles, right axis) using test genomes with increasing 16S rRNA gene sequence similarity towards the query genome (*Escherichia coli* K12 MG1655).

The ncRNA genes are randomly distributed on the different chromosomes, and there are no significant differences between occurrences on the left or the right replicore, or on the leading or the lagging strand. Files containing the coordinates of all predicted ncRNA genes are available as supplementary material.

Properties of putative ncRNAs

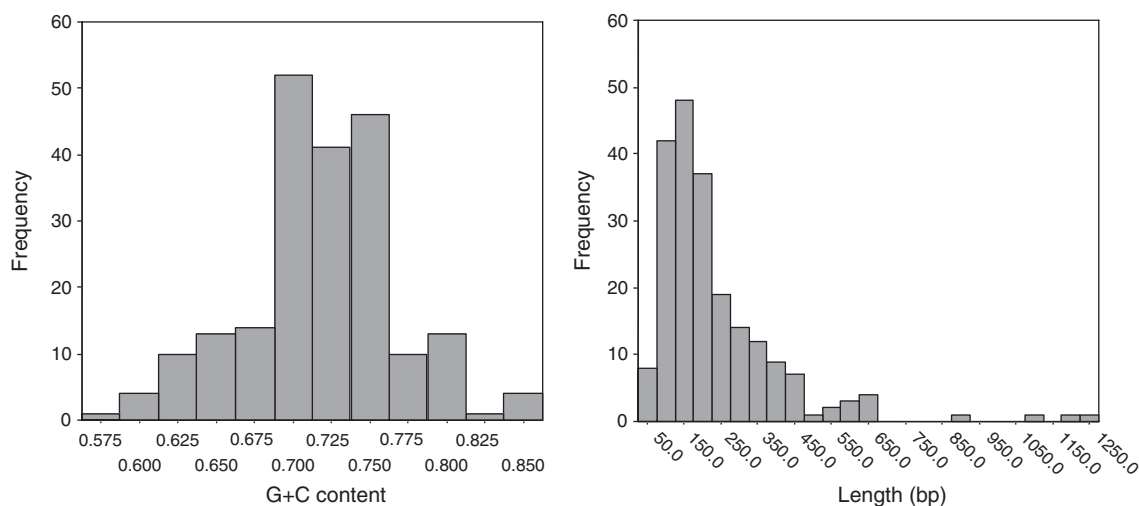
The 213 ncRNA genes identified varied in length between 53 and 1243 nt (Table 1, Fig. 2) and made up 0.63% of the complete genome. The G+C content of the ncRNA genes ranged from 58% to 86%. The average G+C content (72%) was significantly higher than the overall G+C content of the genome (67%) ($P < 0.001$) and also significantly higher than the average G+C content of the protein-coding genes (68%) ($P < 0.001$), the average G+C content of the intergenic regions (63%) ($P < 0.001$) and the average G+C content of ribosomal (53%) and transfer (61%) RNA genes ($P < 0.001$). The majority of the predicted ncRNAs showed significant predicted secondary structure, with free energies ranging from -27.9 to -665.4 kcal mol⁻¹.

Predicted function of ncRNAs in *B. cenocepacia* J2315

The high degree of predicted secondary structure observed in almost all putative ncRNAs identified so far suggests that they are involved in various types of molecular interactions. However, no function was assigned to the majority of putative ncRNAs identified in the last few years (Tjaden et al., 2002; Hershberg et al., 2003). Recently, the Noncoding RNA Database was established (Szymanski et al., 2003). This database includes the sequences of ncRNAs of different organisms, some of which were functionally characterized.

