Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Functional annotations in bacterial genomes based on small RNA signatures

Jayavel Sridhar[1] and Ziauddin Ahamed Rafi[1, *]

[1]Centre of Excellence in Bioinformatics, School of Biotechnology, Madurai Kamaraj University, Madurai 625021, Tamilnadu, India; Ziauddin Ahamed Rafi* – E-mail: zarafi@gmail.com; * Corresponding author

received March 12, 2008; accepted March 25, 2008; published April 04, 2008

Abstract:
One of the key challenges in computational genomics is annotating coding genes and identification of regulatory RNAs in complete genomes. An attempt is made in this study which uses the regulatory RNA locations and their conserved flanking genes identified within the genomic backbone of template genome to search for similar RNA locations in query genomes. The search is based on recently reported coexistence of small RNAs and their conserved flanking genes in related genomes. Based on our study, 54 additional sRNA locations and functions of 96 uncharacterized genes are predicted in two draft genomes viz., *Serratia marcesens Db1* and *Yersinia enterocolitica* 8081. Although most of the identified additional small RNA regions and their corresponding flanking genes are homologous in nature, the proposed anchoring technique could successfully identify four non-homologous small RNA regions in *Y. enterocolitica* genome also. The KEGG Orthology (KO) based automated functional predictions confirms the predicted functions of 65 flanking genes having defined KO numbers, out of the total 96 predictions made by this method. This coexistence based method shows more sensitivity than controlled vocabularies in locating orthologous gene pairs even in the absence of defined Orthology numbers. All functional predictions made by this study in *Y. enterocolitica* 8081 were confirmed by the recently published complete genome sequence and annotations. This study also reports the possible regions of gene rearrangements in these two genomes and further characterization of such RNA regions could shed more light on their possible role in genome evolution.

Key words: functional annotation; sRNA; KO; KOBAS; Bio-ontology; flanking genes

Abbreviations:
eca - *Erwinia caratovora atroseptica*, *eco* - *Escherichia coli* K12, *sma* - *Serratia marcescens* Db1, *ye* - *Yersinia enterocolitica* 8081, *ypk* - *Yersinia pestis* KIM, *yps* - *Yersinia pseudotuberculosis* Db1, KEGG - Kyoto Encyclopedia of Genes and Genomes, KO - KEGG Orthology, KOBAS - KO Based Annotation System.

Background:
Genome sequencing projects are exponentially adding complete genome sequence data to genome databases; but, the identification of functionally important regulatory regions and the rate of annotating functions to the coding genes still needs to improve due to lack of parallel experimental techniques. Massive accumulation of complete genome sequence data in public databases such as NCBI-GenBank creates an opportunity for comparative genomics based identification of functional annotations for draft sequences from known genomes using variety of data mining tools **[1]**. Initial functional characterizations of proteins were carried out with biochemical experiments, which can be extended by matching recently sequenced proteins/genes through computational studies **[2]**. Such theoretical functional annotations for the unknown query sequences are assigned based on the known characteristics of the reference set. Additionally functional annotations of proteins have also been attempted based on several other properties like, proteins sharing common domains connected via related multi domain proteins (super families); proteins in the same pathways **[3]**, networks, or complexes; or protein complexes in their expression pattern **[4]** and proteins correlated in their phylogenetic profiles **[5]**. Though many sophisticated techniques are employed, the identification of homologous relationship is one of the powerful techniques for functional annotations of unknown proteins from known data **[6]**. A rough estimate shows that well over 80% of our biological knowledge concerning protein sequences is inferred from homology **[7]**. Availability of intermediate grade complete genome sequences (without functional characterization) opens up

a prosperous way for comparative genomic analysis based annotations in related genomes **[8]**. This study uses two such draft genomes (at the time of initiating the work) for the functional annotations of proteins based on non-coding small RNA (sRNA) identification in prokaryotic genomes.

In recent years, non-coding RNAs (ncRNAs) have been identified to have variety of regulatory functions. Most of the non-coding sRNAs are of intergenic origin **[9, 10]**. Due to the lack of potent statistical signals their identification has become difficult **[11, 12]**. The redundancy and co-occurrence of sRNAs sandwiched between a pair of conserved flanked genes in prokaryotic genomes is reported **[10]**. This feature has been used to identify both homologous and non-homologous additional sRNA locations in 20 closely related *Enterobacteriaceae* genomes **[13]**. The locus of such additional sRNA locations (ASLs) follows genomic backbone continuity and gene synteny (gene order) conservation. These results also confirm that all experimentally known sRNAs and the ASLs fall within the intergenic regions and follow the redundancy and co-occurrence of sRNAs sandwiched between a pair of conserved flanked genes rule, especially, in *Enterobacteriaceae* genomes. The current study uses the above principle to identify ASLs and functionally annotate their respective conserved flanking genes pair in two draft completed genomes viz., *Serratia marcesens Db1* (*sma*) and *Yersinia enterocolitica* 8081 (*ye*). Preliminary gene prediction data reported as ORFs obtained from the respective genome sequencing groups of *sma* and *ye* in EMBL format **[14]** are analyzed and functional annotations of 96 such ORFs are presented. The Kyoto Encyclopedia of Genes and Genomes Orthology (KO) system has been used to compare the above functional assignments. Throughout the text, KEGG annotations for genome names and gene ids are used for convenience.

