


RESEARCH

Open Access



Identification of small non-coding RNAs from *Rhizobium etli* by integrated genome-wide and transcriptome-based methods

Kasthuri Rajendran¹, Vikram Kumar², Ilamathi Raja¹, Manoharan Kumariah¹ and Jebasingh Tennyson^{2*} 

Abstract

Background: Small non-coding RNAs (sRNAs) are regulatory molecules, present in all forms of life, known to regulate various biological processes in response to the different environmental signals. In recent years, deep sequencing and various other computational prediction methods have been employed to identify and analyze sRNAs.

Results: In the present study, we have applied an improved sRNA scanner method to predict sRNAs from the genome of *Rhizobium etli*, based on PWM matrix of conditional sigma factor 32. sRNAs predicted from the genome are integrated with the available stress specific transcriptome data to predict putative conditional specific sRNAs. A total of 271 sRNAs from the genome and 173 sRNAs from the transcriptome are computationally predicted. Of these, 25 sRNAs are found in both genome and transcriptome data. Putative targets for these sRNAs are predicted using TargetRNA2 and these targets are involved in a wide array of cellular functions such as cell division, transport and metabolism of amino acids, carbohydrates, energy production and conversion, translation, cell wall/membrane biogenesis, post-translation modification, protein turnover and chaperones. Predicted targets are functionally classified based on COG analysis and GO annotations.

Conclusion: sRNAs predicted from the genome, using PWM matrices for conditional sigma factor 32 could be a better method to identify the conditional specific sRNAs which expand the list of putative sRNAs from the intergenic regions (IgRs) of *R. etli* and closely related α -proteobacteria. sRNAs identified in this study would be helpful to explore their regulatory role in biological cellular process during the stress.

Keywords: sRNA, *Rhizobium etli*, Sigma factor 32, Genome, Transcriptome

Background

Small non-coding RNAs are bacterial regulatory molecules, 50-500 nt (bp) in length and contain several stem loops. sRNAs are often located in the intergenic regions, transcribed from their own promoter or promoters of nearby genes and contain *rho*-independent terminator. sRNAs regulate the gene expression by perfect or imperfect base pairing with complementary sequence stretches, generally located in 5'-UTR regions of *trans*-encoded target mRNAs, resulting in altered target

mRNA translation and stability [1–3]. The regulation of sRNAs are mediated with the help of chaperone Hfq, enhance RNA-RNA interaction, through the preferential binding at single-stranded AU-rich regions of the non-coding RNAs and their target mRNAs [4].

Several sRNAs have been identified by genome-wide profiling and transcriptome-based methods. To date, many computational techniques and experimental methods have been used to predict sRNAs in both gram-negative and gram-positive bacteria [5–11]. sRNAs regulate diverse cellular processes and conditionally expressed during oxidative stress, iron uptake, quorum sensing, virulence and heat shock [5, 12–14]. Sigma factors are transcription initiation factors that enable specific binding of RNA polymerase (RNAP) to gene

* Correspondence: jebasinghs@gmail.com

²Department of Plant Sciences, School of Biological Sciences, Madurai Kamaraj University, Madurai, Tamil Nadu 625 021, India
Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

promoters. Bacteria contain different sigma factors, each capable of directing the core polymerase to transcribe a specific set of genes, depending on the environmental or developmental signals [15]. However reports on screening the conditional sRNAs are not yet available for the group of α -proteobacteria, except sRNAs predicted from RNA sequence analysis under heat shock and saline shock; and sRNAs predicted from the genome are integrated with the virulence specific transcriptome data [14, 16]. In our previous work, conditional sigma factor based-sRNAs were predicted from the genome of *Agrobacterium* using an improved sRNA scanner. This method was used to identify the sRNAs that are regulated by several conditional sigma factors, such as, 24, 32 and 54. sRNA scanner identified the sRNAs resided in intergenic regions of the genome, based on the transcriptional signals. This bioinformatic tool uses PWM matrices of sRNA promoter and *rho*-independent terminators signals, through sliding window-based genome scans, using consensus sequences of sigma factor promoter binding sites – 35 and – 10 and *rho*-independent transcription terminator sequences [16].

Rhizobium etli is a gram-negative bacterium that belongs to Rhizobiales of α -proteobacteria, interacts symbiotically with the common beans *Phaseolus vulgaris* to form nitrogen-fixing root nodules. Inside the nodules, bacteria differentiate into bacteroids that are capable of fixing the atmospheric N_2 into NH_3 . The genome of *R. etli* consists of one circular chromosome (6,530,228 bp) and six plasmids: p42a (194,229 bp), p42b (184,338 bp), p42c (250,948 bp), p42d (371,254 bp), p42e (505,334 bp) and p42f (642,517 bp) with 6034 protein-coding genes [17]. Two earlier studies were reported on the identification of sRNA candidates in *R.etli*. Using tiling microarray analysis, 66 novel sRNA candidates comprising 17 putative sRNAs and 49 putative *cis*-regulatory ncRNAs were computationally predicted and 4 of these were confirmed subsequently by wet-lab experiments [9]. Yet another study, identified 13 differentially expressed ncRNAs under heat shock and 9 under saline shock conditions in *R.etli* [14]. However, there is scanty information on stress conditional specific sRNAs in *Rhizobium*.

In the present study, we report the sRNAs predicted from the genome and transcriptome of *Rhizobium etli*. Further, sRNAs predicted from the genome are integrated with the stress-specific transcriptome to identify putative conditional specific sRNAs. The mRNA targets for these sRNAs were identified and data are presented on the functional categorization and regulatory network analysis for the predicted mRNA targets.

Results

Genome-wide sRNA prediction by improved sRNAscanner
Prediction of sRNAs from the nitrogen-fixing *Rhizobium* was performed by genome-wide computational analysis,

based on the PWM matrices of conditional sigma factor 32 (Heat shock sigma factor) using improved sRNA Scanner program [16]. sRNA scanner demarks the transcription units (TUs) using consensus sequences of sigma factor binding sites (– 35 and – 10 (Supplementary file 3)) and *rho*-independent transcription terminator sequences. An earlier version of sRNA Scanner uses PWM matrix; only for housekeeping sigma factor 70 and *rho*-independent transcription termination in which limited numbers of training sequences were used. The total number of sRNAs predicted from each replicons of *Rhizobium etli* is graphically represented in Fig. 1.

