Journal of Microbiology (2014) Vol. 52, No. 6, pp. 482–489
Copyright ⓒ 2014, The Microbiological Society of Korea

DOI 10.1007/s12275-014-3658-3

Prediction of Bacterial microRNAs and Possible Targets in Human Cell Transcriptome[§]

Amir Shmaryahu*, **Margarita Carrasco, and Pablo D.T. Valenzuela**

Fundación Ciencia & Vida. Zañartu 1482, Santiago, Chile

(Received Dec 19, 2013 / Revised Jan 23, 2014 / Accepted Jan 24, 2014)

Recent studies have examined gene transfer from bacteria to humans that would result in vertical inheritance. Bacterial DNA appears to integrate into the human somatic genome through an RNA intermediate, and such integrations are detected more frequently in tumors than normal samples and in RNA than DNA samples. Also, vertebrate viruses encode products that interfere with the RNA silencing machinery, suggesting that RNA silencing may indeed be important for antiviral responses in vertebrates. RNA silencing in response to virus infection could be due to microRNAs encoded by either the virus or the host. We hypothesized that bacterial expression of RNA molecules with secondary structures is potentially able to generate miRNA molecules that can interact with the human host mRNA during bacterial infection. To test this hypothesis, we developed a pipeline-based bioinformatics approach to identify putative microRNAs derived from bacterial RNAs that may have the potential to regulate gene expression of the human host cell. Our results suggest that 68 bacterial RNAs predicted from 37 different bacterial genomes have predicted secondary structures potentially able to generate putative microRNAs that may interact with messenger RNAs of genes involved in 47 different human diseases. As an example, we examined the effect of transfecting three putative microRNAs into human embryonic kidney 293 (HEK293) cells. The results show that the bacterially derived microRNA sequence can significantly regulate the expression of the respective target human gene. We suggest that the study of these predicted microRNAs may yield important clues as to how the human host cell processes involved in human diseases like cancer, diabetes, rheumatoid arthritis, and others may respond to a particular bacterial environment.

***Keywords*:** bioinformatics, RNA structure, microRNA, bacterial RNA, human diseases

Introduction

The regulation of transcription and translation in eukaryotes is complex and is growing ever more complicated with each new discovery. One layer of regulation recently ascribed to these processes is by microRNAs (miRNAs) that are expressed from DNA regions previously thought to be silent. These miRNAs function by altering the expression of endogenous genes. miRNAs were first described in *Caenorhabditis elegans* with the discovery that an RNA molecule, lin-4, downregulates the expression of the *lin-14* gene (Wightman *et al.*, 1993). However, since there is no homolog to lin-4 in other species, this discovery was considered to be unique to *C. elegans*.

Since this discovery, however, specific and potent silencing of genes by double-stranded RNA (dsRNA), RNA interference (RNAi), has been reported (Fire *et al.*, 1998) and found to be a common mode of gene regulation. In addition, another miRNA, named let-7, was later discovered in *C. elegans* (Reinhart *et al.*, 2000). In contrast to lin-4, it has homologs in other species including humans, suggesting that miRNAs may play a common regulatory role in eukaryotes.

RNA silencing is critical in plant and animal development (Ambros, 2004; Baulcombe, 2004). It is also important for immunity against viruses in plants and insects, as the endogenous RNA silencing machinery induced by viral dsRNA degrades the invading genetic material. It has long been unclear whether RNA silencing plays a role in immunity in vertebrates since these organisms have other sophisticated innate mechanisms for responding to viral dsRNA, including the protein kinase R-dependent antiviral response and the Toll-like receptor system (Akira and Takeda, 2004). Recent studies reveal that the genomes of vertebrate viruses encode products that interfere with the host RNA silencing machinery (Voinnet, 2005), suggesting that RNA silencing may indeed be important for antiviral responses in vertebrates. Several virally encoded miRNAs have now been found, but their relevance to infection is not clear. Five miRNAs were identified to be encoded by the genome of the herpesvirus Epstein-Barr virus (EBV) (Pfeffer *et al.*, 2004). One of these miRNAs, miR-BART2, targets for cleavage the messenger RNA (mRNA) for EBV DNA polymerase (BALF5). Recent computational predictions combined with cloning have identified additional miRNAs from other herpesviruses, although their function remains unknown (Cai *et al.*, 2005; Pfeffer *et al.*, 2005; Samols *et al.*, 2005). Interestingly, an miRNA found to be encoded by the papovavirus simian virus 40 (SV40) genome is derived from the late transcript and targets the transcript of the large T antigen for cleavage (Sullivan *et al.*, 2005). This does not affect viral replication *in vitro*, but may func-

*For correspondence. E-mail: amir.shmaryahu@gmail.com; Tel.: +562-2367-2056; Fax: +569-9449-4006

[§]Supplemental material for this article may be found at http://www.springerlink.com/content/120956.

tion to limit the expression of the large T antigen *in vivo*. Abrogating this miRNA-mediated suppression of T antigen increased the recognition of SV40-infected cells by antigen-specific cytotoxic T cells (Sullivan *et al.*, 2005). Thus, the viral miRNA may reduce the susceptibility of the virus to the host immune system.

Although the vast majority of bacteria do not cause human diseases, many species are pathogenic. One of the interesting bacterial diseases with the highest disease burden is ulcer, which is caused by the bacterium *Helicobacter pylori*. This bacterium lives in the stomachs of most people and does little harm (Marshall and Warren, 1983, 1984). However, in 10% to 15% of cases, it can stimulate gastric ulcers. Ulcers are caused by the formation of acidic products by *H. pylori* in the stomach, leading to damage of the inner linings of the intestines. Marshall and Warren have discovered that this bacterium is present in more than 90% of duodenal ulcers and up to 80% of stomach ulcers. *H. pylori* appears to lead to stomach cancer, which has the second highest mortality rate among cancers. However, the mechanisms through which this bacterium leads to cancer are as yet unclear. Pathogenic bacteria contribute to other globally important diseases, such as pneumonia, which can be caused by bacteria such as *Pseudomonas aeruginosa* and *Staphylococcus aureus* (Ebby, 2005), and foodborne illnesses, which can be caused by bacteria such as *Shigella*, *Campylobacter*, and *Salmonella* (Naimi *et al.*, 2003). Pathogenic bacteria also cause infections such as tetanus (Farrar *et al.*, 2000), typhoid fever (Giannella, 1996), diphtheria (Freeman, 1951), syphilis (Eccleston *et al.*, 2008), leprosy (Sasaki *et al.*, 2001), and stomach ulcers. Many pathogenic bacteria produce toxins that aid in their pathogenicity.