Methodology:

Identification of additional sRNA locations (ASLs)

The complete draft genome sequence data for *sma* and *ye* along with their primary gene predictions in EMBL format **[14]** are obtained from Sanger Centre **[15, 16]**. The nucleotide sequences of 60 known sRNAs of *E. coli* K12 (*eco*) (NC_000913), termed as sRNA template sequences (STS), are extracted from Ecocyc **[17]** database. Similar to our earlier work **[13]**, the *eco* STS are used to search for homologous ASLs in *sma* and *ye* genomes using genomic blast **[18, 19]** maintained at Sanger centre with default parameters (Database: genome assembly, Executable: BLASTN, Filter: Low complexity regions and report: top 100 alignments). Blast hits showing >80% identity (cut-off) alone are considered as homologous ASLs. As reported in our earlier study, a few of the *eco* STS had to be replaced with other known STS obtained from *eca* or *yps* genomes to obtain better ASL homologs. This search resulted in the identification of 54 ASLs in both *sma* and *ye* genomes collectively.

Identification of homologous protein sequence locations (HPSLs) that sandwich ASLs in *sma* and *ye*

The protein sequences of the genes that sandwich the known 60 *eco* sRNAs are extracted from KEGG **[20]**. Using these extracted protein sequences as templates, a search for homologous protein sequence locations (HPSLs) in *sma* and *ye* draft genomes are marked using TBLASTN (with default parameters). Although, this resulted in few/several HPSLs in each of the selected draft genomes, HPSLs that sandwich the already identified 54 ASLs alone are considered for further studies.

Selection of preliminary ORFs and assignment of functional annotations

The preliminary gene ids for the ORFs in the draft genomes that either fall or overlap within the HPSLs sandwiching the ASLs alone are extracted. Functional annotations for these preliminary gene ids are assigned based on consensus gene functions reported for the flanking genes in *eco* and other related genomes using KEGG. The sRNA specific conserved flanking gene pairs in 20 *Enterobacteriaceae* genomes already reported **[13]** are used in this study.

Genomic backbone continuity of the regions of interest (identified regions)

Although, the identified ASLs and their conserved flanking genes are picked up for functional annotations based on sequence homology, our earlier studies indicates that the conserved nature of these regions are retained only if the query regions are observed to be within the common genomic backbone. Hence, a multiple genome sequence alignment for the *eco*, *sma* and *ye* genomes is made with Mauve **[21]** aligners under default parameters (default seed weight = 15, determine LCBs, full alignment mode, Extend LCBs, minimum island size = 50, maximum backbone gap size = 50 and minimum backbone size = 50). This alignment is used to verify the genomic backbone continuity of the identified regions.

Example:

The STS of sRNA *tke1* of *eco* extracted from Ecocyc **[17]** database is used to search against the query genome *sma*. Genomic blast is used to identify ASLs within the query genomes. The protein sequences of the genes yfhK (gene id1: b2556) and purL (gene id2: b2557) that sandwich *tke1* sRNA in *eco* are also extracted from KEGG database. These protein sequences are used as template sequences for a similarity sequence search using

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

TBLASTN with default parameters in *sma.* The BLAST search will result in identification of several HPSLs. However, a careful analysis is carried out to extract only HPSLs that sandwich the corresponding *tke1* ASLs. The preliminary gene ids reported in the database for the ORFs located within the HPSLs sandwiching the *tke1* ASLs alone are extracted. These extracted preliminary gene ids refer to the ORFs sma3041 and sma3042 in *sma*. The *tke1* sRNA is already reported to be flanked by conserved yfhK/purL genes pair in sixteen enterobacterial genomes **[13]** and their gene ids are c3079/c3080 (ecc), ecs3422/ecs3423 in *ecs*, b2556/b2557 in *eco*, z3833/z3835 in *ece*, eca3257/eca3258 in *eca*, plu3313/plu3317 in *plu*, spa0302/spa0300 in *spt*, t0292/t0291 in *stt*, sty2811/sty2812 in *sty*, stm2564/stm2565 in *stm*, sf2603/sf2604 in *sfl*, s2775/s2776 in *sfx*, yptb2876/yptb2879 in *yps*, yp2541/yp2536 in *ypm*, ypo2916/ypo2921 in *ype* and y1313/y1309 in *ypk*. Similarly the corresponding *tke1* ASL identified in *sma* is observed to be flanked by two ORFs sma3041/sma3042. These ORFs sma3041 and sma3042 are further analyzed based on their sequence homology, length, KO numbers (from KOBAS) and based on the results obtained, sma3041 is functionally annotated as a *two component sensor protein* (yfhK) and sma3042 is functionally annotated as *phosphoribosyl formylglycinamidine synthase* (purL).

Similarly, the ORFs ye1035 and ye1032 that sandwich the *tke1* ASL identified in *ye* genome are also functionally annotated as *two component sensor protein* (yfthK) and *phosphoribosyl formylglycinamidine synthase* (purL) respectively. The multiple genome sequence alignment of this region comprising the identified *tke1* ASLs flanked by sma3041 and sma3042 in *sma* with *eco* and *ye* using Mauve **[21]** is shown (Figure 1). It can be seen from the figure 1 that the common backbone for the block containing *tke1* sRNA with gene ids b2556 and b2557 in *eco* (marked A) are conserved in both *sma* (marked B) and *ye* (marked C) genomes.