Table 1. Properties of the putative ncRNA genes in the *Burkholderia cenocepacia* J2315 genome

| | Chromosome 1 | Chromosome 2 | Chromosome 3 | Total |
|---|-------------------|--------------------|-------------------|-------------------|
| Total size (bp) | 3 870 082 | 3 217 062 | 875 677 | 7 963 121 |
| No. of ncRNA genes | 78 | 116 | 19 | 213 |
| Density (ncRNA genes/million bp) | 20.16 | 36.06 | 21.70 | 26.75 |
| Length (bp) | | | | |
| Total length | 17 434 | 27 819 | 4528 | 49781 |
| % of chromosome | 0.45 | 0.87 | 0.52 | 0.63 |
| Average (\pm SD) | 228 \pm 167 | 240 \pm 197 | 237 \pm 102 | 236 \pm 179 |
| Range (min–max) | 65–1093 | 53–1243 | 55–466 | 53–1243 |
| Minimal free energy (kcal mol ⁻¹) | | | | |
| Average (\pm SD) | -120.1 \pm 87.8 | -125.6 \pm 104.6 | -112.8 \pm 44.9 | -122.4 \pm 94.4 |
| Range (min–max) | -35.0 to -568.0 | -27.9 to -665.4 | -40.2 to -185.4 | -27.9 to -665.4 |
| G+C content (mol%) | | | | |
| Average (\pm SD) | 72 \pm 5 | 72 \pm 5 | 70 \pm 6 | 72 \pm 5 |
| Range (min–max) | 58–85 | 59–84 | 59–86 | 58–86 |

**Fig. 2.** Distribution of G+C content and length distribution of all putative ncRNAs.

A BLAST search against this database, using the putative ncRNA genes from *B. cenocepacia* J2315 as a query, resulted in several hits, even though the significance of the alignments was low. Two ncRNAs genes located on chromosome 1 produced significant ($E < 0.01$) alignments; one with *HgcE* from *Pyrococcus furiosus* (a ncRNA with unknown function) (Klein *et al.*, 2002), and one with *Acm_lbi* (a phage-derived trans-acting RNA molecule that interferes with lipopolysaccharide biosynthesis) (Mamat *et al.*, 1995). A single *B. cenocepacia* J2315 ncRNA located on chromosome 2 showed significant ($E = 0.0003$) similarity to a 6S RNA gene of *P. aeruginosa* (Vogel *et al.*, 1987). 6S RNA gene molecules interact with different RNA polymerase subunits and these interactions are thought to modulate the sigma70-holoenzyme activity (Wassarman & Storz, 2000). The Rfam database contains a large collection of multiple sequence alignments and covariance models covering many common

noncoding RNA families (Griffiths-Jones *et al.*, 2003). Two ncRNAs genes located on chromosome 1 resulted in significant alignments ($E = 0.003$ and 2.3×10^{-9}) with members of Rfam family RF00174, a group of cobalamin riboswitches involved in the modulation of expression of genes involved in vitamin B12 metabolism (Nahvi *et al.*, 2004). A third ncRNA gene from chromosome 1, as well as two ncRNA genes from chromosome 2, resulted in alignments with low significance ($E = 0.098$, 0.092 and 0.043, respectively) with the Rfam family containing bacterial RNase P class A (RF00010). Interestingly, both the *Burkholderia pseudomallei* K96243 and *Burkholderia mallei* ATCC 23344 genomes also contain ncRNA genes belonging to these Rfam families. Nevertheless, it remains to be determined whether the low-significance matches with entries present in the Noncoding RNA Database and in Rfam truly give an indication of the function of these ncRNAs. No

Table 2. Conservation of the putative *Burkholderia cenocepacia* J2315 ncRNA genes in the genomes of other taxa

| | Number (%) of <i>B. cenocepacia</i> J2315 ncRNAs with matches in all members of | | |
|--------------|---|-----------------------------|---------------------|
| | <i>B. cepacia</i> complex* | Genus <i>Burkholderia</i> † | Betaproteobacteria‡ |
| Chromosome 1 | 14/78 (18.0%) | 1/78 (1.3%) | 0/78 (0.0%) |
| Chromosome 2 | 6/116 (5.2%) | 0/116 (0.0%) | 0/116 (0.0%) |
| Chromosome 3 | 0/19 (0.0%) | 0/19 (0.0%) | 0/19 (0.0%) |
| Total | 20/213 (9.4%) | 1/213 (0.5%) | 0/213 (0.0%) |

**Burkholderia cepacia* complex strain 383, *Burkholderia ambifaria* (strains AMMD and MC40-6), *B. cenocepacia* (strains AU1054, HI2424, MC0-3 and PC184), *Burkholderia dolosa* AU0158, *Burkholderia multivorans* ATCC 17616 and *Burkholderia vietnamiensis* G4.

