Prediction of CRISPR-Cas Targets in Salmonella Typhimurium

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE

PILANI

Palash Sethi
2014B1A80260P
Under the supervision of
Dr. Sandhya Amol Marathe
Dept. Of Biological Sciences
BITS Pilani, Pilani Campus

May 2019

CERTIFICATE

This is to certify that the Thesis entitled - *Prediction of CRISPR-Cas Targets in Salmonella Typhimurium* and submitted by *Palash Sethi/2014B1A80260P* in partial fulfillment of the requirement of BITS F424T Thesis embodies the work done by him under my supervision.

Date - 7 May 2019

Dr. Sandhya Marathe
Assistant Prof
Dept. of Biological Sciences, BITS Pilani

Abstract

CRISPR-Cas systems are abundant antiviral defense systems of bacteria and archaea. CRISPR RNAs are special RNAs that contain spacer sequences, which help in locating and binding of the Cascade complex to the incoming external viral DNA, which gets regulated due to complex formation. In a recent studies, self-targeting of the lasR mRNA based on only nine nucleotide complementarity between the CRISPR RNA and the target mRNA has been shown. We assume the hypothesis that CRISPR-Cas systems regulate the translation of self mRNAs and try to predict the genes that will be regulated by CRISPR 1 and CRISPR 2 systems in Salmonella Typhimurium str. 14028S. We use various methodologies from RNAi prediction tools that predict miRNA and siRNA targets. The predicted regulated gene set is further analysed to infer information about the biological pathways that might get effected due to self targeting of genes by CRISPR-Cas systems.

Acknowledgement

I am thankful to Dr. Sandhya Amol Marathe for acting as my thesis advisor and helping me throughout this project. I am also thankful to bioinformaticians all over the world, whose open source tools push the boundaries of biological research.

Contents

Introduction

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) is an immunity system that was discovered in E. Coli. The CRISPR system consists of short repeats of palindromic DNA which is interspaced by non identical spacer sequences.

In early 2000s, it was found that these spacer sequences were identical to viral DNA, especially bacteriophage DNA. A number of genes, called the Cas genes (CRISPR associated) were also identified. The Cas genes transcribe and translate to Cas proteins. These Cas genes are essentially helicases and nucleases which unwind and cut DNA respectively. Along with CRISPR sequences, they form CRISPR-Cas system. Spacer sequences direct the CRISPR-Cas complex towards a viral DNA, resulting into degradation of viral DNA by Cas proteins.

It has to be noted that CRISPR-Cas system provides acquired immunity to bacteria. This is because the spacer sequences are derived from previous viral infections, and any such viral infection leads to an addition of spacer sequence in the spacer array.

Until recently, it was observed that CRISPR-Cas system only effects external viral DNA. But it has been observed that CRISPR-Cas system can also regulate self genes, in addition to viral DNA.

The study of these endogenous gene targets of CRISPR-Cas system is essential, as these genes may be involved in various biological pathways such as biofilm formation, pathogenecity etc in prokaryotes, similar to how miRNA and smallRNA regulate the transcription of various endogenous genes.

This thesis is an attempt to predict the self gene targets of CRISPR 1 and CRISPR 2 systems in Salmonella Typhimurium str. 14028s. Various methodologies from RNAi prediction tools have inspired this project. This report has been organised in the following manner

1. Chapter 1 describes the important features that are used to predict gene targets in mirna:mrna systems.

2. Chapter 2 and 3 discuss the features of free energy and target accessibility.

3. Chapter 4 describes the extended mean free energy, a feature that is used to shortlist putative target genes.

4. Chapter 5 describes the pathway that the shortlisted genes are involved in.

5. Appendix contains the shortlisted genes in CRISPR 1 and CRISPR 2 systems

Chapter 1

Common features of RNA target prediction tools

Many RNAs (generally, non coding RNAs) have been found to regulate target RNAs by base pair interactions. Examples vary from small RNAs (sRNAs) to small interfering RNAs (siRNAs), and microRNAs (miRNAs). In the following sections, we briefly describe the features that are used to predict RNA-RNA interactions[1]. These features are later elaborated in further chapters.

1.1 Duplex Stability

RNAs interact to form a duplex. Initially the two RNAs are present as individual RNAs, in their secondary structures. When they interact, they lose their individual structures and form a duplex secondary structure. This can be thought of as a chemical reaction in which two RNAs combine to form a single RNA. The Gibbs Free Energy of the forward reaction can be used as a measure of stability of the duplex structure. Since, it is difficult to measure free energy directly, we calculate the change in free energy. The energy calculation of the duplex structure is essentially a restricted and specialized version of full RNA secondary structure prediction.

1.2 Site Accessibility

Site Accessibility is a measure of the easeness with which a miRNA can find and hybridize with a mRNA target. Following translation, mRNA exists in a secondary structure which can meddle with a miRNA's capacity to tie to the target site.

miRNA:mRNA hybridization includes a two-step procedure in which a miRNA ties first to a short open locale of the mRNA. The mRNA secondary structure then unfolds as the miRNA completes binding to a target. Therefore, to assess if the mRNA is a putative target of miRNA, the amount of energy required to make the target site accessible to a miRNA can be evaluated. Accessibility of the target site is the energy required to make target site single stranded. It is defined as the difference between the energy of the ensemble of all structures and the energy of the ensemble of structures, where the target site is single stranded.

1.3 Seed match

A seed match is a Watson-Crick (WC) complimentary match between a miRNA and its target in the seed sequence. A WC match between a miRNA and mRNA nucleotide occurs when adenosine (A) pairs with uracil (U) and guanine (G) pairs with cytosine (C).

