Published online 29 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D233–D238***
doi:10.1093/nar/gks1264

BSRD: a repository for bacterial small regulatory RNA

Lei Li, Dandan Huang, Man Kit Cheung, Wenyan Nong, Qianli Huang and Hoi Shan Kwan*

Biology Programme, School of Life Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China

Received August 12, 2012; Revised September 28, 2012; Accepted November 5, 2012

ABSTRACT

In bacteria, small regulatory non-coding RNAs (sRNAs) are the most abundant class of post-transcriptional regulators. They are involved in diverse processes including quorum sensing, stress response, virulence and carbon metabolism. Recent developments in high-throughput techniques, such as genomic tiling arrays and RNA-Seq, have allowed efficient detection and characterization of bacterial sRNAs. However, a comprehensive repository to host sRNAs and their annotations is not available. Existing databases suffer from a limited number of bacterial species or sRNAs included. In addition, these databases do not have tools to integrate or analyse high-throughput sequencing data. Here, we have developed BSRD (http://kwanlab.bio.cuhk.edu.hk/BSRD), a comprehensive bacterial sRNAs database, as a repository for published bacterial sRNA sequences with annotations and expression profiles. BSRD contains over nine times more experimentally validated sRNAs than any other available databases. BSRD also provides combinatorial regulatory networks of transcription factors and sRNAs with their common targets. We have built and implemented in BSRD a novel RNA-Seq analysis platform, sRNADeep, to characterize sRNAs in large-scale transcriptome sequencing projects. We will update BSRD regularly.

INTRODUCTION

Small regulatory RNAs (sRNAs) in bacteria are a class of non-coding RNA genes. They are usually 50–500 bp long and encoded in an estimated amount of ~200–300 copies in a typical bacterial genome (1). sRNAs can be categorized as *cis*-encoded antisense sRNAs, which are completely complementary to their targets, and *trans*-encoded antisense sRNAs, which are only partially complementary to their targets, with binding facilitated by the RNA-binding protein Hfq (2). sRNAs are important post-transcriptional regulators, they can either inhibit translation of mRNAs by degrading the mRNAs or masking the ribosome binding sites, or activate the translation by opening the ribosome binding sites or increasing mRNA stability (3). They are involved in many crucial cellular processes, including biofilm formation (4) and quorum sensing (5). sRNAs may directly or indirectly regulate most bacterial genes (6).

Since the first discovery of the chromosome-encoded sRNA regulator, MicF, in *Escherichia coli* (7), detection of sRNAs has been hampered by traditional genetic screening methods because of their relatively small size, non–protein-coding nature and locating in intergenic regions (8). As a result, only a limited amount of sRNAs has been identified. Recent advance in computational methods and high-throughput techniques such as genomic tiling microarrays and deep sequencing has discovered many sRNAs and provided invaluable insights into the detection and characterization of bacterial sRNAs.

Although sRNAs play crucial regulatory roles and their discovery has been greatly facilitated in recent years, the quantity and quality of currently available sRNA databases are far from desirable. Databases such as RegulonDB (9) and Ecocyc (10) include only sRNAs in the *E. coli* K12 strain. Rfam (11), which is a database of structural RNA families, contains only a few bacterial sRNA families. sRNAMap (12) contains only data from gram-negative strains and lacks current updates. Other databases also focused on information on sRNA targets, however, whereas sRNATarbase (13) contains exclusively experimentally validated sRNA targets; RNApredator (14) and sRNATarget (15) are solely based on computational predictions. Furthermore, annotations of sRNAs in most of these databases are not up-to-date, and these databases can neither integrate nor analyse next-generation sequencing data. Therefore, a repository that collects sequences of all

*To whom correspondence should be addressed. Tel: +852 3943 6285; Fax: +852 3943 1146; Email: hoishankwan@cuhk.edu.hk

published sRNAs and with information such as their annotations and expression profiles is needed.

Here, we present BSRD, a comprehensive bacterial sRNAs database, to serve as a repository for bacterial sRNA sequences and their annotations and expression profiles. In addition to sRNAs annotated in the public databases, we also include in BSRD manually curated bacterial sRNAs and their annotations from the literature. Besides identification of sRNAs, BSRD also provides extensive information on functional characterizations and expression profiles of sRNAs. Furthermore, we have developed and implemented in BSRD a new RNA-Seq analysis platform, sRNADeep, for characterization of sRNAs from high-throughput deep sequencing data.