†All above-mentioned *B. cepacia* complex species, plus *Burkholderia mallei* (strains 10339, 2002721280, ATCC 23344, FMH, GB8 horse 4, JHU, NCTC 10229, NCTC 10247, and SAVP1), *Burkholderia phymatum* STM815, *Burkholderia phytofirmans* PsJN, *Burkholderia pseudomallei* (strains 1106a, 1106b, 1655, 1710a, 1710b, 406e, 668, K96243, Pasteur, and S13), *Burkholderia thailandensis* E264, and *Burkholderia xenovorans* LB400.

‡All above-mentioned *Burkholderia* species, plus *Bordetella bronchiseptica* RB50, *Bordetella parapertussis* 12822, *Bordetella pertussis* Tohama I, *Cupriavidus necator* H16, *Cupriavidus pinatubonensis* JMP134, *Cupriavidus metallidurans* CH34, *Polynucleobacter* sp. QLW-P1DMWA-1, *Ralstonia pickettii* 12), *Ralstonia solanacearum* (strains GMI1000 and UW551), *Chromobacterium violaceum* ATCC 12472, *Neisseria gonorrhoeae* FA 1090 and *Neisseria meningitidis* (strains FAM18, MC58 and Z2491).

Table 3. Properties of the experimentally confirmed ncRNA genes in the *Burkholderia cenocepacia* J2315 genome

| Name | Replicon | Start | End | Length (bp) | Free energy (kcal mol ⁻¹) | GC content (%) | Conserved in* | 5' gene† | 3' gene‡ |
|------|----------|---------|---------|-------------|---------------------------------------|----------------|-----------------------|-----------|----------|
| Bc1 | Chr 1 | 950306 | 950478 | 172 | -91.73 | 74.4 | <i>Bcc</i> | BCAL0873 | BCAL0874 |
| Bc2 | Chr 1 | 2773902 | 2774032 | 130 | -36.26 | 61.8 | - | BCAL2506a | BCAL2507 |
| Bc3 | Chr 1 | 2924806 | 2925138 | 332 | -101.75 | 62.5 | <i>B. cenocepacia</i> | BCAL2652 | BCAL2653 |
| Bc4 | Chr 2 | 1900468 | 1900668 | 200 | -89.08 | 74.5 | - | BCAM1704 | BCAM1705 |

**Bcc*, *Burkholderia cepacia* complex; *B. cenocepacia*, strains J2315, AU1054, HI2424, MC0-3 and PC184; -, no match in database.

†BCAL0873, import inner membrane translocase, subunit Tim44; BCAL0874, protein of unknown function; BCAL2506a, conserved hypothetical protein; BCAL2507, conserved hypothetical protein; BCAL2652, DoxX family protein; BCAL2653, cobyrinic acid synthase CobQ; BCAM1704, alcohol dehydrogenase GroES domain protein; BCAM1705, periplasmic sensor signal transduction histidine kinase.

meaningful matches were found with *B. cenocepacia* J2315 ncRNA genes in the tmRNA database.