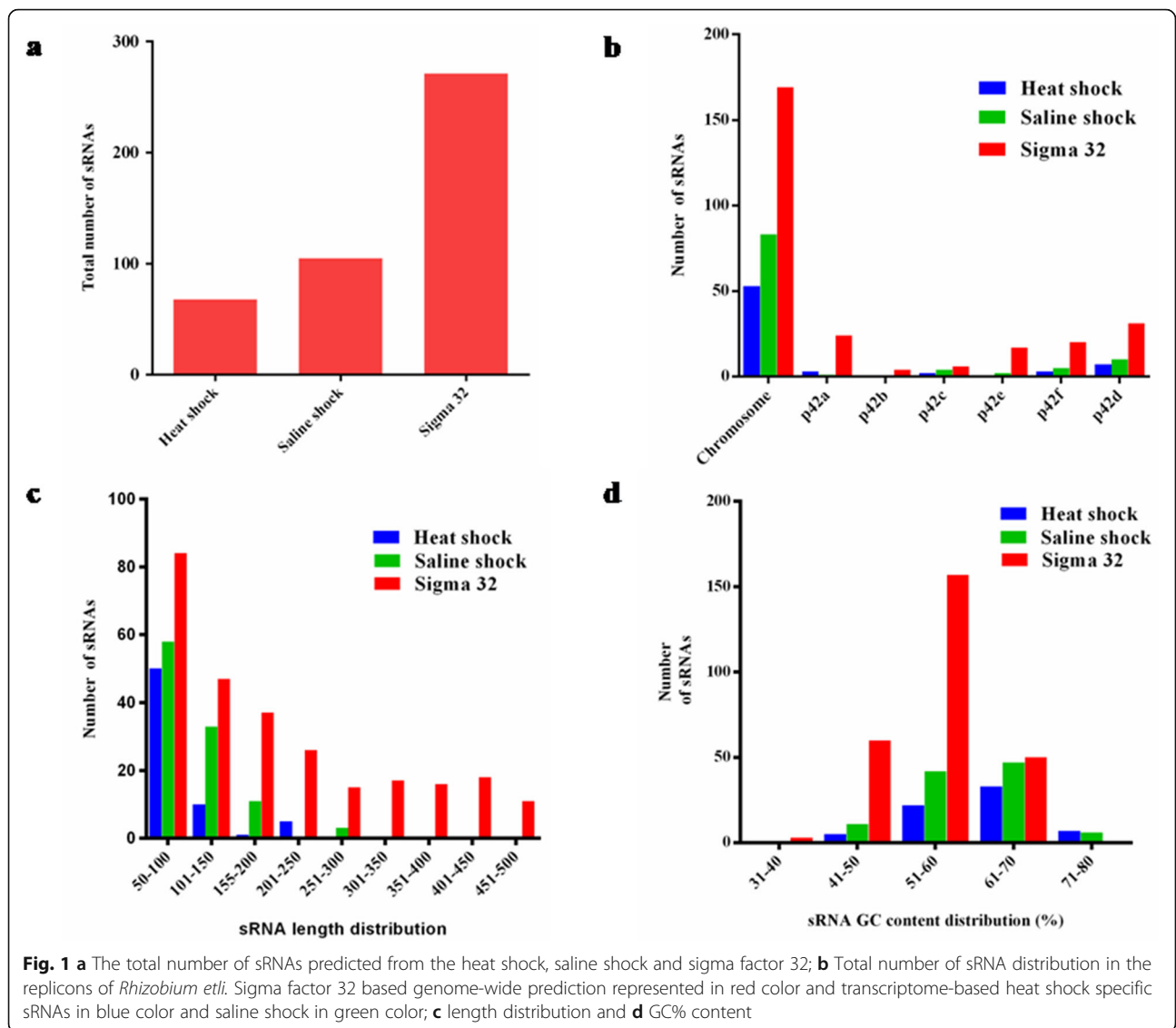
The majority of the sRNA candidates identified varied in length between 50 and 500 nt. GC content for most of the sRNAs of *Rhizobium* found to have 50 to 70%. A total of 247 sRNAs were predicted from the genome of *R. etli* known to be conditionally regulated by sigma factor 32. To find the novel putative sRNAs, predicted sRNAs were searched against Rfam database and BSRD database to eliminate the conserved homologs (Table 1). Seventeen and four sRNA candidates have shown homology with already identified sRNAs in Rfam and BSRD database, respectively (Table 1). The sRNAs predicted from the genome were compared with previously reported sRNAs. Eight sRNA candidates were conserved with the earlier reported sRNAs by Vercruyssen et al. 2010 [9] and one sRNA with the López-Leal et al. 2015 [14].

Transcriptome based sRNAs prediction

The high-quality RNAseq reads of *R. etli* under control, heat and saline shock were aligned to the genome of *R. etli* using Rockhopper. After alignment, transcripts from the intergenic regions and antisense regions from the complementary strand of the protein-coding genes were identified. The intergenic sRNAs having a length of 50-500 nt were taken for further analysis. A total of 68 *trans*-encoded sRNAs under the heat shock and 105 under the saline shock were identified. A relatively larger number of sRNAs were found to be expressed from the chromosomes, has a length of 50 to 150 nt (Fig. 1). Further, predicted sRNA candidates were searched against Rfam and BSRD databases. To find the novel putative sRNA candidates, the above-screened sRNAs were compared with previously reported sRNAs (supplementary file 4). Eight sRNAs from heat shock and fourteen sRNAs from saline shock showed homology with already reported sRNAs by Vercruyssen et al. 2010 [9]; five sRNAs from heat shock and two sRNAs from saline shock with sRNAs reported by López-Leal et al. 2015 [14].

sRNA conservation and comparative analysis

sRNAs are known to be conserved in nature, in order to study the sRNA conservation in the present study, the sRNA conservation analysis was performed between the



rhizobium strains, interestingly the sRNAs of *Rhizobium etli* were highly conserved with identities ranging from 80 to 100% with *Rhizobium leguminosarum*. From the analysis, it was found that 21 sRNAs from the genome-based search (18 sRNAs from chromosome and 1 sRNA from 2nd, 6th and 7th replicons) and in the case of the transcriptome (saline shock), 2 sRNAs (chromosomally encoded) were conserved with the sigma factor 32

regulated sRNAs of *R. leguminosarum* (unpublished data). Three chromosomally encoded sRNAs of *R. etli* were found to be conserved with one specific sRNA of *R. leguminosarum* (94–95% identity), which was further selected for the quantification analysis in *R. leguminosarum* (Table 2). The identified novel sRNAs candidates of the genome and transcriptome were correlated to identify the common sRNAs between conditional

Table 1 sRNAs identified from the genome and transcriptome of *Rhizobium etli*

S. No.	No. of sRNAs predicted	Homologous identified in Rfam	Homologous identified in BSRD	Reported sRNAs	Total identified sRNAs
Transcriptome					
1. Heat shock	68	5	2	10	51
2. Saline shock	105	5	4	19	77
Genome					
Sigma 32	271	17	4	9	241

Table 2 sRNA candidates having homologs with *Rhizobium leguminosarum*

S. No.	<i>Rhizobium etli</i>			Identity	<i>Rhizobium leguminosarum</i>		
	Start	Stop	Length of the sRNA		Start	Stop	Length of the sRNA
Genome							
1.	3,324,460	3,324,583	124	100%	3,807,013	3,807,136	124
2.	3,086,939	3,087,101	158	95.57%			
3.	3,086,838	3,087,151	158	95.57%	3,578,887	3,579,049	163
Transcriptome							
4. (saline shock)	3,086,972	3,087,084	113	94.69%			

specific sigma factor 32 derived sRNAs with the stress-specific sRNA transcripts. A total of 271 sRNAs identified from the genome, of which 241 were novel. Similarly, 173 sRNAs from the transcriptome were identified (Supplementary file 1), of which 128 sRNAs were novel. Based on comparative analysis, 25 novel sRNAs were found to be common between the genome-wide and transcriptome data of *R.etli*.