DATA COLLECTION AND CURATION

Acquisition of sRNA sequences

BSRD contains three kinds of sRNA sequences grouped according to their discovery methods: (i) by experimental validation, (ii) by sequence and structural conservation, and (iii) by RNA-Seq or tiling microarray experiments. We have obtained 79 and 87 experimentally validated sRNAs from RegulonDB and sRNAMap, respectively. By literature mining, we first obtained a bacterial strain list from the NCBI taxonomy and searched the NCBI PubMed database using the keyword 'sRNA and strain name' with the PubCrawler program (16). sRNA information was then extracted manually from all of the resulting 445 relevant articles. These include sRNA name or alias, species, physical position, strand, identification method, growth phase, Gene Expression Omnibus accession, target genes and regulation effect, and regulators. sRNAs will be regarded as experimentally validated if they are identified by either Northern blot or reverse transcription polymerase chain reaction. Finally, 964 experimentally validated sRNAs were retrieved and added to BSRD.

A total of 6266 sRNA homologs were collected from the Rfam database. In addition, 2334 bacteria genes annotated as 'ncRNA' or 'antisense RNA' in the NCBI Gene database were also collected. An additional 310 sRNAs from sRNAMap were also retrieved. As a result, a total of 8248 non-redundant sRNA homologs were added to BSRD. We have also obtained and added to BSRD 507 candidate sRNAs identified from high-throughput sequencing datasets. These sRNAs display either differential expression in various conditions or a high expression in a single condition. However, as the current computational prediction method for novel sRNAs is of low precision (6–12%) and sensitivity (20–49%) (17), datasets solely predicted *in silico* were not included in BSRD. In addition, 20 115 bacteria regulatory elements were also integrated.

For sRNAs found in multiple resources, exact duplicate hits were merged, but we kept others, which need to be further verified by rapid amplification of cDNA ends (RACE) or other experimental techniques. In BSRD, a new sRNA nomenclature system modified from Chen's system (18) is used: a sRNA is indicated by an initial 's', which stands for small RNA, followed by a three-letter genome ID used in the KEGG database, and a number that indicates its genomic location. We also add an ending number that indicates the number of sRNAs identified in this location.

Functional annotations of identified sRNAs

In BSRD, each sRNA entry contains seven sections of descriptions: Basic Info, UCSC Browser, Secondary Structure, Expression Profile, Target Info, Wikipedia and Other Links (Figure 1). The 'Basic Info' section provides sRNA sequence information and information such as identification method, terminators, Hfq binding and growth phase. Positions of sRNAs could be graphically visualized with the popular UCSC Archaeal Genome Browser (19) implemented in BSRD. Secondary structures of sRNAs are visualized by the RNAfold (20) and Mfold (21) programs. The 'Expression Profile' section provides expression evidence of sRNAs in different experimental conditions collected from the NCBI Gene Expression Omnibus (22) database. sRNA pathogenesis profiling data obtained by the recently emerging Tn-Seq approach (23) is also included.

Identification of sRNA targets is an initial step to understand the regulatory function of sRNAs. We have acquired 138 sRNA-target interactions from sRNATa rBase and manually curated 56 new sRNA-target interactions from the literatures. The sRNA-target interactions were then combined with transcription factor-target interactions to form the regulatory networks. Sigma factors, which act as upstream regulators to regulate sRNA transcription (24), were also added into the networks. Moreover, target genes of identified sRNAs predicted using IntaRNA (25) and RNAplex (26) were also provided.

As the Wikipedia-based community annotation platform has been successful in Rfam and miRBase (27), the same platform is also implemented in BSRD. Wikipedia pages for all sRNA entries have been reviewed manually to avoid vandalism before implementing into BSRD. As most sRNAs still do not have annotation pages in Wikipedia, a link to a brief guide for creating and editing a new Wikipedia page is also provided. Finally, we have provided cross-links to a selected list of external databases, including Rfam, the Gene Ontology (28), Sequence Ontology (29), RegulonDB, EchoBASE (30), EcoGene (31) and EcoCyc, for access to additional information of the sRNAs.