Conservation of putative ncRNA genes

It was investigated whether predicted ncRNA genes were conserved between *B. cenocepacia* J2315 and closely and more distantly related taxa. It was previously demonstrated that ncRNA genes found in *E. coli* tend to be conserved only in closely related bacteria like *Shigella* sp. and *Salmonella* sp., but that there is very little (if any) conservation if the phylogenetic distance between two organisms (as measured by 16S rRNA gene sequence similarity) increases beyond a certain threshold (Hershberg et al., 2003). This was confirmed in the present study, as almost no conservation was observed beyond the borders of the *B. cepacia* complex (Table 2). Even within the *B. cepacia* complex (whose members typically share 98–100% 16S rRNA gene sequence similarity; Coenye et al., (2001)) conservation is rather low (9.4% for the whole genome). No conservation was observed outside the genus *Burkholderia*. It should be noted that, because BLAST was used to assess conservation of putative ncRNA genes among various organisms, homologous genes that do not share much sequence similarity but do have a similar secondary structure may have been missed.

Confirmation of expression of ncRNAs

A recently developed two-colour microarray was used to screen for the production of transcripts from intergenic regions (i.e. putative ncRNAs). Twenty-one intergenic regions were found to be up-regulated during growth in human sputum and it could be confirmed that four of them correlated to predicted ncRNAs (designated Bc1, Bc2, Bc3 and Bc4) (Table 3). Noticeable specific production of these four predicted ncRNA was only observed when *B. cenocepacia* J2315 was grown in 10% (w/v) CF sputum, and not when cells were grown in minimal growth medium containing glucose and casamino acids (data not shown). Of these four ncRNAs, three were located on chromosome 1, while one was located on chromosome 2. All confirmed ncRNAs had a marked secondary structure (Fig. 3), again suggesting that they may be involved in various types of molecular interactions. In order to confirm that the signal observed in the microarray experiments was derived from the predicted ncRNA and not from the 3' untranslated region of the mRNA, it was attempted to develop primers for each of the four ncRNA in order to confirm their expression with real-time PCR. Unfortunately, due to the marked secondary structure of the ncRNAs, only primers for Bc4 could be developed. These primers were situated within the predicted