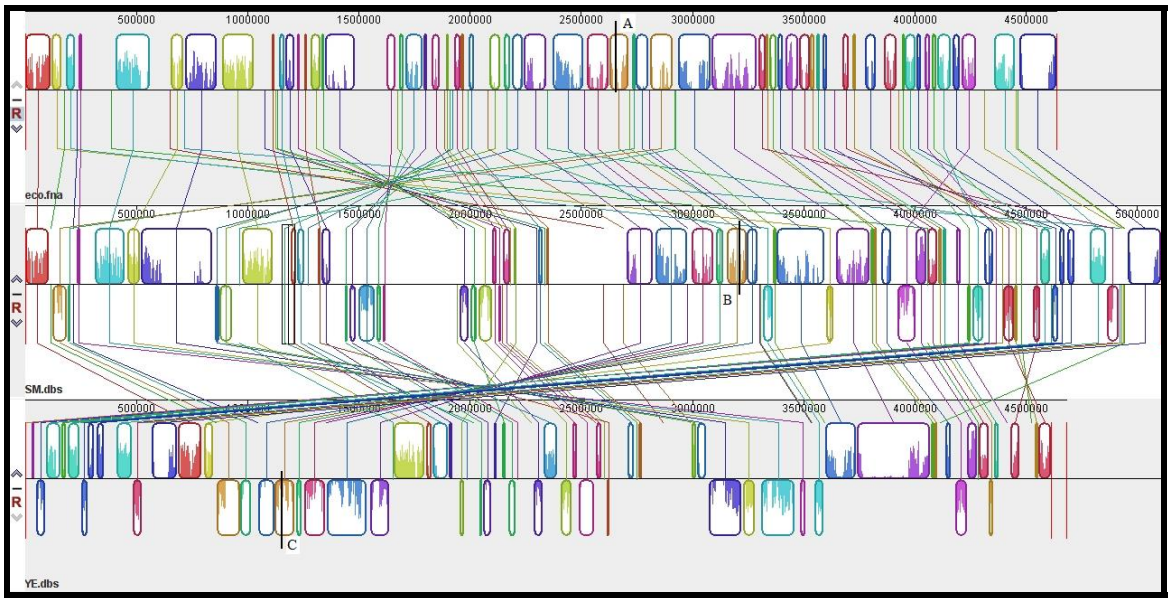

Figure 1: A multiple sequence alignment of *eco, sma* and *ye* genomes using Mauve. (A) The tke1 sRNA and its conserved flanking genes (b2556 and b2557) in *eco* are observed in 'A' block. (B) The identified ASL and its corresponding flanking genes (sma3041 and sma3042) are observed to be retained in the common genomic backbone marked 'B' in *sma* genome. (C) The identified ASL and its corresponding flanking genes (ye1032 and ye1035) are also observed to be retained in the common genomic backbone marked 'C' in *ye* genome.

Results:

The combined results of the BLAST and TBLASTN as described above could identify 54 ASLs with their corresponding preliminary gene ids for the ORFs falling within *sma* and *ye*. The results are tabulated (Table 1a and Table 1b in supplementary material). The table lists known sRNAs selected, the coordinates of ASLs identified in *sma* and *ye* genomes, their strand orientations, the reference genome from which STS are extracted for the search and the identified preliminary flanking gene ids. Based on consensus functions reported for the conserved flanking genes on either side of the known sRNAs, in closely related enterobacterial genomes, the functional annotations for 96 id-affixed

genes as described above are affixed and tabulated (Table 2a and Table 2b under supplementary material). These results also include the percentage of sequence identity and similarity of the id-affixed genes with their template protein sequences used. A perusal of the table indicates that most of the functionally annotated gene ids (except for a few) have high percentage of sequence identity and sequence similarity, indicating the high homologous nature between the template regions and the identified regions. The results obtained using KOBAS are also presented in this table.

A comparison of functional annotations using KO terms

KEGG Orthology (KO) terms or numbers are used to differentiate the universe of all genes in all organisms available in the KEGG database into groups of functionally identical genes (orthologs). KO based automated functional prediction systems is a specialized ontology system developed more specifically for prokaryotic annotations. The recently described KO based annotation system (KOBAS) **[22]** is used to annotate the unknown query (protein) sequences. In order to compare the functional annotations made from this study, the functional annotations of the preliminary gene ids using KO terms are also obtained as described below. The nucleotide sequences of all the identified preliminary gene ids from *sma* and *ye* are extracted using the Emboss *extractseq* code obtained from Embosswin **[23]** 2.10.0.-win-0.8 **[24]**. Strand orientations of the sequences in complement strand are reversed using *revseq* application. The extracted sequences of all these genes with their exact strand orientations (+ or - strand) in fasta format is given as input for KOBAS an automated functional analysis program. The results of KOBAS will contain a major orthologous gene id, obtained from KEGG-GENES **[20]** database, and is considered as a functionally equivalent gene to that of the selected gene id. The KOBAS will assign the KO number under which the gene/protein function grouped. Based on the KO terms, the functions of the specific gene id can be assigned. These KO based functional annotations are used to compare the functional predictions made from the current study.

These KO number under which the identified gene id is grouped with its major orthologous hit (top hit by KOBAS) from KEGG-GENES database is presented. Based on the KO number, the functional annotation of the orthologous group under which the identified gene id is classified is also presented. Since *sma* and *ye* are only draft genomes, the KEGG database is yet to add KO terms for all the preliminary predicted gene ids. The latest version of KEGG database contains only 8032 KO terms and any further additions to this database for the *sma* and *ye* genomes can be used for more confirmations of functional annotations. In the absence of a KO term for a specific gene, KOBAS cannot assign its orthologous group and its function. Hence, only 65 functional annotations alone are presented in the table using KOBAS. All the functional predictions made from this study in *Y. enterocolitica* 8081 are confirmed by the recently published complete genome sequence and annotations **[25]**.