One interesting and recent study has found evidence that lateral gene transfer is possible from bacteria to the cells of the human body, known as human somatic cells. They found that the bacterial DNA was more likely to integrate into the genome in tumor samples than in normal, healthy somatic cells. The phenomenon might play a role in cancer and other diseases associated with DNA damage (Riley *et al.*, 2013). No single bacterial species has been identified as a risk factor for cancer, but an increase in the abundance of *Fusobacterium* in human colorectal tumors compared with controls has been reported (Marchesi *et al.*, 2011; Kostic *et al.*, 2012). These studies suggest that *Fusobacterium* may be associated with the later stages of colon cancer, but it is unknown whether they play a role in the early stages of colorectal carcinogenesis. While the causes of colorectal cancer are not fully known, it is becoming increasingly clear that the gut microbiota provide an important contribution (Azcarate-Peril *et al.*, 2011).

These results prompted the question of whether pathogenic bacteria encode RNAs with secondary structures similar to those of miRNAs and, if so, whether these RNAs are important for the pathogenicity of the bacteria. Based on these questions, we generated the hypothesis that the bacterial expression of RNA molecules with secondary structures is potentially able to generate miRNA molecules that can interact with the human host mRNA and lead to new regulatory mechanisms that may be involved in human diseases. The general goal of this study was to search for and identity possible miRNAs generated from bacterial RNAs that may have the potential to regulate gene expression of the human host cell. To address these questions, we developed a novel bioinformatics approach to identify structures in bacterial genomes that present the properties of miRNAs and have target sequences in the human genome. This involved: 1) the analysis of all bacterial gene sequences from the complete genome, including non-coding RNAs; 2) prediction of the secondary structure of each bacterial RNA and the search for candidates of miRNAs generated from the total bacterial RNA sequences; and 3) the identification of possible targets in the human mRNA sequences that may bind to these predicted bacterial miRNAs. Many studies have been conducted on pathogenic bacteria mechanisms and those affecting human host cells, but this was the first study on a possible novel mechanism of regulation by bacterial RNA.

Materials and Methods

DNA sequence sources and gene annotations

The following genomes were analyzed in this study: 32 complete genomes and six draft genomes from the Actinobacteria group; 28 complete genomes and 10 draft genomes from the Alphaproteobacteria group; five complete genomes from the Bacteroidetes/Chlorobi group; 21 complete genomes and one draft genome from the Betaproteobacteria group; 14 complete genomes from the Chlamydiae/Verrucomicrobia group; one complete genome from the Cyanobacteria group; 17 complete genomes and one draft genome from the Epsilonproteobacteria group; 101 complete genomes and 19 draft genomes from the Firmicutes group; two complete genomes and three draft genomes from the Fusobacteria group; 101 complete genomes and 58 draft genomes from the Gammaproteobacteria group; 13 complete genomes from the Spirochetes group and two complete genomes and nine draft genomes from the 'Other bacteria' group (see Supplementary data Table S2). All sequences were consolidated into one large database named DB.1-PB (database 1 pathogenic bacteria). This database contained a name, project code number, and group and sequence for each bacterial species. All sequences were obtained from NCBI (http://www.ncbi.nlm.nih.gov/). The data included 334 complete genomes and 104 draft genomes. Genes were extracted from DB.1-PB using the BioPerl software (http://www.bioperl.org/). Genes were annotated based on GenBank (http://www.ncbi.nlm.nih.gov/genbank/) (Kostic *et al.*, 2012) and gene sequences were transferred to a new database named DB.2-BG (database 2 bacterial genes) that included a FASTA format barcode identification for each gene.

Nucleic acid folding and miRNA prediction (Supplementary data Table S2)

Using the base UNAFold program (http://mfold.rna.albany.edu/) (Azcarate-Peril *et al.*, 2011), we individually predicted the secondary structure of each gene in DB.2-BG. This program is an integrated collection of programs that simulate folding, hybridization, and melting pathways for one or two single-stranded nucleic acid sequences. Folding (secondary structure) prediction for single-stranded RNA or DNA combines free energy minimization, partition function calculations, and stochastic sampling. For melting simulations, the package computes entire melting profiles, not just melting temperatures. Ultraviolet (UV) absorbance at 260 nm, heat capacity change (Cp), and mole fractions of different molecular species are computed as a function of temperature. The package installs and runs on Linux platforms. Depending on the length of the gene, it took from 3 sec up to 40 min to calculate one secondary structure using one CPU and 4 GB RAM. Between one and 100 predictions of secondary structure were made for each gene. Parameters were left at the default settings of the software apart from temperature, which was set to 37°C. In the second stage, all results were scanned using "*". CT output files from BioPerl to identify predictions with similar secondary structures to pre-miRNA. In the case of a match, the sequence was transferred to the database DB.3-FCT (database 3 folding base CT files).

miRNA targets and disease relationship (Supplementary data Table S3)

Each gene in DB.3-FCT was put into a new database, DB.4-miRNA (database 4 miRNA prediction) in FASTA format. This database also includes the additional information included in DB.3-FCT. For every gene in DB.4-miRNA, we used BioPerl and the BLAST program (http://blast.ncbi.nlm.nih.gov/Blast.cgi) [31] to identify complementary sequence relationships between the miRNA predictions and all human mRNAs, including mitochondrial mRNA, found in the NCBI database. In the case of a match, we checked whether the complementary segment was within the predicted miRNA sequence of the bacterial gene. If this condition was met, the gene, miRNA prediction information, and the target gene, function, and sequence were added to a new database named DB.5-T (database 5 targets). To complete our database, we searched each gene listed in the DB.5-T database in the OMIM database (http://omim.org/) (Hamosh *et al.*, 2005), which contains information on medical studies, associated diseases, and gene mutations for every human gene. This search could provide useful information on diseases that may be caused by an miRNA disrupting the expression of its target gene. This search was performed using the BioPerl software.