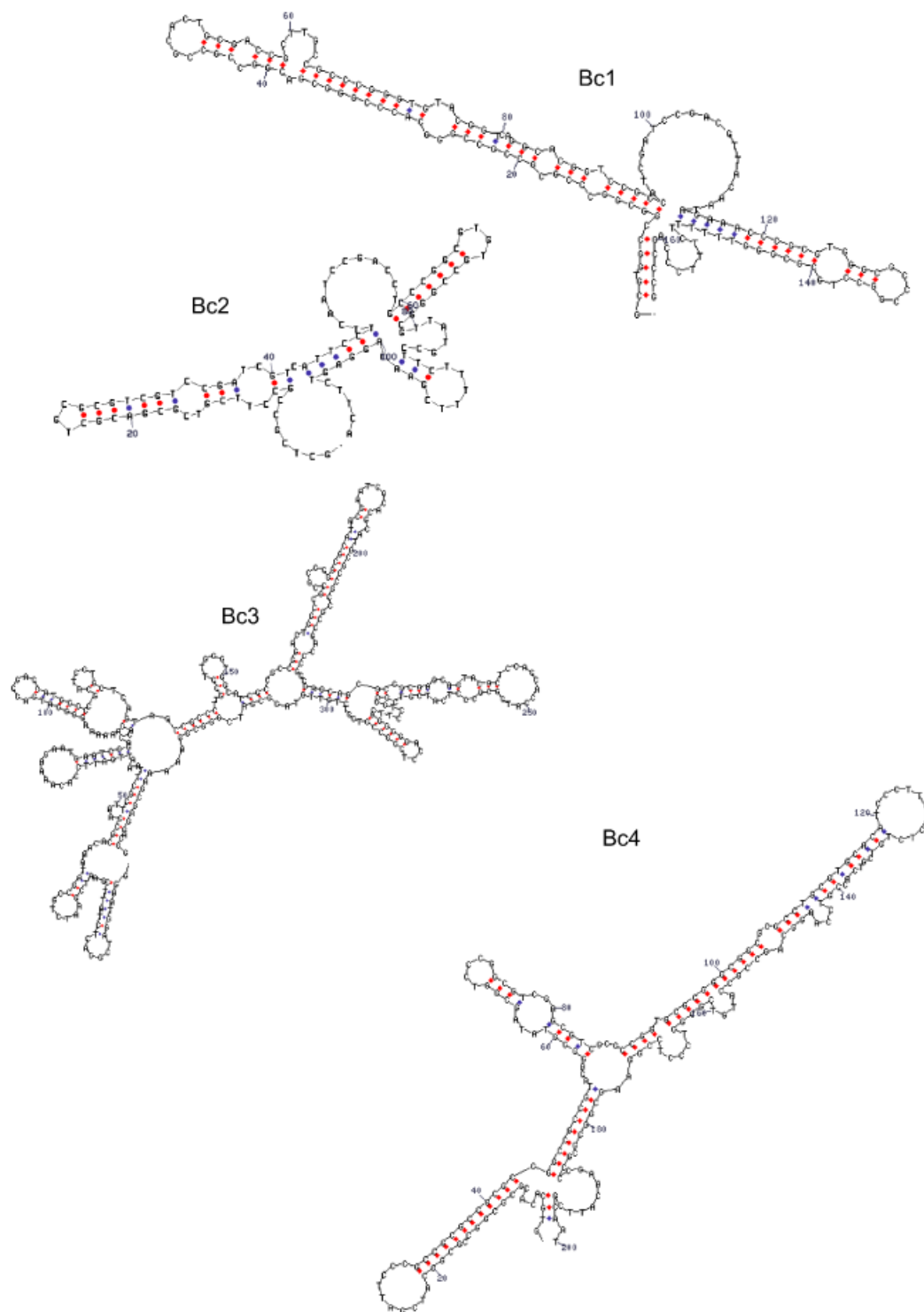


Fig. 3. Secondary structure of the confirmed ncRNAs.

ncRNA sequence (Fig. 4). Using this approach, it could be shown that (1) Bc4 is expressed in both conditions and (2) Bc4 expression is upregulated in sputum compared with control conditions (Fig. 4). It should be noted that inter-

genic regions were only represented by a single 60-mer probe on the microarray, meaning that production of an ncRNA from a part of an intergenic region not covered by the probe would go unnoticed. In addition, not all

↓
 GTGCACACGCGCGGCCGCGGCATCGATTCCCGCCGCGCCGCGCCGGCCCGCGTACAGGCGTATAGCGGT
 CCCGCCGTGAGCGTCGCCCGGTGCGCGGGCGGCGCGCCTGCGTGACATCCCTTCCCTCTGTGCGTCA
CCGTCCAAGGCAGCCGCCATGTCGCGCTCCCTCCGGAAGCGGCGCGCCCGAACATTGCAATCCCTAACA
 TTTACCGCAACGCGCGATCCGCACCGGTGCGCGATCAGGGTTTCTCCGCATCAAGCACGATGCGTCCGGC
 GTCGACAGG

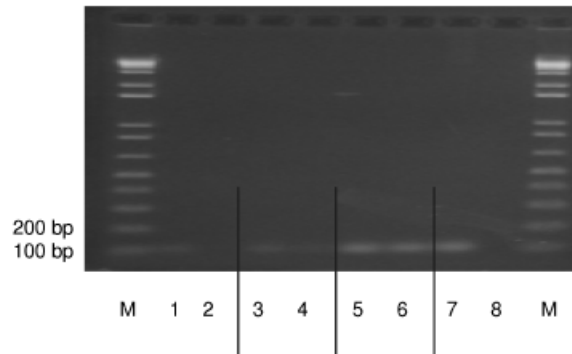


Fig. 4. Confirmation of expression of Bc4. Upper part. Overview of IG2_1900469, the intergenic region comprising ncRNA Bc4. Arrows indicate the 5' and 3' ends of the ncRNA. The boxed regions are the primer-binding sites in the real-time PCR, and the underlined sequence is the sequence of the probe on the microarray. Lower part. Result of real-time PCR on cDNA derived from sputum-grown cells (lanes 1, 3 and 5) and on cDNA derived from cells grown under control conditions (lanes 2, 4 and 6). Lanes 1 and 2, 40 cycles; lanes 3 and 4, 45 cycles; lanes 5 and 6, 50 cycles. Lane M, molecular weight marker; lane 7, positive control (genomic DNA J2315); lane 8, negative control of PCR.