Identification of non-homologous ASLs using conserved flanking genes in *sma* and *ye*

The STS of sRNAs *rygC*, *ryeF*, *sroC* and *sroE* could not identify any ASLs in *sma* and *ye* based on sequence search indicating that the sequences of these sRNAs are non-homologous. However, using the protein sequences that sandwich these sRNAs (*rygC*, *ryeF*, *sroC* and *sroE)* from *eco* as template sequences, a search for HPSLs in *sma* and *ye* is made. Since the genomic backbones within these regions are continuous, an attempt has been made to identify the ASLs in this genome. Table 3 (see supplementary material) gives the preliminary gene ids of the conserved flanking genes that could sandwich the listed sRNA whose sequences are non-homologous. Such non-homologous sRNA regions containing conserved flanking genes have already been reported for 21 different classes of enterobacterial small RNAs **[13]** within the members of the *Enterobacteriaceae* family. Computational approaches like INFERNAL (Rfam) and experimental reports propose these regions as corresponding sRNA locations even in the absence of sequence homology. Hence, for example, we propose that *rygC* sRNA is possible between the conserved flanking genes pair *ye3399* and *ye3400* in *ye* genome, whose sequences are non-homologous to any of the STS obtained from *eco, eca, yps,* etc.

Significance of this method

The above mentioned sRNA based anchoring or coexistence methodology employs simple nucleotide and protein sequence similarity searches along with the knowledge of genomic backbone continuity information towards identification of the functionally important regions. Although the blast searches for the draft genomes will produce redundant similarity hits, the current anchoring of ASLs could effectively identify both the ASLs and their corresponding conserved flanking genes with their functions. This study attempts to link known RNA data, flanking gene conservation and backbone retention to identify the gene functions in draft genomes.

Discussion:

The occurrence of ASLs with conserved flanking genes in *sma* and *ye* indicate that the sRNAs are retained within the conserved flanking genes in related genomes. Only 27

ASLs could be identified from the known 60 STS in each of the *sma* and *ye* genomes. This could be probably because of the retention capacity of specific regions (sRNAs with their corresponding flanking genes) between different genuses within the family varies. If the common backbone is retained, the retention capacities of these regions are high. A few of the ASLs are observed to contain a single conserved flanking gene (marked * in Table 1a and Table 1b in supplementary material). However, the ASLs of these regions are confirmed based on high homologous nature of the STS. This suggests that intergenic regions containing non-coding transcripts may possibly involve in genome shuffling and gene rearrangements. Such ASLs containing a single conserved flanking gene has been reported as possible integration sites for 'alien' genetic pools **[26]**.

Identification of four non-homologous ASLs based on orthologous flanking gene pairs (Table 3 under supplementary material) in the query genomes is based on an earlier study **[13]** that enterobacterial small RNAs are identified to have entirely distinct non-homologous sequences but with non-homologous orthologues. The KO term identification agrees with our functional annotation results for 65 preliminary gene ids. The current anchoring of ASLs to functionally annotate preliminary gene ids could succeed in annotating 96 genes indicating a clear advantage of the anchoring technique compared to other existing ontology based methods. It must be emphasized here that a few of the identified ASLs have already been computationally predicted as small ORF's in the draft genomes. These small ORF having distinct transcription signals must be sRNAs and they cannot be coding genes. Similar anchoring technique involving flanking genes or specific operon or gene cluster based RNA predictions may evolve as a new method for locating sRNAs in closely related genomes. Similar applications for the identification of ncRNAs, conserved flanking genes and their functional annotations in closely related *Archaebacteria* and *Eukaryotes* are also being attempted.

Conclusion:

Functional annotations for coding genes and sRNA predictions in ongoing genome projects can be done with genomic backbone retention information obtained from related genomes. Although this work has been started much earlier, all the functional predictions made by this study have been confirmed by the recently published complete genome sequence and annotations for *Y. enterocolitica* 8081 [NCBI-Refseq Accession number NC_008800]. The above study also indicates the possible identification of non-homologous sRNA regions even in the absence of sequence similarity. Identification of sRNAs with a single conserved flanking gene has been proposed to be regions of 'alien' gene pool integration sites. Characterization of these sRNA regions involved with foreign gene integration and such 'alien' genetic pool may elucidate their role in pathogenesis and survival of the pathogen.

Acknowledgement:

We thank the Department of Biotechnology (DBT), Govt. of India New Delhi for funding the "Centre of Excellence in Bioinformatics". One of the authors, JS thanks the DBT for research fellowship. We acknowledge the Sanger Institute for the use of '*sma*' and '*ye*' draft genome data.