Cell culture

The HEK293 cells was originally derived from human embryonic kidney cells grown in tissue culture. HEK293 cells were obtained from ATCC, catalog# CRL-1573. Cells were grown in Dulbecco's modified Eagle's medium (Invitrogen Cat.#11965-065) supplemented with 10% fetal bovine serum (Invitrogen, Cat.#16000-044) and antibiotics 1X (Invitrogen Cat.#15240). Cells were maintained at between 10% and 90% confluency in a 37°C, 5% CO_2 tissue culture incubator.

HEK293 transfection, cell mRNA extraction, and cDNA synthesis

Transfection was performed using a cationic lipid (Lipofectamine®) according to the specifications of the manufacturer. Briefly, 3.5×10^3 cells were seeded per well of a 12-well plate the day prior to transfection. Cells were transfected with synthetic bacterial miRNA (8 nM, 30 nM, and 60 nM) and cultured in OPTI-MEM (Invitrogen) for 0, 12, 24, and 48 h at 37°C prior to harvesting by removing the medium. Total cellular RNA was isolated by lysing cells in TRIzol® reagent (Ambion®) and using with the Total RNA kit I (e.n.z.a.). RNA was resuspended at 60°C in 40 µl of DEPC- treated water (Ambion®). RNA concentration was determined by spectrophotometry using a NanoDrop (Thermo). An aliquot of 10 µg of total RNA was then treated with 2 U TURBO DNase (Ambion®) for 30 min at 37°C, to remove any traces of genomic DNA. An aliquot of 2 µg of DNase-treated RNA was reverse transcribed with the High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems®) according to the manufacturer's instructions. Samples were diluted 10 times and then stored at -20°C.

Real-time PCR assay

Gene expression was quantified by the KAPA SYBR® FAST qPCR kit (Kapabiosystem®) using a Stratagene Mx300P qPCR System. Each reaction was run in duplicate and contained 2 µl of cDNA template along with 200 or 250 nM primers in a final reaction volume of 20 µl. Cycling parameters were 95°C for 10 min to activate DNA polymerase, then 40 cycles of 95°C for 10 sec, primer-specific annealing temperature for 30 sec, and 72°C for 15 sec, followed by a final extension of 7 min at 72°C. Nonspecific signals caused by primer dimers were excluded by dissociation curve analysis and the use of non-template controls. To normalize for cDNA loading, GADPH was used as an internal control, and gene expression was quantified according to the 2-ΔΔCt method.

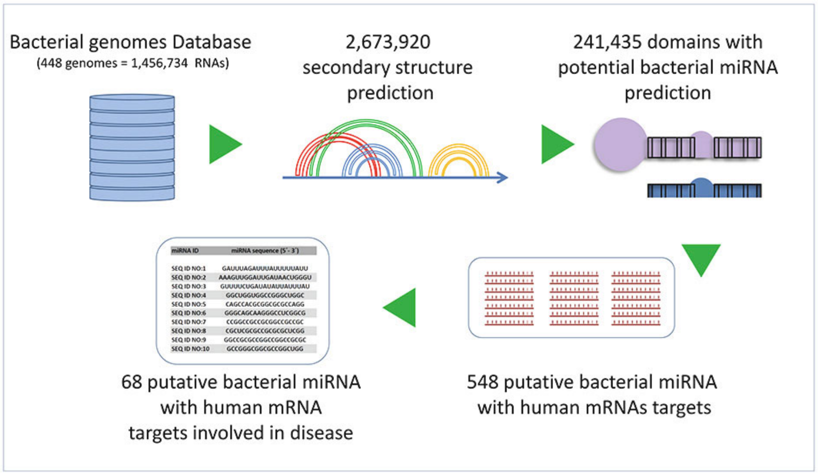
Oligodeoxyribonucleotides used

Sequences of the oligodeoxyribonucleotides used for quantitative real-time PCR were as follows: PTPRJ-F (5′-GTG AAA GCT CTG GAG CCA AC-3′); PTPRJ-F (5′-TGG TTG GAC TGA TGG AAA CA-3′); NFBKBIL1-F (5′-CCA GAT GCC TAC ACC GAT TT-3′); NFBKBIL1-R (5′-TCA CCC TGG AGC TTC TGT CT-3′); DEK-F (5′-ACG GAA CAG TTC TGG AAT GG-3′); DEK-R (5′-TGG TGG CTC CTC TTC ACT TT-3′); GADPH-F (5′-GTC GGA GTC AAC GGA TTT GG-3′); and GADPH-R (5′-CTT GAT TTT GGA GGG ATC TCG-3′). Sequences of the synthetic miRNA used were as follows: PTPRJ-miRNA (5′-GCG GCG GCG GCG GCC GCG GTT-3′), NFKBIL1-miRNA (5′-GAC GUU CUC GGC GGU GGC GTT-3′), DEK-miRNA (5′-GAU UUA GAU UUA UUU UUA UTT-3′) and Control(+)miRNA (5′-UAG UAC UAG UAC UAG UAC UTT-3′).

Results

A method to identify putative miRNAs generated from bacterial RNA

To explore this possibility, we created a bioinformatics pipeline that included a mix of various algorithms, methods, and programs to predict possible miRNAs in bacterial genomes. A summary of the procedure is shown in Fig. 1. The first step was to generate a database (DB.1-PB) that included 448 genomes from nine different bacterial groups (Fig. 2). Of these genomes, 344 were complete and 104 were partial drafts.


Bacterial genomes Database
(448 genomes = 1,456,734 RNAs)
2,673,920
secondary structure
prediction
241,435 domains with
potential bacterial miRNA
prediction
548 putative bacterial miRNA
with human mRNAs targets
68 putative bacterial miRNA
with human mRNA
targets involved in disease

miRNA ID	miRNA sequence (5'- 3')
SEQ ID NO:1	GAUUUAGAUUUAUUUUUAUU
SEQ ID NO:2	AAAGUUGGAUUGAUAACUGGGU
SEQ ID NO:3	GUUUUCUGAUAUAUUAUUUAU
SEQ ID NO:4	GGCUGGUGGCCGGGCUGGC
SEQ ID NO:5	CAGCCACGCGGCGCGCCAGG
SEQ ID NO:6	GGGCAGCAAGGGCCUCGGCG
SEQ ID NO:7	CCGGCCGCCGCGGCCGCCGC
SEQ ID NO:8	CGCUCGCGCCGCGCGCUCGG
SEQ ID NO:9	GGCCGCGCCGGCCGGCCGCGC
SEQ ID NO:10	GCCGGGCGGCGCCGGCUGG

Fig. 1. Procedure of the identification of putative bacterial microRNA.