BSRD INTERFACE AND FUNCTIONALITIES

There are nine sections in the BSRD main menu: Home, Search BSRD, Hierarchical taxonomy, Regulatory network, BLAST BSRD, Download, sRNADeep, Submission and Latest publications. From the 'Home' page, a summary of numbers of sRNAs archived in the latest version of the database is available in the 'Current release' section. BSRD hosts 9579 sRNA entries from 957 bacterial strains. Answers to the most frequently asked questions are provided in the 'FAQ' section, and a help

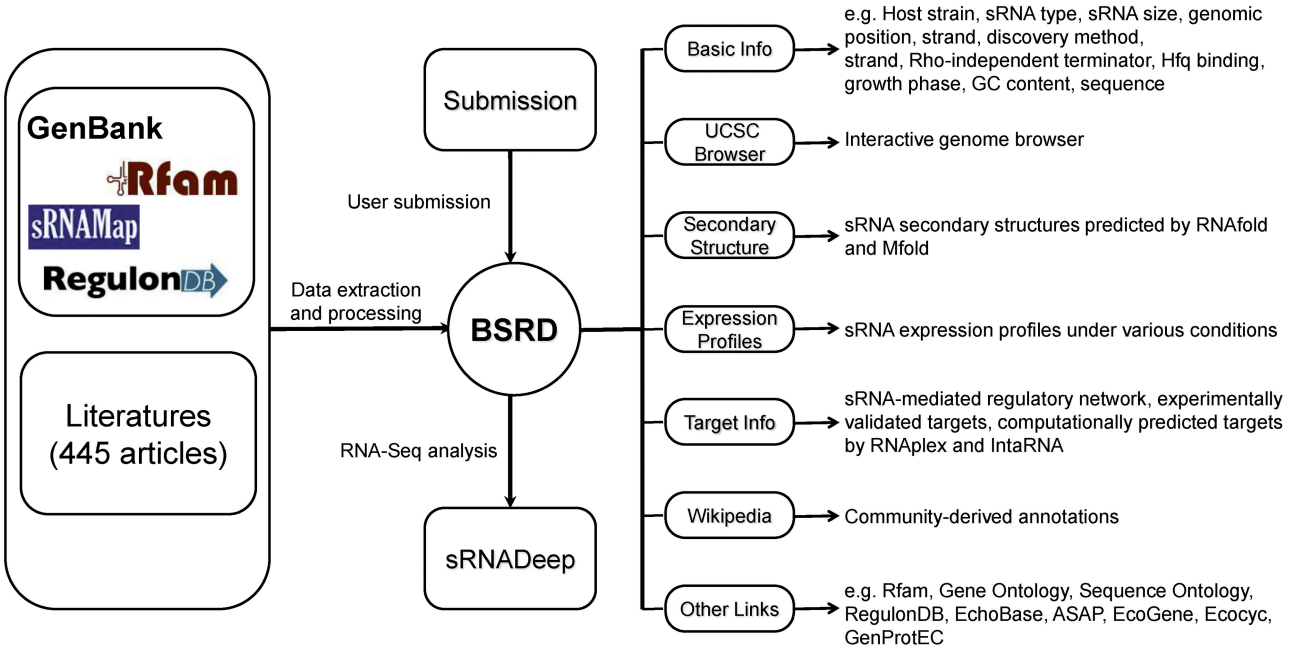

Figure 1. Overview of BSRD design. The three main characteristics of BSRD are (i) comprehensive data collection from external databases and the literatures, (ii) comprehensive annotation and expression profiles for sRNAs and (iii) a novel RNA-Seq analysis platform, sRNADeep, for characterizing sRNAs from high-throughput sequencing data.

documentation is also included in the 'Help' section. The 'Latest Update' section provides news about recent updates of the database.