References:

[01] T. Nandi *et al., J. Biosci.,* 27: 15 (2002) [PMID: 11927774]

[02] P. Bork *et al., J. Mol. Biol.,* 283: 707 (1998) [PMID: 9790834]

[03] Y. Zheng *et al., Genome Biol.,* 3: 0060 (2002) [PMID: 12429059]

[04] T. K. Attwood, *Brief Bioinformatics,* 1: 45 (2000) [PMID: 11466973]

[05] E. M. Marcotte *et al., Nature,* 402: 83 (1999) [PMID: 10573421]

[06] L. Stein, *Nat. Rev. Genet.,* 2: 493 (2001) [PMID: 11433356]

[07] D. G. George, *Meth. Enzymol.*, 266: 41 (1996) [PMID: 8743676]

[08] R. W. Blakesley *et al., Genome Res.,* 14: 2235 (2004) [PMID: 15479945]

[09] K. M. Wassarman *et al., Trends Microbiol.,* 7: 37 (1999) [PMID: 10068996]

[10] R. Hershberg *et al., Nuc. Acids Res.,* 31: 1813 (2003) [PMID: 12654996]

[11] E. Rivas *et al., Curr. Biol.,* 11: 1369 (2001) [PMID: 11553332]

[12] S. R. Eddy, *Cell,* 109: 137 (2002) [PMID: 12007398]

[13] J. Sridhar & Z. A. Rafi, *OMICS*, 11: 74 (2007) [PMID: 17411397]

[14] T. Kulikova *et al., Nuc. Acids Res.,* 35: D16 (2007) [PMID: 17148479]

[15] ftp://ftp.sanger.ac.uk/pub4/pathogens/sma

[16] ftp://ftp.sanger.ac.uk/pub4/pathogens/ye

[17] P. D. Karp *et al., Nuc. Acids Res.* 30: 56 (2002) [PMID: 11752253]

[18] http://www.sanger.ac.uk/cgi-bin/blast/submitblast/s_marcescens

[19] http://www.sanger.ac.uk/cgi-bin/blast/submitblast/y_enterocolitica

[20] M. Kanehisa *et al., Nuc. Acids Res.,* 32: D277 (2004) [PMID: 14681412]

[21] A. C. Darling *et al., Genome Res.,* 14: 1394 (2004) [PMID: 15231754]

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

[22] X. Mao *et al., Bioinformatics,* 21: 3787 (2005) [PMID: 15817693]
[23] P. Rice *et al., Trends Genet.*, 16: 276 (2000) [PMID: 10827456]
[24] ftp://ftp.ebi.ac.uk
[25] N. R. Thomson *et al., PLoS Genet.,* 2: e206 (2006) [PMID: 17173484]
[26] J. Sridhar & Z. A. Rafi, *In Silico Biol.*, 7: 0053 (2007)

Edited by P. Kangueane
Citation: Sridhar and Rafi, Bioinformation 2(7): 284-295 (2008)

Supplementary material

Tables:

sRNA	Start	End	orientation	length nt	Genome from which STS are extracted	flanking genes id
csrB	3377288	3377636	-	349	eco	sma3187/sma3190
csrC	4371671	4371785	-	115	eco	sma4096/sma4098
ffs	388557	388672	+	116	eco	sma0360/sma0362
gcvB	3384355	3384555	+	201	eco	sma3196/sma3197
micF	2809693	2809725	-	33	eco	sma2659/sma2652
oxyS	4269827	4269867	-	41	eco	sma3997/sma4002
rnpB	3795714	3796080	-	367	eco	sma3569/sma3570
rprA	1515148	1515194	-	47	eco	sma1448/sma1449
rtt	2101702	2101831	+	130	eco	*/sma1982
rtt	2101527	2101656	-	130	eco	*/sma1982
rybB	974791	974868	-	77	eco	sma0925/sma0930
ryeA	1278736	1278883	-	147	eco	sma1216/sma1217
rye	3121483	3121568	-	86	eco	sma2954/sma2955
sraE	3418083	3418175	-	93	eca	sma3223/sma3228
spf	4373220	4373328	-	109	eco	sma4098/sma4099
sraA	419341	419369	+	29	eco	sma0343/sma0344
sraB	1226373	1226453	-	81	eco	sma1158/sma1159
sraD	188225	188299		75	eco	sma0166/sma0167
sraF	3780424	3780611	+	188	eco	sma3551/sma3553
sraH	3802264	3802317	-	54	eco	sma3576/sma3577
sraI	4147535	4147593		59	eco	*/sma3898
ssrA	3251988	3252351	+	364	eco	sma3078/ *
ssrS	3518913	3519095	+	183	eco	sma3318/sma3319
tke1	3214595	3214762	+	168	yps	sma3041/sma3042
tp2	3636376	3636534	+	159	eco	sma3419/sma3420
tpke70	1209585	1209953	-	368	eco	*/sma1142
t44	3360413	3360472		60	eco	sma3178/sma3179

Table 1a: List of additional sRNA locations identified in *S.marcescens* Db1 genome (* - indicates the ASL identified containing single conserved flanking gene).