From the sequence information of GenBank (Benson *et al.*, 2008), 1,456,734 genes were extracted from the above database (DB.1-PB) using the BioPerl software. This information was assembled into a new database named DB.2-BG, which included barcode identification for each gene and its sequence in a FASTA format. The average number of gene sequences in DB.2-BG from each bacterium was 3252 but the variance was quite large. For example, the database contained 475 gene sequences from *Mycoplasma genitalium* G37 and 8702 genes from *Burkholderia xenovorans* LB400 (see Supplementary data Table S1).

Next, we predicted the secondary structure of the potential mRNA of each gene separately, using the Unified Nucleic Acid Folding (UNAFold) software package (Markham and Zuker, 2008). With a total of 100 CPUs and 100 GB RAM, it took about 200 days to calculate the secondary structure of 1,456,734 bacterial genes. The analysis was performed at 37°C while the other parameters were left at the default settings of the software. Several predictions of secondary structure were made for each gene product. All secondary structure predictions were scanned automatically using the BioPerl software to find a wild form similar to those described below for pre-miRNA. The secondary structure predictions were extracted into a new database named DB.3-FCT that included, for each gene, the bacterial name, genomic location, function, gene product sequence, secondary structure, and the coordinates of the fragment with the predicted pre-miRNA secondary structure. In this work, the pre-miRNA secondary structure was calculated using a hairpin secondary structure, a minimum length of 19 nucleotides, and an accepted internal loop with up to 25% more nucleotides than those of the complementary strand (Han *et al.*, 2004; Kurreck, 2006).

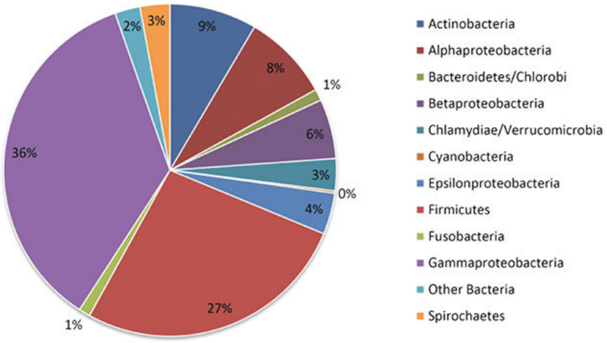


Fig. 2. The bacterial genomes used in this study. A total of 448 bacterial genomes were obtained from the NCBI database. More information can be found in 'Material and Methods'.

Targets and disease relationships

As an extension of the above results, we looked for potential target genes in the human cell. To do this, we first introduced the results of all predictions, previously obtained, into the UNAFold program, which showed the feasibility of over 3000 genes including a secondary structure similar to pre-miRNAs (data not shown). These were jointed into a new database named DB.4-miRNA and processed using the BioPerl software and BLAST program (Altschul *et al.*, 1990) to find any complementary sequence relationship between the predicted miRNAs and any of the 42,753 human mRNAs predicted from the NCBI human genome sequence. A total of 548 miRNA secondary structures were predicted from 122 different pathogenic bacteria, which had complementary sites between a putative miRNA sequence and a human mRNA (Supplementary data Table S2). All of these genes were added to a new database named DB.5-T, together with target gene data. The average quantity of miRNA forecast per organism was 15 miRNA predictions (Fig. 3A). The results showed that 40 bacteria had only one miRNA prediction; the highest yield of miRNA predictions (28 miRNAs) was in *Burkholderia cenocepacia* J2315 and the average number of total miRNA predictions per organism was 15. Statistically, the length range of all miRNAs predicted with a known target was between 19 to 50 nucleotides, while 96.6% of these miRNA predictions had a length of between 19 to 29 nucleotides (Fig. 3B).

Next, we searched the human mRNA targets obtained from DB.5-T in the Online Mendelian Inheritance in Man (OMIM) database (Hamosh *et al.*, 2005), which catalogs all the known diseases with a genetic component and contains information of all the medical studies published for each human gene

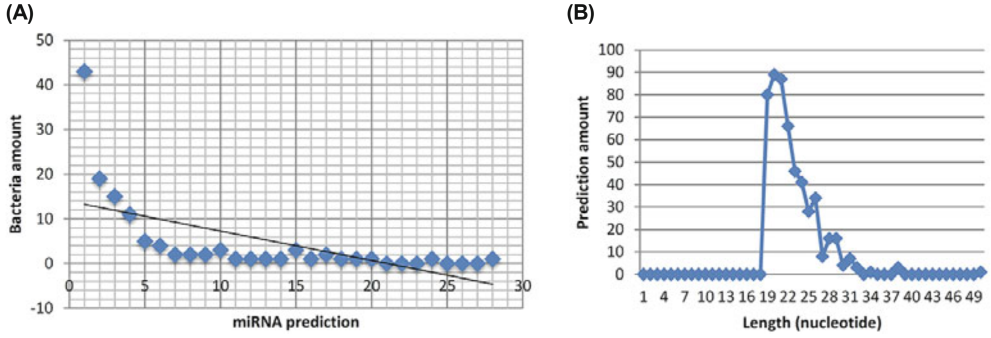


Fig. 3. Number and length of predicted microRNAs (miRNAs). (A) Number of miRNAs predicted versus the number of bacteria. (B) Length of predicted miRNAs. The blue line shows that the length of most of the predicted miRNAs (96.6%) was 19–29 nucleotides (see Supplementary data Table S2).

and the disease involved, including cases of gene mutations. A total of 37 bacteria with 361 predicted miRNAs had potential targets in gene products involved in known human diseases, such as cancer, diabetes, spinal atrophy, and others (Supplementary data Table S3). For example, in the case of *Arcobacter butzleri* RM4018, a member of the epsilon subdivision of the Proteobacteria and a close taxonomic relative of established pathogens, such as *Campylobacter jejuni* and *H. pylori* (Miller *et al.*, 2007), we found a 20 nucleotide miRNA (Fig. 4A) with the targets DEK oncogene (DEK), transcript variant 1, and transcript variant 2 (Fig. 4B and 4C), which encode a protein with one SAP domain. This protein binds to cruciform and super helical DNA and induces positive super coils into closed circular DNA, and is also involved in splice site selection during mRNA processing. Chromosomal aberrations involving this region increase expression of this gene, and the presence of antibodies against this protein is associated with various diseases (von Lindern *et al.*, 1993).