From the 'Search BSRD' page, users can search for sRNAs in BSRD with three options: by sRNA id or name, by sRNA class or by genomic position. Alternatively, users could examine sRNAs according to the host organism from the list of bacteria in the 'Hierarchical taxonomy' page. Additionally, users could go to the 'BLAST BSRD' page for direct input or upload of sRNA sequences to do quick search against BSRD using the BLAST program. Results will be sorted by alignment scores with single nucleotide variations highlighted (Supplementary Figure S1). The 'Download' page provides options for users to download sRNA sequences in FASTA format according to the sRNA id or name, the bacterial host or batch. The 'Submit' page allows users to submit new sRNAs or annotations to BSRD. The 'Latest publications' page enables users to access the latest articles related to 'bacterial sRNAs' in PubMed and Microsoft Academic Research Databases.

Regulatory network

Transcription factor and sRNA can bind to the same target, whereas the clearance rates, steady-state concentrations and response curves can determine the dynamics of these regulatory networks (32). Regulatory networks in BSRD are constructed using cytoscape web (33) (Figure 2). Different colours were assigned for different elements of the networks: sRNA (yellow), target gene (orange), sigma factor (red) and transcription factor (blue). Regulatory relationships were also differentiated with different line patterns: repression (T-shaped) and inducement (Arrow).

sRNADeep

sRNADeep is a novel platform for sRNA expression profiling from RNA-Seq data (Supplementary Figure S2). It can not only annotate expressed sRNAs from a single set of transcriptome data, but also identify differentially expressed sRNAs from two different conditions sRNADeep accepts compressed clean reads archives, which will then be mapped against the non-redundant sRNA set in BSRD using Burrows-Wheeler Alignment (BWA) (34), with a maximum of one mismatch allowed. Clean reads means filtered raw reads after adapter removal and quality trimming. The expectation maximization-based SEQ-EM algorithm (35) is used to handle multi-mapped reads. For a single dataset, the number of reads for each sRNA will be calculated and normalization will be performed using the reads per kilo-base per million method (36). For analysis of two datasets, DESeq (37) will be used to identify differentially expressed sRNAs between the samples.

On job submission, users should provide a valid email address for receiving a job ID, which sRNADeep assigns, for result retrieval. A typical output of sRNADeep includes the length distribution of clean reads, the distribution of mapped reads and expression levels of sRNAs or differentially expressed sRNAs (Supplementary Figure S3).

DISCUSSION

Compared with other currently available sRNA-related databases, BSRD is more advanced in three aspects. First, BSRD hosts the largest collection of sRNAs (Table 1). It encompasses 964 experimentally validated sRNAs, 8248 sRNA homologs and 507 candidate sRNAs from high-throughput datasets. sRNAMap, for