Target identification

The sRNAs regulate diverse cellular processes by interacting with complementary sequence stretches of *trans*-encoded mRNAs. The target mRNAs were predicted for the sRNAs identified from genome-wide and transcriptome data by using TargetRNA2. Based on the thermodynamic interaction energy (kcal/mol) of hybridization between the sRNA and mRNA targets and significant *p*-value (< 0.05), 30 sRNA candidates were selected from the transcriptome of heat and saline shock for further analysis (Table 3). Target mRNAs for the 25 common sRNAs predicted from the genome of *R. etli* are provided in supplementary file 6. Fifty-five sRNAs were taken for further analysis.

Functional categorization of sRNA target genes

To study the role of target mRNAs, selected target mRNAs of thirty sRNA candidates from the transcriptome of heat and saline shock conditions were functionally annotated by COG and GO analysis. Under heat and saline shock conditions, mRNA targets were enriched in COG categories of transport and metabolism of amino acids, carbohydrates, lipids, energy production, and conversion, post-translation modification, protein turn over and chaperons, cell wall/membrane biogenesis and translation (Fig. 2). Enriched GO terms for the target mRNAs were widely distributed about their respective biological cellular processes. The target genes were annotated in 3 classes, *viz.*, biological processes, molecular functions and cellular components. The targets categorized under biological processes include the genes involved in metabolic and cellular processes, molecular function, catalytic activity and binding, such as transporter activity, DNA binding, RNA binding and ion binding (Fig. 3).

GO regulatory network

GO regulatory network (GRN) was constructed for the sRNA target genes for the sRNAs predicted from the transcriptome profiled under conditions of heat and saline. The GO network of mRNA targets of heat shock sRNAs is shown in Fig. 4. The regulation of cell shape was the central node in the GRN. Regulation of cell shape protein MviN (RHE_CH0386) showed interaction with many other GO terms such as regulation of DNA replication, signal transduction, protein folding, cellular amino acid metabolic process, carbohydrate metabolic process, fatty acid biosynthetic process, nitrogen metabolic process, and cell division. In the case of mRNA targets of saline shock sRNAs, phosphorelay signal transduction system was the central node in the network (Fig. 5) governed by *feuP* (RHE_CH01286) which showed interaction with other GO terms, such as, positive regulation of transcription, regulation of cell shape, metabolic process, transmembrane transport, translation, nucleotide catabolic process, cell wall organization, nitrate assimilation and cell cycle.

Promoter, terminator, secondary structure prediction

Promoter and *rho*-independent terminator sequences were predicted for the identified putative novel sRNAs (Table 4 and Supplementary file 6). Secondary structure was predicted for the selected sRNAs using RNAfold server. The predicted minimum free energy for the majority of the sRNAs ranges from -20 to -70 kcal/mol.

Discussion

sRNAs are known to regulate diverse cellular processes in prokaryotes [2, 18, 19]. To date, many computational based methods have been used to identify small regulatory RNAs in bacteria, but there are only a few reports available on the functional roles of sRNAs in *Rhizobium*. In 2016 Borella et al. have reported that the small RNA gene *mmgR* is controlled by nitrogen source in *Sinorhizobium meliloti* [20]. Recently, the function and mechanism of *Sinorhizobium meliloti trans*-sRNA *NfeR1* (Nodule formation efficiency RNA) was experimentally studied on the effect of osmoadaptation and symbiotic efficiency in *Alfa alfa* [21]. In the present study, we have combined genome-wide and transcriptome based

computational methods to identify the novel putative sRNA candidates in *R. etli*. Particularly, we have focused on the identification of sRNAs that are differentially expressed during heat and saline shock and its regulatory role. Genome-wide sigma factor 34 based sRNA predictions provided a total of 271 sRNA candidates. While comparing with the transcriptome based data, many sRNAs were predicted from the genome-wide prediction. A higher number of sRNA candidates were expressed from the chromosome than other replicons. One hundred sixty-nine sRNAs were predicted in the chromosome and 31 sRNAs were found in symbiotic plasmid p42d. A total of 128 novel sRNAs were predicted from the transcriptome data and we found number of sRNAs expressed under saline shock is more than the heat shock condition (Table 1). Although Lopez-Leal et al. 2015 have already reported novel sRNA in the previously published transcriptomic analysis, however only a small number of sRNAs have been reported in their study [14]. Our out of interest has led to the identification of more than 100 new sRNAs from RNA sequence data analysis. Besides, we have compared the identified sRNA candidates with previously reported sRNA data of *R. etli* [14], a total of 9 sRNAs from the sigma 32 genome-based method, 10 from the heat shock and 19 from the saline shock were conserved with the already reported sRNAs. While performing sRNA conservation analysis, 21 sRNAs were found to be overlapped with the *R. leguminosarum*, in which 3 sRNAs of *R. etli* were conserved (share 94–95% homology) with a single sRNA candidate and interestingly another chromosomally encoded sRNA (124 nt) share 100% homology with the sRNA of *R. leguminosarum*.

sRNAs regulate the gene expression by perfect or imperfect base pairing with the target mRNA. Single sRNA is known to regulate multiple mRNA targets, either it upregulates or downregulates based on the binding sites of a set of genes. Earlier findings have shown that sRNAs regulate diverse biological and cellular processes, such as energy metabolism, quorum sensing (QS) and biofilm formation, stress responses and adaptation to adverse growth conditions, and pathogenesis [19, 22, 23]. In the present study, we have identified potential targets of sRNAs and analyzed its role using different computational methods. Target prediction method revealed that 15 sRNAs of heat shock sRNAs have complementary binding sites with heat shock specific genes such as *groES*, *groESch3*, *groEL*, *ibpA*, serine proteases- *degPch1*, *degPch2* and also with the virulence factor coding gene *MviN* which codes for a transmembrane protein. Among the 15 selected sRNAs of the saline shock group, a few sRNAs have a significant binding site on serine proteases (*degPch1*, *degPch2*) and *mviN*. Besides, we could infer that the identified sRNAs might regulate several