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

sRNA	Start	End	orientation	length nt	Genome from which STS are extracted	flanking genes id
c0465	2802523	2802547	-	25	*eco*	ye2575/ ye2576
csrB	3595388	3595696	+	309	*ypk*	ye3291/ ye3294
csrC	31332	31582	+	251	*ypk*	ye0022/ *
Ffs	3388105	3388006	-	100	*ypk*	ye3115/ ye3116
gcvB	3602753	3602955	+	203	*ypk*	ye3300/ ye3303
Is102	2564059	2564206	-	148	*eco*	ye2376/*
micF	1571499	1571589	+	91	*ypk*	ye1398/ ye1401
rnpB	4067726	4067362	-	365	*ypk*	ye3720/ ye3724
rprA	2382211	2382303	.	93	*ypk*	ye2175/ ye2176
Rtt	2460983	2461139	.	157	*eco*	ye2247/ ye2253
Rtt	2460809	2460965	-	157	*ypk*	ye2247/ ye2253
rybB	1661052	1660972	+	81	*eco*	ye1477/ ye1478
ryeA	1991664	1991493	+	172	*ypk*	ye1798/ ye1799
Rye	1229573	1220660	+	88	*eco*	ye1099/ ye1100
sraE	3638400	3638314	-	87	*eco*	ye3327/ ye3331
Spf	30017	30134	.	118	*eco*	ye0021/ y20022
sraA	3408175	3408231	-	57	*eco*	ye3132/ ye3133
sraD	966506	966434	+	73	*eco*	ye0838/ ye0839
sraF	1966847	1966735	-	112	*ypk*	ye1772/ ye1773
sraH	4074104	4074141	-	37	*ypk*	ye3732/ ye3733
sraI	4382210	4382153	+	57	*eco*	*/ ye4024
ssrA	1116170	1116457	-	287	*eco*	ye0993/ ye0994
ssrS	3711867	3712049	-	183	*ypk*	ye3398/ ye3399
tke1	1156238	1156074	+	165	*ypk*	ye1032/ ye1035
tp2	809140	809298	-	159	*eco*	ye0698/ ye0699
tpke11	3242087	3242129	-	43	*eco*	ye0609/ ye0610
t44	3586880	3586741	-	140	*ypk*	ye3284/ ye3285

Table 1b: List of additional sRNA locations in *Y.enterocolitica* 8081 genome (* - indicates the ASL identified containing single conserved flanking gene).

Id-affixed genes	Annotation based on KOBAS			Expected functional annotation based on ASL anchoring method	Id-affixed gene similarity with the Template protein sequence			Predicted putative function
	KO no.	Top hit by KOBAS	Functional annotation by KOBAS		Template protein id	% of identity	% of similarity	
sma3190	ko not found	---	---	secY interacting protein	b2793	55	69	secY interacting protein
sma3187	ko not found	---	---	yqcC related protein	b2792	36	55	yqcC related protein
sma4099	ko2335	yp0018	dna polymerase I	dna polymerase I	b3863	72	80	dna polymerase I
sma4098	ko3978	yp0020	gtp binding protein	gtp binding protein	b3865	85	89	gtp binding protein
sma0360	ko not found	---	---	methyl transferase	b0454	69	80	methyl transferase
sma0362	ko5839	plu3853	haemolysin expression modulating protein	haemolysin expression modulating protein	yptb0978	92	98	haemolysin expression modulating protein
sma3196	ko3566	t2890	gcva regulatory protein	regulator protein of glycine cleavage pathway	b2808	91	95	regulator protein of glycine cleavage pathway
sma3197	ko not	---	---	ygdI related protein	b2809	81	90	ygdI related protein

sma2659	ko not found	---	---	outer membrane protein	b2215	63	78	outer membrane protein
sma2662	ko0936	y2968	two component sensor protein	two component sensor protein	b2216	40	56	two component sensor protein
sma3997	ko4761	eca4243	hydrogen peroxide inducible genes	hydrogen peroxide inducible genes	b3961	89	93	hydrogen peroxide inducible genes
sma4002	ko1755	stm4123	arginosuccinate lyase	arginosuccinate lyase	b3960	84	87	arginosuccinate lyase
sma3569	ko6911	stm3236	putative cytoplasmic protein	tdca related regulatory protein/cytoplasmic related protein	y0115	71	80	cytoplasmic related protein
sma3570	ko7056	yp3801	putative tetrapyrrole methylase	tetrapyrrole methylase/glycerate kinase	y0117	78	80	tetrapyrrole methylase
sma1449	ko not found	---	---	ydik related membrane protein	b1688	59	71	ydik related membrane protein
sma1448	ko1007	y1930	phosphoenol pyruvate synthase	phosphoenol pyruvate synthase	yptb2318	87	93	phosphoenol pyruvate synthase
sma1982	ko1433	eca2333	Formyl tetrahydrofolate deformylase	tetrahydrofolate deformylase related protein	b1232	84	90	Tetrahydrofolate deformylase
sma0925	ko not found	---	---	putative membrane protein	yptb1356	65	74	putative membrane protein
sma0930	ko7085	yp1266	potassium channel protein	transport /channel protein	b0847	78	83	channel protein
sma1217	ko not found	---	---	integrase related protein	stm1871	60	60	integrase related protein
sma1216	ko not found	---	---	Putative periplasmic/exported protein	stm1873	59	78	putative periplasmic protein
sma2954	ko1423	yp2721	putative protease	putative protease	b2081	80	88	putative protease
sma2955	ko7029	yp2722	putative diacylglyerol kinase	glycerol kinase	yptb2081	68	81	glycerol kinase
sma3223	ko1909	yp2864	long-chain-fatty-acid--[acyl-carrier-protein] ligase / acyl-[acyl-carrier-protein]-phospholipid o-acyltransferase	long-chain-fatty-acid--[acyl-carrier-protein] ligase / acyl-[acyl-carrier-protein]-phospholipid o-acyltransferase	b2836	68	81	long-chain-fatty-acid--[acyl-carrier-protein] ligase / acyl-[acyl-carrier-protein]-phospholipid o-acyltransferase
sma3228	ko2529	yp2862	Galactose operon repressor; laci regulator	Transcriptional regulatory protein/laci family	b2837	76	85	Transcriptional regulatory protein galactose operon repressor, laci
sma1158	ko not found	---	---	maf like-inhibitor of septum formation	b1087	70	79	maf like-inhibitor of septum formation
sma1159	ko7040	yp2260	predicted metal binding protein	metal binding protein	b1088	75	83	metal binding protein
sma0166	ko1919	yp0385	glutamate-cysteine ligase	glutamate cysteine ligase	b2688	76	85	glutamate cysteine ligase
sma0167	ko7173	eca3362	autoinducer 2 production protein	autoinducer related protein	b2687	84	90	Autoinducer related protein