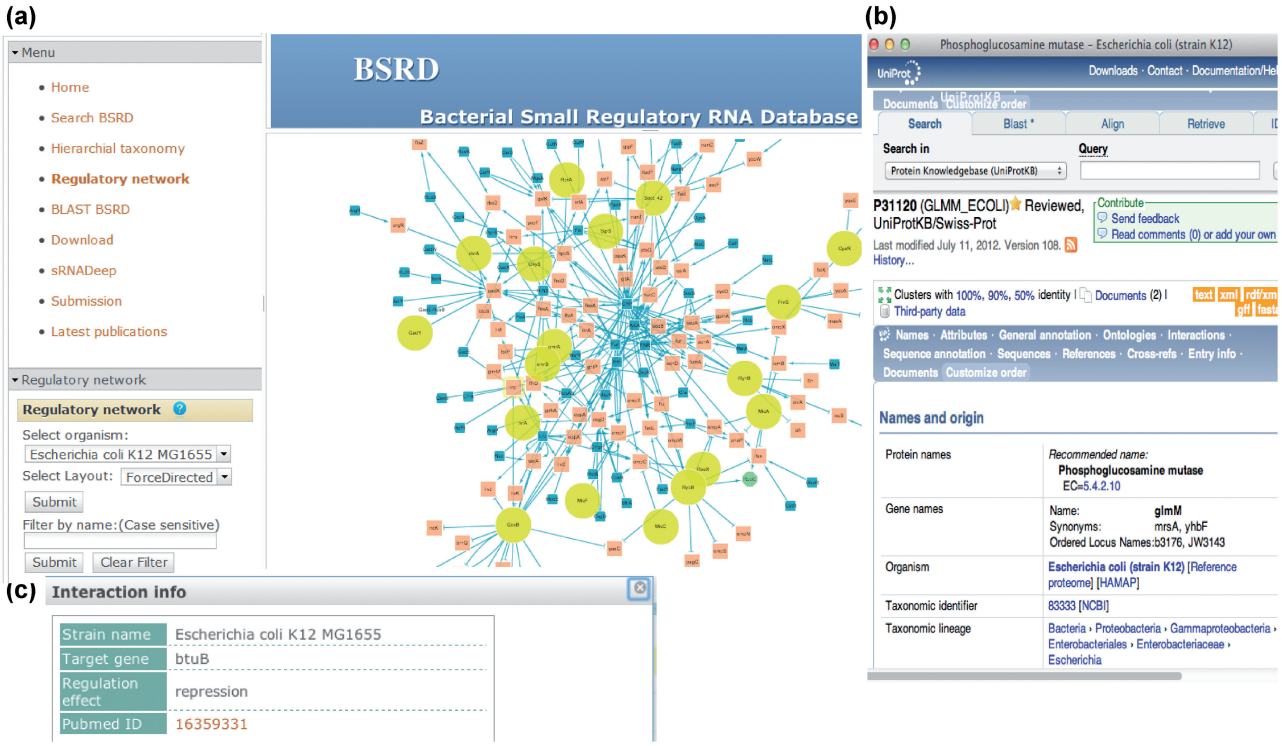

Figure 2. Screenshots of the regulatory network page. (**a**) Overview of the regulatory network. (**b**) Each target gene in the network could be linked out to the UniProt database. (**c**) When clicking on relationship lines, the description of regulatory effects will be shown.

Table 1. Comparison of BSRD with other available resources

	RegulonDB	sRNAMap	Wikipedia	Rfam	BSRD
No. of experimentally validated sRNAs	79	87	99	–	964
No. of sRNA homologs	–	310		6266	8248
No. of sRNA-target interactions	26	60	–	–	194
No. of genomes	1	70	–	–	957
Growth phase	–	–	–	–	373
Secondary structure of sRNAs	–	Yes	–	Yes	Yes
Expression profiles supported	–	Yes	–	–	Yes
Fitness of sRNAs	–	–	–	–	Yes
Sequence homology search	–	Yes	–	Yes	Yes
Wikipedia-derived community annotation	–	–	Yes	Yes	Yes
Deep sequencing read analysis supported	–	–	–	–	Yes
Computational interactions of sRNAs and targets	–	–	–	–	Yes
Regulatory network	–	Yes	–	–	Yes
Candidate sRNAs from high-throughput transcriptome studies	-	–	–	–	Yes

instance, collects only 87 validated sRNAs and 310 sRNA homologs.

Second, BSRD not only provides extensive functional descriptions for sRNAs, but also includes multiple new sRNA annotations from manually curated literature mining, including growth phase, Hfq binding and Rho-independent terminators. It also gives access to large-scale target search prediction of identified sRNAs. We have also integrated information of upstream regulon sigma factors to sRNA regulatory networks for a more comprehensive visualization of regulatory functions.

Third, although recent developments of deep sequencing technology have advanced sRNA researches, web-based tools for annotating sRNAs from high-throughput sequencing data are unavailable. We have thus developed sRNADeep to meet this need. We evaluated the performance of sRNADeep with the transcriptome data of *Listeria monocytogenes* (38). All 13 differentially expressed sRNAs previously reported were successfully recovered by sRNADeep. In addition, nine previously uncharacterized sRNAs were also identified by sRNADeep. sRNADeep could be a useful tool for characterizing sRNAs from deep sequencing data.

FUTURE DEVELOPMENTS

We will continue to import information concerning new bacterial genomes and update sRNA annotations in

BSRD. We also welcome submissions of novel sRNAs or annotations. Because of the expanded use of high-throughput deep sequencing, we expect to develop functions such as the evaluation of the effects of sRNA binding from transcriptome data, and prediction of novel sRNAs by an improved version of sRNADeep.

AVAILABILITY

BSRD is freely available at http://kwanlab.bio.cuhk.edu.hk/BSRD. All sRNA sequences are also available for download in FASTA format. There are no access restrictions for academic and commercial use. The content of BSRD is freely available under the ODC Open Database License.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1 and Supplementary Figures 1–3.