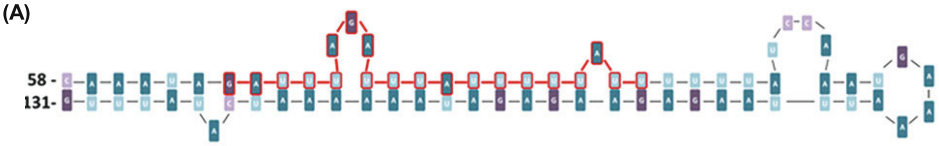
Based on the information in OMIM, we found that the downregulation of these genes may cause acute myeloid leukemia (AML), a cancer of the myeloid line of blood cells, characterized by the rapid growth of abnormal white blood cells that accumulate in the bone marrow and interfere with the production of normal blood cells. AML is the most common acute leukemia affecting adults, and its incidence increases with age (Alexiadis *et al.*, 2000). Additionally, Fu *et al.* (1997) identified DEK as the 43-kD factor that recognizes the TG-rich peri-ets regulatory element in the HIV-2 genome, which is involved in transcriptional regulation and signal transduction. They suggest that their discovery has implications for multiple pathogenic processes, including hematologic malignancies, arthritis, ataxia-telangiectasia, and AIDS.

The effect of bacterially derived putative miRNAs in human cells

To confirm our prediction, specific assays were carried out with bacterial miRNAs that were complementary to mRNAs of the human genes protein tyrosine phosphatase receptor type J (*PTPRJ*), nuclear factor nuclear factor kappa-light-chain-enhancer of activated B cells (NF-kappa-B; *NFKBIL1*), and *DEK* oncogene variants 1 and 2. For this, human embryonic kidney (HEK293) cells were transfected with synthetic miRNA derived from *Burkholderia vietnamiensis G4*, *Burkholderia mallei***,** and *A. butzleri.* After 0, 12, 24, and 48 h, the cells were collected, total RNA was prepared, and total complementary DNA (cDNA) was synthesized. Changes in the expression of human target mRNAs (downregulation) was measured by real-time PCR using primers specific for the above mRNAs, in all three cases.

A first assay was carried out with a putative miRNA from *B. vietnamiensis G4* by measuring the amount of the *PTPRJ* transcript variant 2 mRNA in human cells (Fig. 5). The results showed that the putative miRNA sequence generated from *B. vietnamiensis G4* significantly reduced the expre-

(A)



(B)

```
Target 1 (2838)   5´- AATAAAAATAAATCTAAATC-ACTGGTGTTTTAAAAAAAAA...... - 3´
                      ||||||||||||||||||||
miRNA              3´- TTATTTTTATTTAGATTTAG- 5´
```

(C)

```
Target 2 (2736)   5´- AATAAAAATAAATCTAAATC-ACTGGTGTTTTAAAAAAAAA...... - 3´
                      ||||||||||||||||||||
miRNA              3´- TTATTTTTATTTAGATTTAG- 5´
```

Fig. 4. ***Arcobacter butzleri*** **RM4018 microRNA (miRNA) and its target.** (A) The predicted structure of the pre-miRNA generated from *A. butzleri* RM4018, with a length of 31 nucleotides. The core of the miRNA is shown in red. The numbers (left side) indicate the initial and end nucleotides of the miRNA in the gene. (B) The miRNA sequence is complementary to target gene DEK transcript variant 1, starting from base 2838. (C) The same miRNA sequence is complementary to a second target gene, DEK transcript variant 2, starting from base 2736.

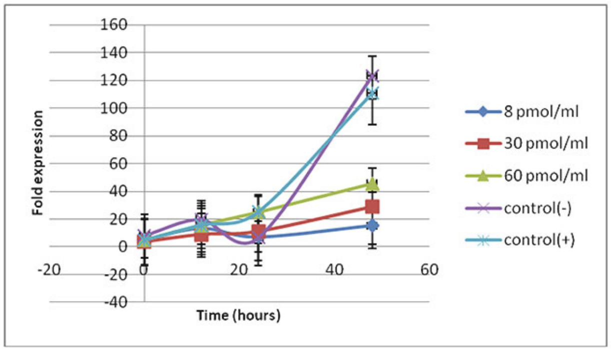


Fig. 5. Bacterially derived microRNA regulate the *PTPRJ* gene expression: Changes in the synthesis of human protein tyrosine phosphatase receptor type J (*PTPRJ*) messenger RNA (mRNA), after transfection with three different concentrations of *Burkholderia vietnamiensis G4* synthetic microRNA. The target *PTPRJ* gene was repressed by microRNA. The control(-) plate had no oligonucleotide and miRNA control(+) plate had mRNAs not targeted by putative miRNA. The error bars were derived from 3 independent experiments.

ssion of its target *PTPRJ* gene. The protein encoded by this gene is a member of the protein tyrosine phosphatase (PTP) family. PTPs are known to be signaling molecules that regulate a variety of cellular processes, including cell growth, differentiation, mitotic cycle, and oncogenic transformation (Honda *et al*., 1994). The protein PTPRJ is a tumor suppressor gene that has been implicated in a range of cancers, including colon cancer and breast cancer (Ruivenkamp *et al*., 2002). We suggest that reducing *PTPRJ* mRNA expression by putative bacterial miRNA may increase the risk of developing tumor cells.