"Table 2a. (Continued)"

sma3551	ko0540	c3845	hypothetical oxidoreductase	hypothetical oxidoreductase /nad dehydrogenase	b3087	72	83	Hypothetical oxidoreductase
sma3553	ko not found	---	---	putative transmembrane protein	b3088	76	86	putative transmembrane protein
sma3576	ko not found	---	---	sigma cross reacting protein	b3209	66	77	sigma cross reacting protein
sma3577	ko not found	---	---	aerobic respiratory sensor protein	b3210	73	81	aerobic respiratory sensor protein
sma3898	ko3825	s4299	putative acetyl transferase	putative acetyl transferase	b3441	53	69	putative acetyl transferase
sma3078	ko3664	eca0836	ssra-binding protein	ssrA- binding protein	b2620	78	85	ssrA- binding protein
sma3318	ko not found	---	---	putative cytoplasmic protein	b2910	86	91	putative cytoplasmic protein
sma3319	ko1934	ypo0913	putative 5-formyltetrahydrofolate cyclo-ligase-family protein	putative ligase	b2912	65	75	putative ligase
sma3041	ko7711	---	---	2 component sensor protein	b2556	79	84	two component sensor protein
sma3042	ko1952	yptb2879	phosphoribosylformylglycinamidine synthase	Phosphoribosyl formylglycinamidine synthase	b2557	87	92	phosphoribosylformylglycinamidine synthase
sma3419	ko0163	yptb0713	pyruvate dehydrogenase decarboxylase component	pyruvate dehydrogenase decaryboxylase component	b0114	89	95	pyruvate dehydrogenase decarboxylase component
sma3420	ko5799	yp0264	pyruvate dehydrogenase complex repressor	pyruvate dehydrogenase complex repressor	b0113	87	92	pyruvate dehydrogenase complex repressor
sma0011	ko3686	c0020	chaperone protein dnaJ	chaperone protein dnaJ	b0015	82	84	chaperone protein dnaJ
sma0010	ko4043	yp3712	chaperone protein dnaK	chaperone dnaK	b0014	90	93	chaperone dnaK
sma1142	ko2457	yptb2490	Lipid a biosynthesis lauroyl acyltransferase	acp dependent/putative acyltransferase	b2378	52	74	putative aceyltransferase
sma3178	ko2967	yp2807	30s ribosomal protein s2	ribosomal protein small subunit	b0169	93	95	ribosomal protein small subunit
sma3179	ko1265	yp2808	methionine aminopeptidase	methionine aminopeptidase	b0168	84	92	methionine aminopeptidase
sma0343	ko3544	yp0775	atp-dependent clp protease	ATP-dependent clp protease	b0438	94	97	ATP- dependent clp protease
sma0344	ko1338	yp0776	atp dependent protease la	ATP- dependent protease lon	b0439	91	95	ATP- dependent lon protease

Table 2a: List of annotated id-affixed genes in *S.marcescens* Db1 (sma)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Id-affixed genes	Annotation based on KOBAS: KO no.	Top hit by KOBAS	Functional annotation by KOBAS	Expected functional annotation based on ASL anchoring method	Id-affixed gene similarity with the Template protein sequence: Template protein id	% of identity	% of similarity	Predicted putative function
ye3133	Ko3544	yp0775	ATP-dependent clp protease	ATP- dependent clp protease	b0438	93	96	ATP- dependent clp protease
ye3132	Ko1338	yp0776	ATP- dependent lon protease	ATP- dependent lon protease	b0439	89	94	ATP- dependent lon protease
ye2575	Ko5874	yptb2401	methyl ccepting chemotaxis protein	methyl accepting chemotaxis protein	b1886	45	58	methyl accepting chemotaxis protein
ye2576	Ko 3408	yp1797	chemotaxis protein chew	chemotaxis protein cheW	b1887	75	83	chemotaxis protein cheW
ye3294	Ko not found	---	---	secY interacting protein	b2793	59	71	secY interacting protein
ye3291	Ko not found	---	---	putative cytoplasmic protein	yptb3010	49	65	putative cytoplasmic protein
ye0021	Ko 2335	yp0018	dna polymerase I	dna polymerase I	yptb0018	87	90	dna polymerase I
ye3115	Ko5839	yp0793	haemolysin expression modulating protein	haemolysin expression modulating protein	yptb0978	100	100	haemolysin expression modulating protein
ye3116	Ko7443	yp0792	methyl guanine methyl transferase	methyl guanine methyl transferase	yptb0977	56	70	methyl guanine methyl transferase
ye3300	Ko3566	yp2822	glycine leavage system transcriptional activator, gcvA	gcvA transcriptional regulator	yptb3017	87	93	gcvA transcriptional regulator
ye3303	Ko1766	yptb3018	putative aminotransferase	putative aminotransferase	yptb3018	82	84	putative aminotransferase
ye2376	Ko7285	yp1916	putative outer membrane lipo protein	outer memnbrane lipo/fluffing protein	b1806	68	82	outer membrane lipo protein
ye1398	Ko0936	y2968	two component sensor protein	two component sensor protein	yptb1259	77	84	two component sensor protein
ye1401	Ko not found	---	---	outer membrane protein	yptb1261	83	91	outer membrane protein
ye3720	Ko not found	---	---	lysr type regulatory protein	yptb3491	95	98	lysr – type regulatory protein
ye3724	Ko not found	---	---	tetrapyrole methylase	yptb3492	81	82	tetrapyrrole methylase
ye2175	Ko not found	---	---	putative membrane protein	yptb2317	52	69	putative membrane protein
ye2176	Ko1007	y1930	phosphoenol pyruvate synthase	phosphoenol pyruvate synthase	yptb2318	91	94	phosphoenol pyruvate synthase
ye2253	Ko not found	---	---	lysr type regulatory protein	b2095	87	92	lysr – type regulatory protein
ye2247	Ko1433	yp1970	formyltetrahydrofolate deformylase	formyltetrahydrofolate deformylase	yptb2397	91	94	formyltetrahydrofolate deformylase
ye1477	Ko not found	---	---	putative membrane protein	yptb1356	72	78	putative membrane protein
ye1478	Ko7085	yp1266	potassium channel family protein	transport/channel protein	yptb1357	89	90	channel protein