ACKNOWLEDGEMENTS

The authors thanks Michael Galperin and also the anonymous referees for critical reviews and constructive suggestions on BSRD.

FUNDING

Funding for open access charge: a Research Fund for the Control of Infectious Diseases, Food and Health Bureau of the Hong Kong SAR, China [RFCID CHP-PH-06].

Conflict of interest statement: None declared.

REFERENCES

1. Hershberg,R., Altuvia,S. and Margalit,H. (2003) A survey of small RNA-encoding genes in Escherichia coli. *Nucleic Acids Res.*, **31**, 1813–1820.
2. Vogel,J. and Luisi,B.F. (2011) Hfq and its constellation of RNA. *Nat. Rev. Microbiol.*, **9**, 578–589.
3. Brantl,S. (2012) Small regulatory RNAs (sRNAs): key players in prokaryotic metabolism, stress response, and virulence. In: Mallick,B. (ed.), *Regulatory RNAs*. Springer Berlin Heidelberg, pp. 73–109.
4. Thomason,M.K., Fontaine,F., De Lay,N. and Storz,G. (2012) A small RNA that regulates motility and biofilm formation in response to changes in nutrient availability in Escherichia coli. *Mol. Microbiol.*, **84**, 17–35.
5. Lenz,D.H., Mok,K.C., Lilley,B.N., Kulkarni,R.V., Wingreen,N.S. and Bassler,B.L. (2004) The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in Vibrio harveyi and Vibrio cholerae. *Cell*, **118**, 69–82.
6. Holmqvist,E., Unoson,C., Reimegård,J. and Wagner,E.G.H. (2012) A mixed double negative feedback loop between the sRNA MicF and the global regulator Lrp. *Mol. Microbiol.*, **84**, 414–427.
7. Mizuno,T., Chou,M.Y. and Inouye,M. (1984) A unique mechanism regulating gene expression: translational inhibition by a complementary RNA transcript (micRNA). *Proc. Natl Acad. Sci. USA*, **81**, 1966–1970.
8. Koo,J.T., Alleyne,T.M., Schiano,C.A., Jafari,N. and Lathem,W.W. (2011) Global discovery of small RNAs in Yersinia pseudotuberculosis identifies Yersinia-specific small, noncoding RNAs required for virulence. *Proc. Natl Acad. Sci. USA*, **108**, E709–E717.
9. Gama-Castro,S., Salgado,H., Peralta-Gil,M., Santos-Zavaleta,A., Muñiz-Rascado,L., Solano-Lira,H., Jimenez-Jacinto,V., Weiss,V., García-Sotelo,J.S., López-Fuentes,A. *et al.* (2011) RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res.*, **39**, D98–D105.
10. Keseler,I.M., Collado-Vides,J., Santos-Zavaleta,A., Peralta-Gil,M., Gama-Castro,S., Muñiz-Rascado,L., Bonavides-Martínez,C., Paley,S., Krummenacker,M., Altman,T. *et al.* (2011) EcoCyc: a comprehensive database of *Escherichia coli* biology. *Nucleic Acids Res.*, **39**, D583–D590.
11. Gardner,P.P., Daub,J., Tate,J., Moore,B.L., Osuch,I.H., Griffiths-Jones,S., Finn,R.D., Nawrocki,E.P., Kolbe,D.L., Eddy,S.R. *et al.* (2011) Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res*, **39**, D141–D145.
12. Huang,H.Y., Chang,H.Y., Chou,C.H., Tseng,C.P., Ho,S.Y., Yang,C.D., Ju,Y.W. and Huang,H.D. (2009) sRNAMap: genomic maps for small non-coding RNAs, their regulators and their targets in microbial genomes. *Nucleic Acids Res.*, **37**, D150–D154.
13. Cao,Y., Wu,J., Liu,Q., Zhao,Y., Ying,X., Cha,L., Wang,L. and Li,W. (2010) sRNATarBase: a comprehensive database of bacterial sRNA targets verified by experiments. *RNA*, **16**, 2051–2057.
14. Eggenhofer,F., Tafer,H., Stadler,P.F. and Hofacker,I.L. (2011) RNApredator: fast accessibility-based prediction of sRNA targets. *Nucleic Acids Res*, **39**, W149–W154.
15. Cao,Y., Zhao,Y., Cha,L., Ying,X., Wang,L., Shao,N. and Li,W. (2009) sRNATarget: a web server for prediction of bacterial sRNA targets. *Bioinformation*, **3**, 364–366.
16. Hokamp,K. and Wolfe,K.H. (2004) PubCrawler: keeping up comfortably with PubMed and GenBank. *Nucleic Acids Res.*, **32**, W16–W19.
17. Lu,X., Goodrich-Blair,H. and Tjaden,B. (2011) Assessing computational tools for the discovery of small RNA genes in bacteria. *RNA*, **17**, 1635–1647.
18. Chen,Y., Indurthi,D.C., Jones,S.W. and Papoutsakis,E.T. (2011) Small RNAs in the genus Clostridium. *MBio*, **2**, e00340–10.
19. Chan,P.P., Holmes,A.D., Smith,A.M., Tran,D. and Lowe,T.M. (2012) The UCSC Archaeal Genome Browser: 2012 update. *Nucleic Acids Res.*, **40**, D646–D652.
20. Hofacker,I.L. and Stadler,P.F. (2006) Memory efficient folding algorithms for circular RNA secondary structures. *Bioinformatics*, **22**, 1172–1176.
21. Chen,S., Lesnik,E.A., Hall,T.A., Sampath,R., Griffey,R.H., Ecker,D.J. and Blyn,L.B. (2002) A bioinformatics based approach to discover small RNA genes in the Escherichia coli genome. *BioSystems*, **65**, 157–177.
22. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.
23. Mann,B., van Opijnen,T., Wang,J., Obert,C., Wang,Y.-D., Carter,R., McGoldrick,D.J., Ridout,G., Camilli,A., Tuomanen,E.I. *et al.* (2012) Control of virulence by small RNAs in Streptococcus pneumoniae. *PLoS Pathog.*, **8**, e1002788.
24. Fröhlich,K.S., Papenfort,K., Berger,A.A. and Vogel,J. (2012) A conserved RpoS-dependent small RNA controls the synthesis of major porin OmpD. *Nucleic Acids Res.*, **40**, 3623–3640.
25. Busch,A., Richter,A.S. and Backofen,R. (2008) IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, **24**, 2849–2856.
26. Tafer,H. and Hofacker,I.L. (2008) RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics*, **24**, 2657–2663.
27. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
28. Gene Ontology Consortium. (2012) The Gene Ontology: enhancements for 2011. *Nucleic Acids Res.*, **40**, D559–D564.