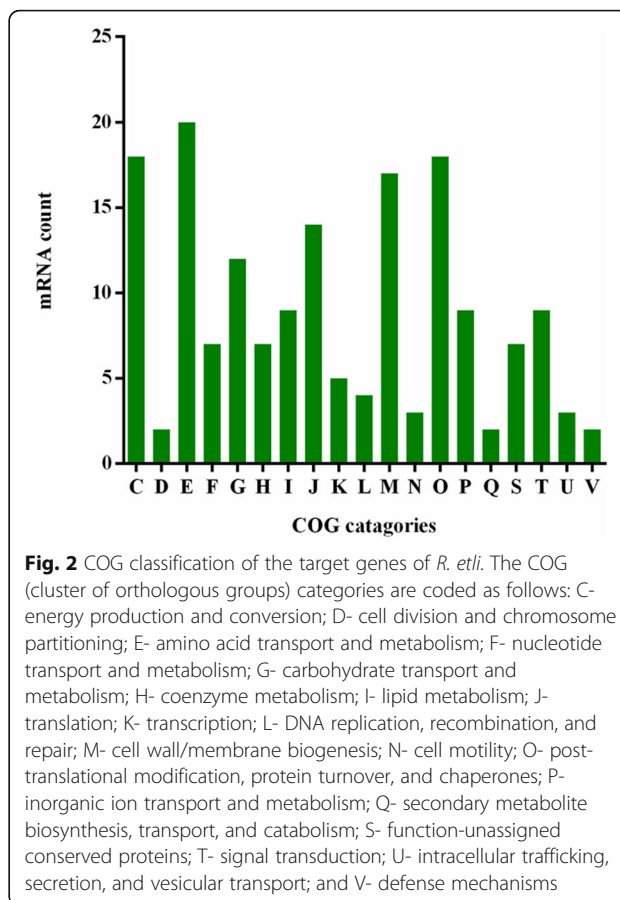
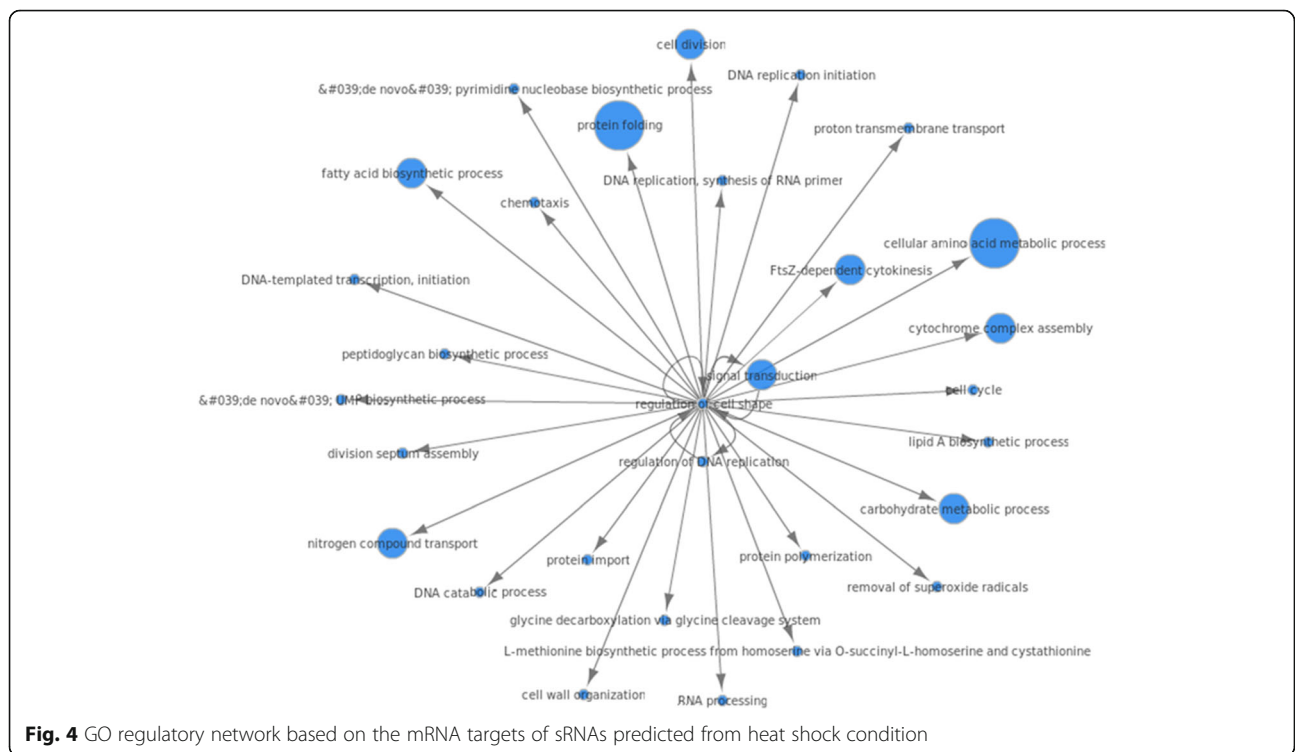
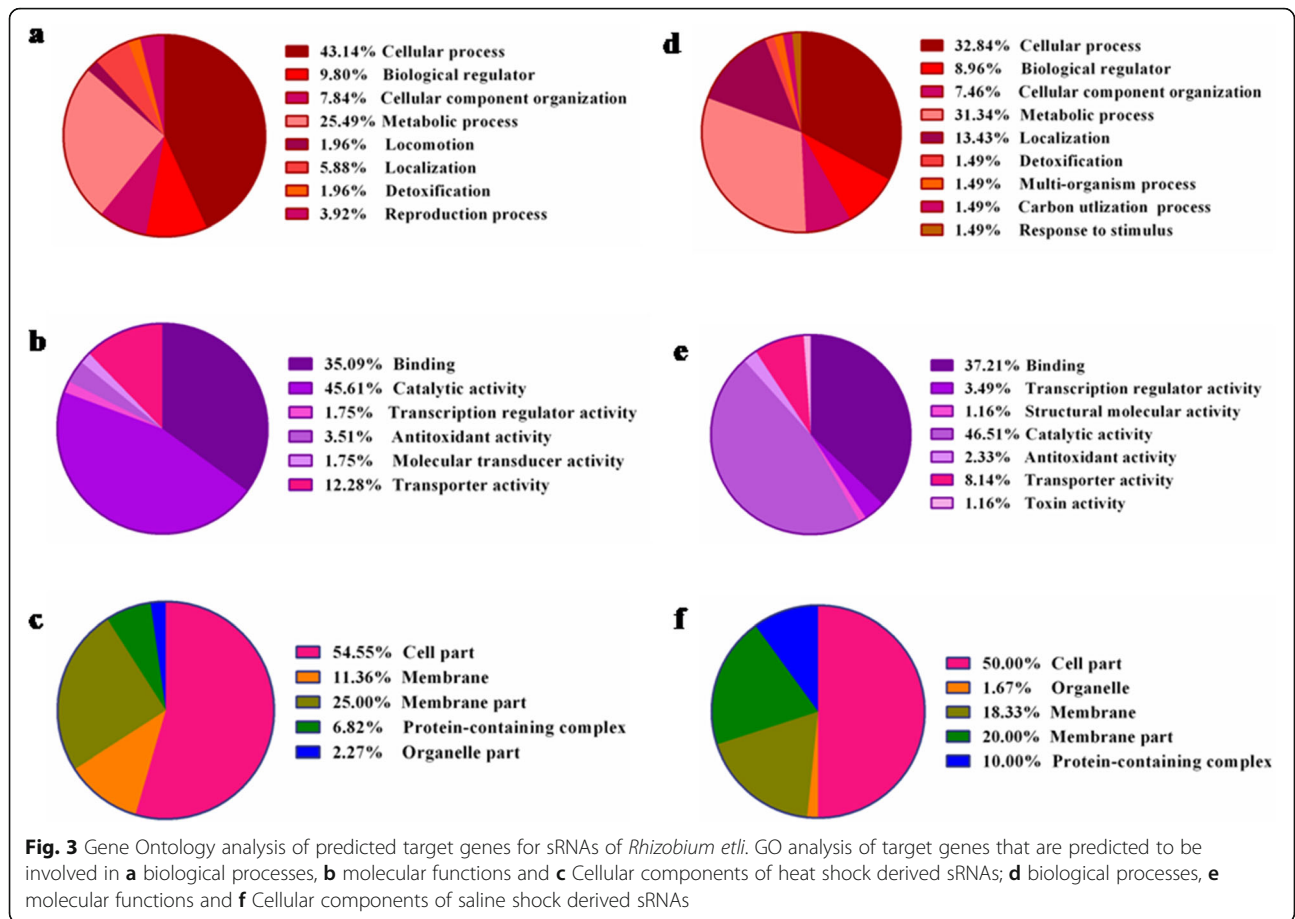
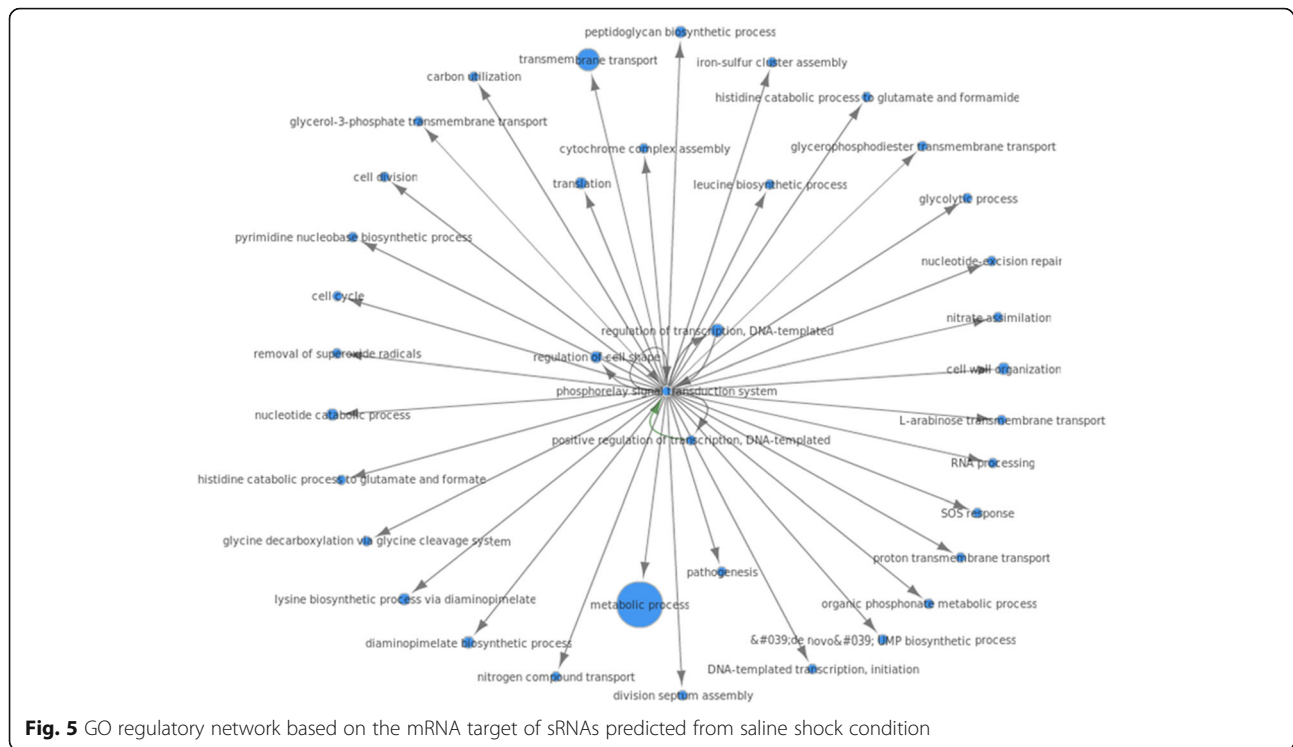