intergenic regions were included on the array. The focus was also on the intergenic regions that showed significant upregulation, and hence only detected ncRNAs with strong evidence of expression in comparison with those that may have had constant expression under the growth conditions applied. Overall, the combination of these factors may have contributed to the low number of experimentally confirmed ncRNAs using the whole genome microarray approach. The production of transcripts from intergenic regions in which QRNA did not identify ncRNA genes (under both conditions tested) was also noticed. However, these were not further investigated in the present study.

Use of QRNA to predict ncRNAs in microbial genomes

Analysis of data from previous studies (Rivas *et al.*, 2001) showed that the use of secondary structure conservation as a sole predictor of ncRNA genes in *P. aeruginosa* may result in a lower proportion of predicted ncRNA genes actually being confirmed using other methods, suggesting a higher proportion of false-positive predictions. In addition, some experimentally confirmed *P. aeruginosa* ncRNA genes were not detected, either because they lacked a conserved secondary

structure or because the structure went undetected by QRNA (Rivas *et al.*, 2001; Livny *et al.*, 2006). Our own microarray experiments confirmed that it is likely that several ncRNA genes were not detected by QRNA as the production of transcripts from intergenic regions in which QRNA did not identify ncRNA genes was noticed. A more refined approach to ncRNA detection would be to use QRNA as a first screen and then use operon modelling (e.g. finding a transcription start site and potential terminator) to reduce the number of false-positives. The incorporation of some of this knowledge into hidden Markov model-based gene finders may allow the development of automated methods for the reliable prediction of ncRNAs in bacterial genomes. Recently, a novel tool (sRNAPredict2) was developed (Livny *et al.*, 2005, 2006), in which QRNA predictions are combined with predictions based on BLAST E and score values, the presence of Rho-independent terminators and the presence of putative promoters. The use of this (and other) novel prediction algorithm will likely allow the more efficient prediction of ncRNAs in microbial genomes. The conclusion of the recent paper by Frehult *et al.* (2007) that 'the most popular homology search methods (on noncoding RNA) are often the least accurate' again highlights the necessity for further research in this area.

Conclusion

In the present study 213 putative ncRNA genes in the *B. cenocepacia* J2315 genome were identified, of which we confirmed the expression of four by microarray hybridizations. Many of the ncRNA gene transcripts have a marked predicted secondary structure that may allow interaction with other molecules. Several *B. cenocepacia* J2315 ncRNAs seem to be related to previously characterized ncRNAs involved in regulation of various cellular processes, while the function of many others remains unknown at present. The presence of a large number of ncRNA genes in this organism may help to explain its complexity, phenotypic variability and ability to survive in a remarkably wide range of environments. As qRNA has a relatively low sensitivity and is particularly insensitive to ncRNAs with little or no conserved intramolecular secondary structure and as the production of transcripts from intergenic regions in which no ncRNA genes were predicted was noticed, it is highly unlikely that this screen has revealed all ncRNA genes in the *B. cenocepacia* J2315 genome.

Acknowledgements

T.C. and P.V. are indebted to the Fund for Scientific Research – Flanders (Belgium) for funding. T.C. also acknowledges the support from the Belgian Federal Government (Federal Office for Scientific, Technical and Cultural Affairs). P.D. and E.M. acknowledge funding from the United Kingdom Wellcome Trust (grant 075586), R.T.G. acknowledges funding from the United States Cystic Fibrosis Foundation to enable the development of the microarray and D.W.U. acknowledges funding from the Danish Center for Scientific Computing.