29. Eilbeck,K., Lewis,S.E., Mungall,C.J., Yandell,M., Stein,L., Durbin,R. and Ashburner,M. (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.*, **6**, R44.
30. Misra,R.V., Horler,R.S.P., Reindl,W., Goryanin,I.I. and Thomas,G.H. (2005) EchoBASE: an integrated post-genomic database for Escherichia coli. *Nucleic Acids Res.*, **33**, D329–D333.
31. Rudd,K.E. (2000) EcoGene: a genome sequence database for Escherichia coli K-12. *Nucleic Acids Res.*, **28**, 60–64.
32. Hussein,R. and Lim,H.N. (2012) Direct comparison of small RNA and transcription factor signaling. *Nucleic Acids Res.*, **40**, 7269–7279.
33. Lopes,C.T., Franz,M., Kazi,F., Donaldson,S.L., Morris,Q. and Bader,G.D. (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–2348.
34. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
35. Paşaniuc,B., Zaitlen,N. and Halperin,E. (2011) Accurate estimation of expression levels of homologous genes in RNA-seq experiments. *J. Comput. Biol.*, **18**, 459–468.
36. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
37. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
38. Oliver,H.F., Orsi,R.H., Ponnala,L., Keich,U., Wang,W., Sun,Q., Cartinhour,S.W., Filiatrault,M.J., Wiedmann,M. and Boor,K.J. (2009) Deep RNA sequencing of L. monocytogenes reveals overlapping and extensive stationary phase and sigma B-dependent transcriptomes, including multiple highly transcribed noncoding RNAs. *BMC Genomics*, **10**, 641.