Fig. 2 COG classification of the target genes of *R. etli*. The COG (cluster of orthologous groups) categories are coded as follows: C- energy production and conversion; D- cell division and chromosome partitioning; E- amino acid transport and metabolism; F- nucleotide transport and metabolism; G- carbohydrate transport and metabolism; H- coenzyme metabolism; I- lipid metabolism; J- translation; K- transcription; L- DNA replication, recombination, and repair; M- cell wall/membrane biogenesis; N- cell motility; O- post-translational modification, protein turnover, and chaperones; P- inorganic ion transport and metabolism; Q- secondary metabolite biosynthesis, transport, and catabolism; S- function-unassigned conserved proteins; T- signal transduction; U- intracellular trafficking, secretion, and vesicular transport; and V- defense mechanisms

hypothetical proteins. In 2014 López-Leal et al. reported, *groESch2*, *groEL*, and *ibpA* heat shock genes were up-regulated in *R. etli* during heat shock and two serine proteases, viz., *degPch1* and *degPch2* were significantly over-expressed during saline shock [14]. Based on the results of the present study, we suggest these newly identified sRNAs might regulate the expression of heat and saline shock specific genes. Further, the target mRNAs of these sRNAs were taken for the functional categorization using COG and GO analysis.

In the GO enrichment analysis, most of the target genes were associated with cellular, metabolic and transport processes. COG analysis revealed that most of the target mRNAs of sRNAs of this study were involved in amino acid transport and metabolism, energy production and conversion, post-translational modification, protein turnover, chaperones and cell wall/membrane biogenesis. Particularly, heat shock sRNAs are firmly categorized in post-translational modification, protein turnover and chaperones in COG analysis. Further, we have constructed the GRN of predicted target mRNAs using the biological process GO terms. The transmembrane protein *MviN* constitutes the central node in the regulatory network in the heat shock condition. It is well





documented that subjecting the cells to heat shock can disrupt the cell membrane integrity. Regulation of cell shape protein MviN was shown to be up-regulated under heat shock condition as compared to control besides the down-regulation of DNA replication proteins *dnaA* and *dnaB* [14]. sRNAs identified in the present study have complementary binding sites with these target proteins, which might down or up-regulate the target proteins. Network analysis revealed that many target genes mainly involved in protein folding, cellular amino acid, carbohydrate metabolic processes, signal transduction, cell division, cell cycle and cell wall organization. Under saline shock conditions, many target mRNAs were found to be involved in the metabolic process, transmembrane transport, cell organization, translation and regulation of transcription.

Conclusion

In this study, for the first time, we reported novel sRNAs expressed differentially under stress conditions. The mRNA targets of these sRNAs were identified, functionally classified and found that these sRNAs are involved in different cellular metabolic processes including protein folding. GO network analysis of *Rhizobium* revealed a new biological role of sRNAs. Several reports are available regarding the sRNA identification but the reports on the biological roles of sRNAs in *Rhizobium* are quite limited. This work begins to address the new biological insights in sRNAs function and its roles in a bacterial

system. It's possible that the above applied genome-wide computational methods can be used to identify the conditional specific sRNAs in other *Rhizobium* or closely related α -proteobacteria. However, the precise role of sRNAs reported in the preset study needs to be validated experimentally in future studies.

Materials and methods

Genome-wide prediction of sRNAs from *Rhizobium etli* by using improved sRNAscanner

Rhizobium etli complete genome sequence and annotation files were retrieved from the National Centre for Biotechnology Information (NCBI) ftp site. Genome sequences and annotation files were downloaded in Fasta nucleic acid (.fna) and protein data file (.ptt) formats, respectively. Accession numbers of *Rhizobium etli* with their respective replicons used in our study are listed in the supplementary file 1. In the present study, we employed the improved version of the sRNA scanner to predict conditional sigma factor 32 specific sRNAs. This bioinformatic tool uses PWM matrices of sRNA promoter and *rho*-independent terminators signals (Supplementary file 2), through sliding window-based genome scans, using consensus sequences of sigma factor promoter binding sites -35 and -10 and *rho*-independent transcription terminator sequences.