A second assay was carried out by measuring the effect of a putative miRNA from *B. mallei* on the expression of *NFKBIL1* mRNA in human cells (Fig. 6). The results showed that the predicted putative miRNA sequence from *B. mallei* downregulated, by up to five-fold, its target *NFKBIL1* mRNA. NFKBIL1 is part of the NFKB complex. The activated NFKB complex translocates into the nucleus and binds to DNA at kappa-B-binding motifs. *NFKBIL1* gene mutations have been shown to cause rheumatoid arthritis (Okamoto *et al*., 2003). In additional, the NFKB complex is involved in cellular responses to stimuli such as stress, and bacterial or viral antigens (Gilmore, 1999). Given the above, we can suggest that disruption of this system facilitates pathogen penetration and general defense suppression.

A third assay was carried out by measuring the effect of a putative miRNA from *A. butzleri* in modifying the expression of *DEK* oncogene mRNA in human cells (Fig. 7). The results showed that the putative miRNA sequence generated from *A. butzleri* bacteria strongly reduced the expression of its target DEK mRNA. *DEK* is a highly abundant, evolutionarily conserved, and ubiquitous nuclear protein that can be regulated at the level of transcription and post-translationally. Evidence suggests that it may function as a nuclear architectural protein. Many studies support distinct intracellular functions for DEK in DNA replication, positive and negative regulation of gene transcription, histone acetylation, mRNA splicing, and nucleosome assembly. Several reports have already suggested that DEK may modulate genome stability. Undoubtedly, the silencing of this protein can cause abundant mutations and damage in the cell (Kavanaugh *et al*., 2011).

Discussion

We have developed a novel bioinformatics procedure able to analyze 448 known pathogen bacterial genomes and identify the preferred secondary structures of all of the bacterial transcriptomes. From these secondary structures, we identified all double-stranded regions that, through processing, could give rise to putative bacterial miRNAs. Then, we analyzed all human mRNA sequences to identify those with an exact sequence complementary to the putative bacterial miRNAs. We found 548 miRNA secondary structure predictions (Supplementary data Table S2) of 122 different patho-

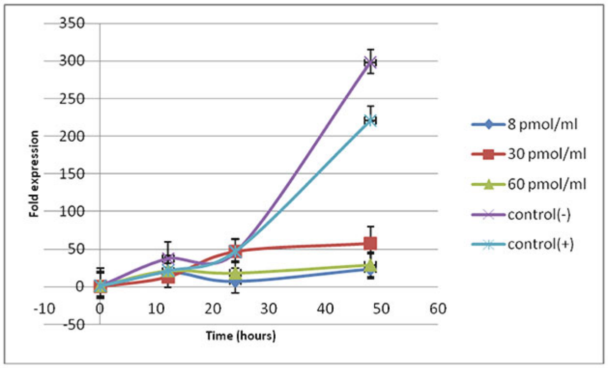


Fig. 6. Bacterially derived microRNA regulate the *NFKBIL1* gene expression: Changes in the synthesis of *NFKBIL1* messenger RNA, after transfection with three different concentrations of *Burkholderia mallei* synthetic microRNA. The target *NFKBIL1* gene was repressed by microRNA. The control(-) plate had no oligonucleotide and miRNA control(+) plate had mRNAs not targeted by putative miRNA. The error bars were derived from 3 independent experiments.

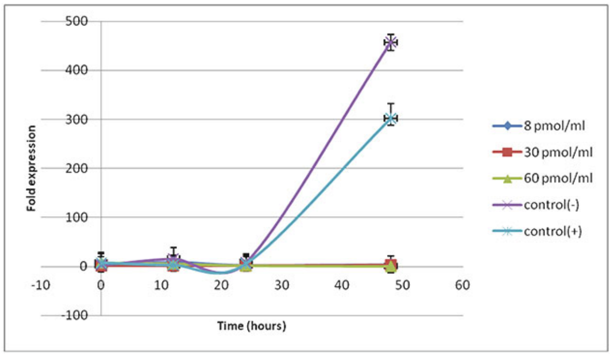


Fig. 7. Bacterially derived microRNA regulate the *DEK* gene expression: Changes in the synthesis of human *DEK* oncogene messenger RNA variants 1 and 2, after transfection with three different concentrations of *Arcobacter butzleri* synthetic microRNA. The target *DEK* gene was repressed by microRNA. The control(-) plate had no oligonucleotide and miRNA control(+) plate had mRNAs not targeted by putative miRNA. The error bars were derived from 3 independent experiments.

genic bacteria with a length range of between 19 to 50 nucleotides and presenting complementary sites between miRNA sequence and human mRNA. Additionally, our results show that 68 miRNA predictions (Supplementary data Table S3) from 37 different bacterial genomes may downregulate gene expression of human mRNA targets known to be involved in 47 different human diseases, such as leukemia (eight predictions), diabetes (two predictions), colon cancer (two predictions), and other diseases including Diamond-blackfan anemia, renal glucosuria, retinal degeneration, rheumatoid arthritis, Joubert syndrome, pseudoxanthoma elasticum, and epilepsy. This bacterial miRNA binds to its mRNA targets, causing their degradation by the host RNase H. It is important to emphasize that the predictions of miRNA structure may not be accurate due to tow main reasons. First, RNA folding varies according to a wide variety of parameters such as temperature, salt concentration, and the presence of chaperones, which affect directly the accuracy of the miRNA secondary structure (Kurreck, 2006). The second reason is that studies comparing the diversity of miRNA indicate that pre-RNA secondary structure can affect the efficiency of an miRNA (Kurreck, 2006). Additionally, RNA silencing in response to bacterial infection, as we hypothesize, could occurs only if there are evidences of the existence of specific bacterial RNA containing pre-RNA secondary structure, which are transferred or expressed with sufficient copy amount in their precursor to repress the host gene target expression.

To evaluate our predictions, we transfected human cells with three different oligonucleotides corresponding to bacterial miRNA sequences, which were complementary to the mRNA of the human genes *PTPRJ*, *NFKBIL1*, and *DEK* oncogene variants 1 and 2. In all three cases, the results indicate significant reduction in target gene expression. Then, we hypothesized that changes in the expression of human target *PTPRJ* mRNA, a tumor suppressor gene that has been implicated in a range of cancers, induced by the putative miRNA from *B. vietnamiensis G4* might participate in the generation of tumor cells. We note that *B. vietnamiensis G4* uses the type IV secretion system to translocate DNA and protein macromolecules to a diverse range of bacterial and eukaryotic cells (Zhang *et al.*, 2009). Our results suggest that sequences derived from bacterial RNA can downregulate specific human target genes. This regulation may affect processes such as metabolism, cell division, etc., of the human cell, which may be causes of different diseases. The results of this research indicate the possibility of studying in more detail the molecular basis of bacterial pathogenicity.