References

- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W & Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.
- Argaman L, Hershberg R, Vogel J, Bejerano G, Wagner EGH, Margalit H & Altuvia S (2001) Novel small RNA-encoding genes in the intergenic region of *Escherichia coli*. *Curr Biol* **11**: 941–950.
- Bonnet E, Wuyts J, Rouze P & Van de Peer Y (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics* **20**: 2911–2917.
- Carter RJ, Dubchak I & Holbrook SR (2001) A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Res* **29**: 3928–3938.
- Chen S, Lesnik EA, Hall TA, Sampath R, Griffey RH, Ecker DJ & Blyn LB (2002) A bioinformatics based approach to discover small RNA genes in the *Escherichia coli* genome. *BioSystems* **65**: 157–177.
- Coenye T & LiPuma JJ (2003) Molecular epidemiology of *Burkholderia* species. *Fronti Biosci* **8**: e55–e67.
- Coenye T & Vandamme P (2003) Diversity and significance of *Burkholderia* species occupying diverse ecological niches. *Env Microbiol* **5**: 719–729.
- Coenye T, Vandamme P, Govan JRW & LiPuma JJ (2001) Taxonomy and identification of the *Burkholderia cepacia* complex. *J Clin Microbiol* **39**: 3427–3436.
- Eddy SR (2001) Non-coding RNA genes and the modern RNA world. *Nature Rev Gen* **2**: 919–929.
- Fickett JW (1982) Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res* **11**: 5303–5318.
- Frehult EK, Bollback JP & Gardner PP (2007) Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. *Gen Res* **17**: 117–125.
- Griffiths-Jones S, Bateman A, Marshall M, Khanna A & Eddy SR (2003) Rfam: an RNA family database. *Nucleic Acids Res* **31**: 439–441.
- Hershberg R, Altuvia S & Margalit H (2003) A survey of small RNA-encoding genes in *Escherichia coli*. *Nucleic Acids Res* **31**: 1813–1820.
- Hu Z, Zhang A, Storz G, Gottesman S & Leppla SH (2006) An antibody-based microarray assay for small RNA detection. *Nucleic Acids Res* **34**: e52.
- Huttenhofer A & Vogel J (2006) Experimental approaches to identify non-coding RNAs. *Nucleic Acids Res* **34**: 635–646.
- Klein RJ, Misulovin Z & Eddy SR (2002) Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc Natl Acad Sci USA* **99**: 7542–7547.
- Larsen TS & Krogh A (2003) EasyGene—a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinform* **4**: 21.
- Le SY, Zhang K & Maizel JV (2002) RNA molecules with structure dependent functions are uniquely folded. *Nucleic Acids Res* **30**: 3574–3582.
- Leiske DL, Karimpour-Fard A, Hume PS, Fairbanks BD & Gill RT (2006) A comparison of alternative 60-mer probe designs in an *in-situ* synthesized oligonucleotide microarray. *BMC Genomics* **7**: 72.
- Lenz DH, Miller MB, Zhu J, Kulkarni RV & Bassler BL (2005) CsrA and three redundant small RNAs regulate quorum sensing in *Vibrio cholerae*. *Mol Microbiol* **58**: 1186–1202.
- Livny J, Fogel A, Davis BM & Waldor MK (2005) sRNAPredict: an integrative computational approach to identify sRNAs in bacterial genomes. *Nucleic Acids Res* **33**: 4096–4105.
- Livny J, Brencic A, Lory S & Waldor MK (2006) Identification of 17 *Pseudomonas aeruginosa* sRNAs and prediction of sRNA-encoding genes in 10 diverse pathogens using the bioinformatics tool sRNAPredict2. *Nucleic Acids Res* **34**: 3484–3493.