Sigma factor 32 specific Position weight matrices were used for identifying sRNAs from the complete bacterial genome using sRNA Scanner [8, 16, 24]. sRNA Scanner

Table 3 sRNAs predicted from transcriptome data of heat and saline shock

S.No.	sRNA	Start	Stop	Strand	sRNAs length	Distance of up gene	Upstream gene	Distance of down gene	Downstream gene
1.	REH1	64,166	64,238	-	72	68	16S Ribosomal RNA	76	Ile tRNA
2.	REH2	135,830	135,893	-	63	14	hypothetical protein	100	hypothetical protein
3.	REH3	236,107	236,217	+	110	11	hypothetical protein	374	hypothetical protein
4.	REH4	351,037	351,205	+	168	91	GntR family transcriptional regulator	246	glutamine amidotransferase
5.	REH5	351,126	351,199	-	73	180	GntR family transcriptional regulator	252	glutamine amidotransferase
6.	REH6	489,951	490,025	+	74	50	trehalose-6-phosphate synthase	270	glucose-6-phosphate isomerase
7.	REH7	648,703	648,855	-	152	224	two-component response regulator protein	100	LuxR family transcriptional regulator
8.	REH8	730,097	730,188	-	91	230	oxidoreductase	250	methyl-accepting chemotaxis protein
9.	REH9	868,406	868,463	-	57	1	chaperonin GroEL	1	co-chaperonin GroES
10.	REH10	909,921	909,973	+	52	31	ribonuclease HII	85	hypothetical protein
11.	REH11	3,940,457	3,940,565	-	108	184	glutaredoxin protein	246	heavy metal-transporting ATPase
12.	REH12	4,100,672	4,100,733	-	61	109	F0F1 ATP synthase subunit epsilon	18	F0F1 ATP synthase subunit beta
13.	REH13	4,122,429	4,122,544	+	115	82	2-oxoglutarate dehydrogenase E1 component	32	succinyl-CoA synthetase subunit alpha
14.	REH14	55,880	55,946	+	66	244	Hypothetical protein	898	Hypothetical protein
15.	REH15	94,510	94,577	-	67	136	GntR family transcriptional regulator	37	6-phosphogluconate dehydrogenase
16.	RES1	236,090	236,217	+	127	0	hypothetical protein	127	hypothetical protein
17.	RES2	351,043	351,203	+	160	98	GntR family transcriptional regulator	248	glutamine amidotransferase
18.	RES3	370,123	370,180	-	57	469	30S ribosomal protein S20	314	chromosomal replication initiation protein
19.	RES4	562,123	562,230	+	107	89	hypothetical protein	349	hypothetical protein
20.	RES5	562,200	562,297	-	97	166	hypothetical protein	282	hypothetical protein
21.	RES6	630,669	630,835	-	166	48	ribose ABC transporter, substrate-binding protein	1	ribose ABC transporter, ATP-binding protein
22.	RES7	730,014	730,067	+	53	147	oxidoreductase	403	methyl-accepting chemotaxis protein
23.	RES8	909,943	910,038	-	95	53	ribonuclease HII	20	hypothetical protein
24.	RES9	1,926,293	1,926,364	+	71	106	hypothetical protein	284	hypothetical protein
25.	RES10	3,051,143	3,051,203	-	60	206	N-acyl-L-homoserine lactone (AHL) synthase	54	two-component response regulator protein
26.	RES11	3,776,485	3,776,552	+	67	169	hypothetical protein	207	hypothetical protein
27.	RES12	1,754,046	1,754,157	+	111	118	50S ribosomal protein L7/L12	30	DNA-directed RNA polymerase subunit beta
28.	RES13	1,548,328	1,548,402	+	74	133	molybdopterin converting factor subunit 2 protein	3	molybdopterin converting factor subunit 1 protein
29.	RES14	127,134	127,190	-	56	1203	hypothetical protein	1323	hypothetical protein
30.	RES15	572,116	572,220	-	104	189	cytochrome C oxidase, fixN chain protein	189	nitrogen fixation transcriptional regulator protein

Table 4 Promoter, terminator and secondary structure of sRNAs identified from the transcriptome data

sRNA No.	Promoter	Terminator	Secondary structure	Reference
0001	CAATCAAGT	TTTATG		[27] sRNA1
0002	CTCCAGCA	TTTATG		[27] sRNA2
0003	CGTAAATC	TTTATG		[27] sRNA3
0004	CGTAAATC	TTTATG		[27] sRNA4
0005	AGAGATCA	TTTATG		[27] sRNA5
0006	CGTAAATC	TTTATG		[27] sRNA6
0007	TTTAAATC	TTTATG		[27] sRNA7
0008	CGTAAATC	TTTATG		[27] sRNA8
0009	CGTAAATC	TTTATG		[27] sRNA9
0010	CGTAAATC	TTTATG		[27] sRNA10
0011	CGTAAATC	TTTATG		[27] sRNA11
0012	CGTAAATC	TTTATG		[27] sRNA12
0013	CGTAAATC	TTTATG		[27] sRNA13
0014	CGTAAATC	TTTATG		[27] sRNA14
0015	CGTAAATC	TTTATG		[27] sRNA15
0016	CGTAAATC	TTTATG		[27] sRNA16
0017	CGTAAATC	TTTATG		[27] sRNA17
0018	CGTAAATC	TTTATG		[27] sRNA18
0019	CGTAAATC	TTTATG		[27] sRNA19
0020	CGTAAATC	TTTATG		[27] sRNA20
0021	CGTAAATC	TTTATG		[27] sRNA21
0022	CGTAAATC	TTTATG		[27] sRNA22
0023	CGTAAATC	TTTATG		[27] sRNA23
0024	CGTAAATC	TTTATG		[27] sRNA24
0025	CGTAAATC	TTTATG		[27] sRNA25
0026	CGTAAATC	TTTATG		[27] sRNA26
0027	CGTAAATC	TTTATG		[27] sRNA27
0028	CGTAAATC	TTTATG		[27] sRNA28
0029	CGTAAATC	TTTATG		[27] sRNA29
0030	CGTAAATC	TTTATG		[27] sRNA30
0031	CGTAAATC	TTTATG		[27] sRNA31
0032	CGTAAATC	TTTATG		[27] sRNA32
0033	CGTAAATC	TTTATG		[27] sRNA33
0034	CGTAAATC	TTTATG		[27] sRNA34
0035	CGTAAATC	TTTATG		[27] sRNA35
0036	CGTAAATC	TTTATG		[27] sRNA36
0037	CGTAAATC	TTTATG		[27] sRNA37
0038	CGTAAATC	TTTATG		[27] sRNA38
0039	CGTAAATC	TTTATG		[27] sRNA39
0040	CGTAAATC	TTTATG		[27] sRNA40
0041	CGTAAATC	TTTATG		[27] sRNA41
0042	CGTAAATC	TTTATG		[27] sRNA42
0043	CGTAAATC	TTTATG		[27] sRNA43
0044	CGTAAATC	TTTATG		[27] sRNA44
0045	CGTAAATC	TTTATG		[27] sRNA45
0046	CGTAAATC	TTTATG		[27] sRNA46
0047	CGTAAATC	TTTATG		[27] sRNA47
0048	CGTAAATC	TTTATG		[27] sRNA48
0049	CGTAAATC	TTTATG		[27] sRNA49
0050	CGTAAATC	TTTATG		[27] sRNA50

was used with CSS of 12 and search length with 50–500 nt. To ensure the non-coding nature of the sRNA, the protein-coding potentials of the transcripts were assessed based on coding potential score (CPS) using the coding potential calculator (<http://cpc.cbi.pku.edu.cn/server>). Accordingly, CPS score – 1 represents weak non-coding and + 1 means weak coding of the transcript [25]. Transcripts with a true non-coding nature were considered for further annotation of sRNA. Length and GC content of the putative non-coding transcripts were analyzed using customized PERL script. To refine the data, every sRNA was checked in Rfam database and Bacterial small Small Regulatory RNA Database (BSRD) [26] to identify the already reported sRNAs. The sRNAs were also compared with previous reports to assess and confirm their novelty. Filtered putative non-coding RNAs (sRNAs) were used for further analysis.