Genetic transfer can occur during human cell infection by intercellular bacteria, the ingestion of bacteria by macrophages, or via a vesicular system to transfer genes from its property cell. Previous studies have examined gene transfer from bacteria to humans that would result in vertical inheritance (Lander *et al.*, 2001; Huerta-Cepas *et al.*, 2007). One of the best studied examples of genetic transfer from bacteria to eukaryotes is genetic transfer to plants from the bacteria *Agrobacterium tumefaciens*. This bacterium uses a type IV secretion system to inject a tumor-inducing plasmid into plant cells. Through illegitimate recombination, the plasmid integrates into the plant genome, and plasmid-encoded transcripts are produced using endogenous promoters, creating a tumor environment that promotes the bacterium's own growth (Talya *et al.*, 2001).

Non-coding RNAs (ncRNAs) may play an important role in the regulation of gene expression, as well as function as structural elements in a cell. In the past few years, an increasing number of small ncRNAs have been identified in many different organisms, ranging from prokaryotes to more complex eukaryotes (Huttenhofer *et al.*, 2001; Markeret *et al.*, 2002). In *Escherichia coli*, more than 100 small ncRNAs have already been identified (Saetrom *et al.*, 2005). The broad phylogenies and high expression levels of small ncRNAs suggest that they may be carrying out important tasks within a given organism. Given their versatile roles in transcriptional and translational control of gene expression and in quality control of macromolecular products, it is suggested that the study of these predicted miRNAs will yield important clues in the future as to how the processes of the human host cell can facilitate human diseases like cancer in response to changing bacterial environments.

Acknowledgements

Work supported by Fondecyt 3110015, Conicyt Basal CCTE PFB-16 and Fundación Ciencia & Vida. We thank Cristina Padilla, and Paulina Saavedra for their excellent technical assistance.

References

Akira, S. and Takeda, K. 2004. Toll-like receptor signalling. *Nat. Rev. Immunol.* **4**, 499–511.

Alexiadis, V., Waldmann, T., Andersen, J., Mann, M., Knippers, R., and Gruss, C. 2000. The protein encoded by the proto-oncogene DEK changes the topology of chromatin and reduces the efficiency of DNA replication in a chromatin-specific manner. *Genes Dev.* **14**, 1308–1312.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Ambros, V. 2004. The functions of animal microRNAs. *Nature* **431**, 350–355.

Azcarate-Peril, M.A., Sikes, M., and Bruno-Barcena, J.M. 2011. The intestinal microbiota, gastrointestinal environment and colorectal cancer: a putative role for probiotics in prevention of colorectal cancer? *Am. J. Physiol. Gastrointest. Liver Physiol.* **301**, G401–424.

Baulcombe, D. 2004. RNA silencing in plants. *Nature* **431**, 356–363.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. 2008. GenBank. *Nucleic Acids Res.* **36**, D25–30.

Cai, X., Lu, S., Zhang, Z., Gonzalez, C.M., Damania, B., and Cullen, B.R. 2005. Kaposi's sarcoma-associated herpesvirus expresses an array of viral microRNAs in latently infected cells. *Proc. Natl. Acad. Sci. USA* **102**, 5570–5575.

Eccleston, K., Collins, L., and Higgins, S.P. 2008. Primary syphilis. *Int. J. STD AIDS* **19**, 145–151.

Ebby, O.L. 2005. Community-acquired pneumonia: from common pathogens to emerging resistance. *Emerg. Med. Pract.* v7n12.

Farrar, J.J., Yen, L.M., Cook, T., Fairweather, N., Binh, N., Parry, J., and Parry, C.M. 2000. Tetanus. *J. Neurol. Neurosurg. Psychiatry* **69**, 292–301.

Fire, A., Xu, S.Q., Montgomery, M.K., Kostas, S.A., Driver, S.E.,

and Mello, C.C. 1998. Potent and specific genetic interference by double-stranded RNA in *C. elegans*. *Nature* **391**, 806–811.

Freeman, V.J. 1951. Studies on the virulence of bacteriophage-infected strains of *Corynebacterium diphtheriae*. *J. Bacteriol.* **61**, 675–688.

Fu, G.K., Grosveld, G., and Markovitz, D.M. 1997. DEK, an autoantigen involved in a chromosomal translocation in acute myelogenous leukemia, binds to the HIV-2 enhancer. *Proc. Natl. Acad. Sci. USA* **94**, 1811–1815.

Giannella, R.A. 1996. *Salmonella*, chap. 21 pp. 295–302. *In* Baron, S. (ed.), Medical Microbiology, 4th ed. University of Texas Medical Branch. Galveston, Tx., USA.

Gilmore, T.D. 1999. The Rel/NF-κB signal transduction pathway: Introduction. *Oncogene* **18**, 6842–6844.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–517.

Han, J., Lee, Y., Yeom, K.H., Kim, Y.K., Jin, H., and Kim, V.N. 2004 The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev.* **18**, 3016–3027.

Honda, H., Inazawa, J., Nishida, J., Yazaki, Y., and Hirai, H. 1994. Molecular cloning, characterization, and chromosomal localization of a novel protein-tyrosine phosphatase, HPTP eta. *Blood* **84**, 4186–4194.

Huerta-Cepas, J., Dopazo, H., Dopazo, J., and Gabaldon, T. 2007. The human phylome. *Genome Biol.* **8**, R109.

Huttenhofer, A., Kiefmann, M., Meier-Ewert, S., O'Brien, J., Lehrach, H., Bachellerie, J.P., and Brosius, J. 2001. RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBRO J.* **20**, 2943–2953.

Kavanaugh, G.M., Wise-Draper, T.M., Morreale, R.J., Morrison, M.A., Gole, B., Schwemberger, S., Tichy, E.D., Lu, L., Babcock, G.F., Wells, J.M., and *et al.* 2011. The human DEK oncogene regulates DNA damage response signaling and repair. *Nucleic Acids Res.* **39**, 7465–7476.