ye1798	Ko not found	---	---	---	yptb1662	74	82	putative periplasmic protein
ye1799	Ko not found	---	---	putative integrase	sty2077	46	67	Putative integrase
ye1099	Ko7029	yp2722	diacylglycerol kinase	diacylglycerol kinase	yptb2821	76	87	diacylglycerol kinase

"Table 2b. (Continued)"

ye1100	Ko1423	yp2721	putative protease	putative protease	yptb2820	94	95	putative protease
ye3327	ko1909	yptb3042	long-chain-fatty-acid-phospholipid o-acyltransferase	long-chain-fatty-acid-phospholipid o-acyltransferase	yptb3042	69	82	long-chain-fatty-acid-phospholipid o-acyltransferase
ye3331	ko2529	yp2862	galactose operon repressor	galR-regulatory protein	yptb3044	77	86	galR-regulatory protein
ye0022	ko3978	yp0020	Predicted gtpase	gtp binding protein	yptb0019	94	98	gtp binding protein
ye0839	ko7173	yp0386	autoinducer-2 production protein	autoinducer related protein	yptb0830	88	92	Autoinducer related protein
ye0838	ko1919	yptb0829	glutamate-cysteine ligase	glutamate cysteine ligase	yptb0829	77	87	glutamate cysteine ligase
ye1772	ko not found	---	---	yebN family membrane /transport protein	yptb1630	78	83	yebN family membrane protein
ye1773	ko not found	---	---	putative membrane protein	yptb1631	69	75	putative membrane protein
ye3732	ko not found	---	---	sigma cross reacting protein	yptb3498	62	76	sigma cross reacting protein
ye3733	ko not found	---	---	aerobic respiratory sensor protein	yptb3500	71	81	aerobic respiratory sensor protein
ye4024	ko3825	spa3398	putative acetyl transferase	putative acetyl transferase	b3441	53	72	putative acetyl transferase
ye0994	ko3664	yp1055	ssra binding protein	ssra binding protein	yptb1135	79	86	ssra binding protein
ye0993	ko not found	---	---	integrase related protein	b2622	48	65	integrase related protein
ye3398	ko not found	---	---	putative cytoplasmic protein	yptb3187	85	91	putative cytoplasmic protein
ye3399	ko1934	ypo0913	putative 5-formyltetrahydrofolate cyclo-ligase-family protein	Putative methynyl tetrahydrofolate synthetase/ ligase family protein	yptb3188	76	83	Putative methynyl tetrahydrofolate synthetase
ye1035	ko7711	yp2541	two component system sensor kinase	two component sensor kinase	yptb2876	84	88	two component sensor kinase
ye1032	ko1952	yptb2879	phosphoribosylformylglycinamidine synthase	phosphoribosylformylglycinamidine synthase	yptb2879	93	95	phosphoribosylformylglycinamidine synthase
ye0698	ko5799	yp0264	pyruvate dehydrogenase	pyruvate dehydrogenase	b0113	97	99	pyruvate dehydrogenase
ye0699	ko0163	yptb0713	putative dehydrogenase	putative dehydrogenase	b0114	95	98	putative dehydrogenase
ye0609	ko4043	yp3712	molecular chaperone protein dnak	molecular chaperone protein dnak	b0014	90	93	molecular chaperone protein dnak
ye0610	ko3686	yp3711	chaperone protein	chaperone protein	b0015	79	81	chaperone protein dnaj

ye3285	ko1265	yptb3004	dnaj methione aminopeptidase	dnaj methione aminopeptidase	yptb3004	98	99	methione aminopeptidase
ye3284	ko2967	yp2807	30s ribosomal subunit	30s ribosomal subunit	yptb3003	96	98	30s ribosomal subunit

Table 2b: List of annotated id-affixed genes in *Y.enterocolitica* 8081 (ye).

sRNA	flanking gene id1	Proposed function	flanking gene id2	Proposed function
rygC	ye3399	Predicted Ligase	ye3400	Phosphoglycerate dehydrogenase
ryeF	ye2402	yecP Methyl transferase	ye2403	Copper homeostasis protein
sroC	ye2992	gltJ, transport protein	ye2991	Periplasmic binding protein
sroE	ye1074	Histidine tRNA synthetase	ye1073	GcpE, Diphosphate synthase

Table 3: Identified flanking genes pair between which the corresponding non-homologous ASL is expected in *Yersinia enterocolitica* genome.