- Lowe TM & Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**: 955–964.
- Lung B, Zemann A, Madej MJ, Schuelke M, Techritz S, Ruf S, Bock R & Huttenhoffer A (2006) Identification of small non-coding RNAs from mitochondria and chloroplasts. *Nucleic Acids Res* **34**: 3842–3852.
- Mamat U, Rietschel ET & Schmidt G (1995) Repression of lipopolysaccharide biosynthesis in *Escherichia coli* by an antisense RNA of *Acetobacter methanolicus* phage Acml. *Mol Microbiol* **15**: 1115–1125.
- Mathews DH, Sabina J, Zuker M & Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288**: 911–940.
- Mattick JS (2001) Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep* **2**: 986–991.
- Mattick JS (2003) Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *BioEssays* **25**: 930–939.
- McCutcheon JP & Eddy SR (2003) Computational identification of non-coding RNAs in *Saccharomyces cerevisiae* by comparative genomics. *Nucleic Acids Res* **31**: 4119–4128.
- Nahvi A, Barrick JE & Breaker RR (2004) Coenzyme B12 riboswitches are widespread genetic control elements in prokaryotes. *Nucleic Acids Res* **32**: 143–150.
- Pedersen AG, Jensen LJ, Brunak S, Staerfeldt HH & Ussery DW (2000) A DNA structural atlas for *Escherichia coli*. *J Mol Biol* **299**: 907–930.
- Puerta-Fernandez E, Barrick JE, Roth A & Breaker RR (2006) Identification of a large noncoding RNA in extremophilic eubacteria. *Proc Natl Acad Sci USA* **103**: 19490–19495.
- Rivas E & Eddy SR (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinform* **2**: 8.
- Rivas E, Klein RJ, Jones TA & Eddy SR (2001) Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr Biol* **11**: 1369–1373.
- Salanoubat M, Genin S, Artiguenave F et al. (2002) Genome sequence of the plant pathogen *Ralstonia solanacearum*. *Nature* **415**: 497–502.
- Schattner P (2002) Searching for RNA genes using base-composition statistics. *Nucleic Acids Res* **30**: 2076–2082.
- Seffens W & Digby D (1999) mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res* **27**: 1578–1584.
- Silvaggi JM, Perkins JB & Losick R (2006) Genes for small, noncoding RNAs under sporulation control in *Bacillus subtilis*. *J Bacteriol* **188**: 532–541.
- Staden R (1984) Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes. *Nucleic Acids Res* **12**: 551–567.
- Szymanski M, Erdmann VA & Barciszewski J (2003) Noncoding regulatory RNAs database. *Nucleic Acids Res* **31**: 429–431.
- Thompson JD, Higgins DG & Gibson TJ (1994) Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput Appl Biosci* **10**: 19–29.
- Tjaden B, Saxena RM, Stolyar S, Haynor DR, Kolker E & Rosenow C (2002) Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. *Nucleic Acids Res* **30**: 3723–3728.
- Vogel DW, Hartmann RK, Struck JCR, Ulbrich N & Erdmann VE (1987) The sequence of the 6S RNA gene of *Pseudomonas aeruginosa*. *Nucleic Acids Res* **15**: 4583–4591.
- Vogel J, Bartels V, Tang TH, Churakov G, Slatger-Jäger JG, Hüttenhofer A & Wagner EGH (2003) RNomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res* **31**: 6435–6443.
- Wassarman KM & Storz G (2000) 6S RNA regulates *E. coli* RNA polymerase activity. *Cell* **101**: 613–623.
- Wassarman KM, Repoila F, Rosenow C, Storz G & Gottesman S (2001) Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev* **15**: 1637–1651.
- Williams KP (2002) The tmRNA website: invasion by an intron. *Nucleic Acids Res* **30**: 179–182.
- Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**: 3406–3415.

Supplementary material

The following supplementary material is available for this article online:

Table S1. List of ncRNA found on chromosome 1 of *B. cenocepacia* J2315.

Table S2. List of ncRNA found on chromosome 2 of *B. cenocepacia* J2315.

Table S3. List of ncRNA found on chromosome 3 of *B. cenocepacia* J2315.

This material is available as part of the online article from: <http://www.blackwell-synergy.com/doi/abs/10.1111/j.1574-6968.2007.00916.x> (This link will take you to the article abstract).

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.