Kostic, A.D., Gevers, D., Pedamallu, C.S., Michaud, M., and Duke, F. 2012. Genomic analysis identifies association of *Fusobacterium* with colorectal carcinoma. *Genome Res.* **22**, 292–298.

Kurreck, J. 2006. siRNA efficiency: structure or sequence–That is the question. *J. Biomed. Biotechnol.* **83757**, 1–7.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., and Zody, M.C. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Marchesi, J.R., Dutilh, B.E., Hall, N., Peters, W.H.M., Roelofs, R., Boleij, A., and Tjalsma, H. 2011. Towards the human colorectal cancer microbiome. *PLoS ONE* **6**, e20447.

Marker, C., Zemann, A., Terhörst, T., Kiefmann, M.J.P., Kastenmayer Green, P., Bachellerie-Brosius, J.P., and Huttenhofer, A. 2002. Experimental RNomics: identification of 140 candidates for small non-messenger RNAs in the plant *Arabidopsis thaliana*. *Curr. Biol.* **12**, 2002–2013.

Markham, N.R. and Zuker, M. 2008. UNAFold: software for nucleic acid folding and hybridization, chap. 1, pp. 3–31. *In* Keith, J.M. (ed), Bioinformatics, volume II. Structure, function and applications, number 453 in Methods in Molecular Biology. Human Press. Totowa, N.J., USA.

Marshall, B.J. and Warren, J.R. 1983. Unidentified curved bacillus on gastric epithelium in active chronic gastritis. *Lancet* **1**, 1273–1275.

Marshall, B.J. and Warren, J.R. 1984. Unidentified curved bacilli in the stomach patients with gastritis and peptic ulceration. *Lancet* **1**, 1311–1315.

Miller, W.G., Parker, C.T., Rubenfield, M., Mendz, G.L., and Wösten, M.M.S.M. 2007. The complete genome sequence and analysis of the Epsilonproteobacterium *Arcobacter butzleri*. *PLoS ONE* **2**, e1358.

Naimi, T.S., Wicklund, J.H., Olsen, S.J., Krause, G., Wells, J.G., Bartkus, J.M., Boxrud, D.J., Sullivan, M., Kassenborg, H., Besser, J.M., and *et al.* 2003. Concurrent outbreaks of *Shigella sonnei* and enterotoxigenic *Escherichia coli* infections associated with parsley: implications for surveillance and control of foodborne illness. *J. Food Prot.* **66**, 535–541.

Okamoto, K., Makino, S., Yoshikawa, Y., Takaki, A., Nagatsuka, Y., Ota, M., Tamiya, G., Kimura, A., Bahram, S., and Inoko, H. 2003. Identification of I-kappa-BL as the second major histocompatibility complex-linked susceptibility locus for rheumatoid arthritis. *Am. J. Hum. Genet.* **72**, 303–312.

Pfeffer, S., Sewer, A., Lagos-Quintana, M., Sheridan, R., Sander, C., Grasser, F.A., van Dyk, L.F., Ho, C.K., Shuman, S., and Chien, M. 2005. Identification of microRNAs of the herpesvirus family. *Nat. Methods* **2**, 269–276.

Pfeffer, S., Zavolan, M., Grässer, F.A., Chien, M., Russo, J.J., Ju, J., John, B., Enright, A.J., Marks, D., Sander, C., and *et al.* 2004. Identification of virus-encoded microRNAs. *Science* **304**, 734–736.

Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. 2000. The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* **403**, 901–906.

Riley, D.R., Sieber, K.B., Robinson, K.M., White, J.R., Ganesan, A., Nourbakhsh, S., and Dunning Hotopp, J.C. 2013. Bacteria-human somatic cell lateral gene transfer is enriched in cancer samples. *PLoS Comput. Biol.* **9**, e1003107.

Ruivenkamp, C.A., Wezel, T., Zanon, C., Stassen, A.P.M., Vlcek, C., Csikos, T., Klous, A.M., Tripodis, N., Perrakis, A., Boerrigter, L., and *et al.* 2002. Ptprj is a candidate for the mouse colon-cancer susceptibility locus Scc1 and is frequently deleted in human cancers. *Nat. Genet.* **31**, 295–300.

Saetrom, P., Sneve, R., Kristiansen, K.I., Snove, O.Jr., Grunfeld, T., Rognes, T., and Seeberg, E. 2005. Predicting non-coding RNA genes in *Escherichia coli* with boosted genetic programming. *Nucleic Acids Res.* **33**, 3263–3270.

Samols, M.A., Hu, J., Skalsky, R.L., and Renne, R. 2005. Cloning and identification of a microRNA cluster within the latency-associated region of Kaposi's sarcoma-associated herpesvirus. *J. Virol.* **79**, 9301–9305.

Sasaki, S., Takeshita, F., Okuda, K., and Ishii, N. 2001. *Mycobacterium leprae* and leprosy: a compendium. *Microbiol. Immunol.* **45**, 729–736.

Sullivan, C.S., Grundhoff, A.T., Tevethia, S., Pipas, J.M., and Ganem, D. 2005. SV40-encoded microRNAs regulate viral gene expression and reduce susceptibility to cytotoxic T cells. *Nature* **435**, 682–686.

Talya, K., Tzvi, T., Yoram, K., Yedidya, G., Colin, D., and Vitaly, C. 2001. Genetic transformation of HeLa cells by *Agrobacterium*. *Proc. Natl. Acad. Sci. USA* **98**, 1871–1876.

Voinnet, O. 2005. Induction and suppression of RNA silencing: insights from viral infections. *Nat. Rev. Genet.* **6**, 206–220.

Von Lindern, M., Fornerod, M., and Soekarman, N. 1993. Translocation t(6;9) in acute non-lymphocytic leukaemia results in the formation of a DEK-CAN fusion gene. *Baillieres Clin. Haematol.* **5**, 857–879.

Wightman, B., Ha, I., and Ruvkun, G. 1993. Posttranscriptional regulation of the heterochronic gene lin-14 mediates temporal pattern formation in *C. elegans*. *Cell* **75**, 855–862.

Zhang, R., LiPuma, J.J., and Gonzalez, C.F. 2009. Two type IV secretion systems with different functions in *Burkholderia cenocepacia* K56-2. *Microbiology* **155**, 4005–4013.