Identification of sRNAs from transcriptome

The RNA-seq dataset was obtained from the NCBI Gene Expression Omnibus (GEO) (Accession No: GSM1212456) [14]. The raw reads of *R. etli* CE3 under three different conditions (control, heat shock, and saline shock) downloaded from the sequence read archive (SRA) database (Accession No.: SRP028924). The SRA tool kit was used for extracting the transcriptome reads from SRA files in FASTQ format [27]. PolyA, polyT and Illumina adapters were removed with cutadapt tool [28]. Sequence quality was analyzed using FastQC. Sequence reads having phred score > 20 were used for further analysis. Trimmed reads were aligned to the genome of *R. etli* by using Rockhopper tools for transcriptome read counting [29, 30]. Based on the alignment data, non-coding transcripts are considered as sRNA. The RPKM (reads per kilobase of transcript per million mapped reads) values of experimental conditions (heat and saline shock) were compared with control for calculating the fold change. Reads of the coding and non-coding transcripts were separated and aligned to the reference genome. The sRNA sequence was aligned to the genome and visualized using the Integrative genome viewer (IGV). Genomic coordinates of predicted sRNA were extracted from the genome using either Samtools or bedtools. Genomic coordinates of these predicted RNAs are provided in the Rockhopper output file.

Target and secondary structure prediction for sRNAs

TargetRNA2 Software was used to predict the mRNA targets for the predicted *trans*-encoded sRNAs (<http://cs.wellesley.edu/~btjaden/TargetRNA2/>). TargetRNA2 is a web server that identifies mRNA targets of sRNA regulatory action in bacteria. As input, TargetRNA2 takes the sequence of an sRNA and the name of a sequenced bacterial replicon and it uses a variety of features, including conservation of the sRNA in other bacteria, the secondary structure of the sRNA, the secondary structure of each

candidate mRNA target and the hybridization energy between the sRNA and mRNA targets [31].

RNAfold web server (<http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi>) was used to predict the secondary structure of sRNAs. sRNA FASTA sequences were used for calculating minimum free energy (ΔG) based on the partition function (default parameter) [32].

Functional enrichment analysis of novel putative sRNAs

Novel putative sRNAs were screened by the integration of the sRNAs predicted from the genome and transcriptome. Sigma factor 32 based sRNAs predicted from the genome were blasted against the sRNAs identified from the transcriptome data of shock conditions [14]. Further, selected sRNAs from genome and transcriptome were functionally annotated based on the target of these sRNAs.

Functional categorization of the predicted target mRNAs was done by clusters of orthologous group (COG) analysis using the EggNOG database [R]. Gene ontology (GO) annotations and regulatory relationships among the biological processes were analyzed through the GO regulatory network by using the comparative GO web server [33].

Prediction of promoter and terminator

The promoter and *rho*-independent terminator regions of sRNAs were analyzed from the region upstream of the transcription start site (TSS) and downstream of the transcription end site (TES), respectively. Genomic coordinates of 150-nt sequences upstream of TSS and 150-nt sequences downstream of TES were extracted using 'Bedtools' [34]. Further, 'BPROM' was used to identify the binding sites of $\sigma 70$ [35] and 'Arnold' for *rho*-independent terminators [36].

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s41544-020-00054-1>.

Additional file 1: Supplementary file 1 List of Accession numbers of *Rhizobium etli* with their respective replicons used in our study.

Supplementary file 2 Position Weight Matrix (PWM) log-odds ratio of nucleotides at each position of the sigma factor 32 binding motif. **Supplementary file 3** Consensus sequence logos of sigma 32 matrix used for sRNA Scanner program. **Supplementary file 4** List of sRNAs overlapped with the published data. a) Overlapped sequences with Ver-

cruysse et al. (2010) data, b) with López-Leal et al. (2014) data. **Supplementary file 5** List of flanking genes of the predicted common sRNAs under heat shock and saline shock conditions and sigma factor 32 based of *Rhizobium etli*. **Supplementary file 6** Promoter, terminator and secondary structures of the predicted common sRNAs from the genome and transcriptome data.

Abbreviations

sRNAs: Small non-coding RNAs; PWM: Positional weigh matrix; hfq: Host factor q; nt: Nucleotide; bp: Basepair; IgRs: Intergenic regions; ncRNAs: Non-coding RNAs; DNA: Deoxyribonucleic acid; RNA: Ribonucleic acid; mRNA: Messenger RNA; ftp: File transfer protocol; fna: Fasta nucleic acid;

ptt: Protein data file; CSS: Cumulative sum of score; CPS: Coding potential score; GC: Guanine and cytosine; GEO: Gene Expression Omnibus; SRA: Sequence read archive; IGV: Integrative genome viewer; BSRD: Bacterial small Small Regulatory RNA Database; Rfam: RNA family data base; GO: Gene Ontology; GRN: GO regulatory network; COG: Clusters of orthologous groups; TSS: Transcription start site; TES: Transcription end site

Acknowledgements

We thank UGC-BSR for the financial support of KR.

Authors' contributions

Jebasingh Tennyson and Manoharan Kumariah conceived the idea. Kasthuri Rajendran planned and performed the experiments. Vikram Kumar and Ilamathi Raja created PWM matrix of improved sRNAScanner. The authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

All data generated or analysed during this study are included in this published article.

Ethics approval and consent to participate

This article does not contain any studies with human participants performed by any of the authors.

Consent for publication

Not applicable.

Competing interests

All authors declare that they have no competing interests.

Author details

¹Department of Plant Morphology and Algology, School of Biological Sciences, Madurai Kamaraj University, Madurai, Tamil Nadu 625 021, India.

²Department of Plant Sciences, School of Biological Sciences, Madurai Kamaraj University, Madurai, Tamil Nadu 625 021, India.

Received: 15 October 2019 Accepted: 20 August 2020

Published online: 07 September 2020

References

- Mizuno T, CHOU MY, Inouye M. Regulation of gene expression by a small RNA transcript (micRNA) in *Escherichia coli* K-12. *Proc Japan Acad Ser B*. 1983;59(10):335–8.
- Gottesman S. Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet*. 2005;21(7):399–404.
- Vogel J, Papenfort K. Small non-coding RNAs and the bacterial outer membrane. *Curr Opin Microbiol*. 2006;9(6):605–11.
- Møller T, Franch T, Højrup P, Keene DR, Bächinger HP, Brennan RG, Valentin-Hansen P. Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction. *Mol Cell*. 2002;9(1):23–30.
- Wassarman KM, Repoila F, Rosenow C, Storz G, Gottesman S. Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev*. 2001;15(13):1637–51.
- Rivas E, Klein RJ, Jones TA, Eddy SR. Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr Biol*. 2001;11(17):1369–73.
- Del Val C, Rivas E, Torres-Quesada O, Toro N, Jiménez Zurdo JL. Identification of differentially expressed small non-coding RNAs in the legume endosymbiont *Sinorhizobium meliloti* by comparative genomics. *Mol Microbiol*. 2007;66(5):1080–91.
- Sridhar J, Narmada SR, Sabarinathan R, Ou HY, Deng Z, Sekar K, Rafi ZA, Rajakumar K. sRNAScanner: a computational tool for intergenic small RNA detection in bacterial genomes. *PLoS One*. 2010;5(8):e11970.
- Verduysse M, Fauvart M, Cloots L, Engelen K, Thijs IM, Marchal K, Michiels J. Genome-wide detection of predicted non-coding RNAs in *rhizobium etli* expressed during free-living and host-associated growth using a high-resolution tiling array. *BMC Genomics*. 2010;11(1):53.

10. Schlüter JP, Reinkensmeier J, Daschkey S, Evguenieva-Hackenberg E, Janssen S, Jänicke S, Becker JD, Giegerich R, Becker A. A genome-wide survey of sRNAs in the symbiotic nitrogen-fixing alpha-proteobacterium *Sinorhizobium meliloti*. *BMC Genomics*. 2010;11(1):245.
11. Fuli X, Wenlong Z, Xiao W, Jing Z, Baohai H, Zhengzheng Z, Bin-Guang M, Youguo L. A genome-wide prediction and identification of intergenic small RNAs by comparative analysis in *Mesorhizobium huakuii* 7653R. *Front Microbiol*. 2017;8:1730.
12. Wilms I, Voss B, Hess WR, Leichert LJ, Narberhaus F. Small RNA-mediated control of the agrobacterium *tumefaciens* GABA binding protein. *Mol Microbiol*. 2011;80(2):492–506.
13. Lee K, Huang X, Yang C, Lee D, Ho V, Nobuta K, Fan JB, Wang K. A genome-wide survey of highly expressed non-coding RNAs and biological validation of selected candidates in agrobacterium *tumefaciens*. *PLoS One*. 2013;8(8):e70720.
14. López-Leal G, Tabche ML, Castillo-Ramírez S, Mendoza-Vargas A, Ramírez-Romero MA, Dávila G. RNA-Seq analysis of the multipartite genome of *rhizobium etli* CE3 shows different replicon contributions under heat and saline shock. *BMC Genomics*. 2014;15(1):770.
15. Kazmierczak MJ, Wiedmann M, Boor KJ. Alternative sigma factors and their roles in bacterial virulence. *Microbiol Mol Biol Rev*. 2005;69(4):527–43.
16. Raja I, Kumar V, Sabapathy H, Kumariah M, Rajendran K, Tennyson J. Prediction and identification of novel sRNAs involved in *Agrobacterium* strains by integrated genome-wide and transcriptome-based methods. *FEMS Microbiol Lett*. 2018;365(23):fny247.
17. González V, Santamaría RI, Bustos P, Hernández-González I, Medrano-Soto A, Moreno-Hagelsieb G, Janga SC, Ramírez MA, Jiménez-Jacinto V, Collado-Vides J, Dávila G. The partitioned *rhizobium etli* genome: genetic and metabolic redundancy in seven interacting replicons. *Proc Natl Acad Sci*. 2006;103(10):3834–9.
18. Storz G, Vogel J, Wassarman KM. Regulation by small RNAs in bacteria: expanding frontiers. *Mol Cell*. 2011;43(6):880–91.
19. Michaux C, Verneuil N, Hartke A, Giard JC. Physiological roles of small RNA molecules. *Microbiology*. 2014;160(6):1007–19.
20. Ceizel Borella G, Lagares A Jr, Valverde C. Expression of the *Sinorhizobium meliloti* small RNA gene *mmgR* is controlled by the nitrogen source. *FEMS Microbiol Lett*. 2016;363(9):fnw069.
21. Robledo M, Peregrina A, Millán V, García-Tomsig NI, Torres-Quesada O, Mateos PF, Becker A, Jiménez-Zurdo JI. A conserved α -proteobacterial small RNA contributes to osmoadaptation and symbiotic efficiency of rhizobia on legume roots. *Environ Microbiol*. 2017;19(7):2661–80.
22. Gottesman S, McCullen CA, Guillier M, Vanderpool CK, Majdalani N, Benhammou J, Thompson KM, FitzGerald PC, Sowa NA, FitzGerald DJ. Small RNA regulators and the bacterial response to stress. *Cold Spring Harb Symp Quant Biol*. 2006;71:1–11.
23. Azhikina TL, Ignatov DV, Salina EG, Fursov MV, Kaprelyants AS. Role of small noncoding RNAs in bacterial metabolism. *Biochem Mosc*. 2015;80(13):1633–46.
24. Sridhar J, Gunasekaran P. Computational small RNA prediction in bacteria. *Bioinform Biol Insights*. 2013;7:BBI–S11213.
25. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res*. 2007;35(suppl_2):W345–9.
26. Li L, Huang D, Cheung MK, Nong W, Huang Q, Kwan HS. BSRD: a repository for bacterial small regulatory RNA. *Nucleic Acids Res*. 2012;41(D1):D233–8.
27. Leinonen R, Sugawara H, Shumway M. International Nucleotide Sequence Database Collaboration The sequence read archive. *Nucleic Acids Res*. 2010;39(suppl_1):D19–21.
28. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J*. 2011;17:10–2.
29. McClure R, Balasubramanian D, Sun Y, Bobrovskyy M, Sumbly P, Genco CA, Vanderpool CK, Tjaden B. Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Res*. 2013;41(14):e140.
30. Tjaden B. De novo assembly of bacterial transcriptomes from RNA-seq data. *Genome Biol*. 2015;16(1):1.
31. Kery MB, Feldman M, Livny J, Tjaden B. TargetRNA2: identifying targets of small regulatory RNAs in bacteria. *Nucleic Acids Res*. 2014;42(W1):W124–9.
32. Hofacker IL. Vienna RNA secondary structure server. *Nucleic Acids Res*. 2003;31(13):3429–31.
33. Fruzangohar M, Ebrahimie E, Ogunniyi AD, Mahdi LK, Paton JC, Adelson DL. Correction: comparative GO: a web application for comparative gene ontology and gene ontology-based gene selection in bacteria. *PLoS One*. 2015;10(4):e0125537.
34. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–2.
35. Salamov VS, Solovyevand A. Automatic annotation of microbial genomes and metagenomic sequences. *Metagenomics and its applications in agriculture*. Hauppauge: Nova Science Publishers; 2011. p. 61–78.
36. Naville M, Ghuillot-Gaudeffroy A, Marchais A, Gautheret D. ARNold: a web tool for the prediction of rho-independent transcription terminators. *RNA Biol*. 2011;8(1):11–3.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://www.biomedcentral.com/submissions)