1.4 Conservation

Conservation refers to the maintenance of a sequence across species. The role of conservation in miRNA tar- get prediction is broad and analysis of conserved elements can be incorporated into miRNA target prediction in a variety of ways.

We have used the features of Duplex Stability and Site Accessibility to predict our gene targets.

Chapter 2

Free Energy

Free energy (or Gibbs free energy) can be used as a measure of the stability of a biological system. On the off chance that the binding of spacer and target mRNA is anticipated to be stable, it is viewed as a valid spacer-target arrangement.

2.1 Basic secondary structure prediction algorithms

RNA secondary structure is denoted by a graph, with it's edges and vertices corresponding to base pairs and nucleotides respectively. This graph follows the rules

1. only complementary nucleotides(A and T, G and C), including GU base pairs can for edges
2. A vertex can not have more than two edges
3. edges span at least three unpaired bases
4. if the vertices are placed on the circumference of a circle in a circular order from 5' to 3', two edges can not cross each other

The last condition guarantees that the graph avoids pseudo-knots. "Turing Model" of energy calculation states that the thermodynamic energy of RNA is a sum of various energy contributions initiating from the stacking of base pairs, the entropic strain of loops, namely hairpin loop, bulge loop,interior loop and multi loop, as well as partial stacking of dangling ends. These have been calculated in detail based experimental evidence.

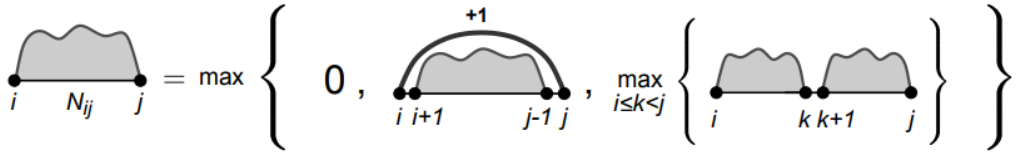

Figure 2.1: Graphical depiction of the Nussinov-like recursion

$$f(x) = max \begin{cases} 0, & : \text{if } (i + s_l) \geq j \\ \mathrm{N}_{i+1,j-1} + 1, & : \text{if } R_i, R_j \text{ can form base pair} \\ max_{i \leq k \leq j} \{ \ N_{i,k} + N_{k+1,j} \ \} & : \text{decomposition.} \end{cases} \quad (2.1)$$

The dynamic programming algorithmic paradigm for RNA secondary structure prediction follows the assumption that the structure can be broken down into smaller components, each with it's own energy contribution. Fig 2.1 outlines this scheme in a graphical manner.

2.2 RNA-RNA interactions

There are essentially two distinct methodologies for deciding the interaction between two RNAs that considers both the sequence and structure of the RNAs. The first methodology predicts the secondary structure of the duplex, made from the two participating RNAs. This is an NP-complete problem. Therefore, existing approaches implement a partial structure model. The simplest model is implemented in RNAcofold.

2.3 ViennaRNA[2]

The ViennaRNA Package is a set of standalone programs and libraries used for prediction and analysis of RNA secondary structures. It includes command-line tools like RNAfold, RNAplfold, RNAalifold etc.

2.4 RNAcofold

We use RNAcofold, a tool from the Vienna RNA package. RNAcofold takes in the sequences of spacer and the gene, and predicts the secondary structure of the most stable duplex, along with it's minimum free energy (MFE).

2.4.1 RNAcofold Input

RNAcofold can be used from the vienna rna web-server or, an offline command line tool can be used as well. We use the offline version to calculate MFE. Input to the tool is a .fasta file containing a single RNA sequence. This RNA sequence consists of the sequence obtained by transcription of gene and spacer. Finally, the two sequences are concatenated by a '&'. This concatenation-based approach makes extensive use of the predefined algorithms for single RNA secondary structure prediction.

2.4.2 RNAcofold Output

The output of the tool is a text file containing the predicted secondary structure in a dot bracket notation along with it's Δ G value. The output file is parsed using regex '[-+]?[0-9]*?[0-9]+$' to get the Δ G.

2.5 Normalisation

ΔG value is proportional to the gene's length. This might result into random spacer-target pairs showing high free energy. Therefore we normalise ΔG value with respect to the spacer and gene length using Eq 2.2 from [3]

$$\Delta G = \frac{\Delta G_0}{log(mn)} \tag{2.2}$$

2.6 Threshold Selection

To shortlist genes from a list of genes based on MFE, a threshold value is decided upon by looking at the lab validated spacer-mRNA data. In our case, there is no such pre-existing data, hence, calculating a threshold value is not possible. We try to overcome this shortcoming by including another feature of 'Site-Accessibility'.

Chapter 3

Site-Accessibility

The idea of site-accessibility approach is to investigate the ensemble properties of the individual sequences that are important for a RNA-RNA interaction. Basically, the target site must be accessible for binding to the interaction partner[4]. Therefore, for any two positions i ¡ k in a sequence, we compute the free energy that is required to make the sequence stretch between i and k single stranded, so that it is accessible.

3.1 Calculating ED Accessibility Energy

Estimation of ED can be performed by calculating the probability function $Pr^{unpaired}(i..k)$ for the ensemble of structures that leave the interaction site ranging from i to k single-stranded[5]. We further calculate the energy $ED(i,k)$ that is required to make the interaction site accessible as

$$ED(i,k) = -RTln(Pr^{unpaired}(i..k)). \tag{3.1}$$

This term is generally positive, as it is the external energy that is required to make the site accessible. $Pr^{unpaired}(i..k)$ can be calculated using RNAplfold, a tool from the RNA Vienna package[2].

3.2 RNAplfold

RNAplfold computes local pair probabilities for base pairs.

The -u option makes it possible to compute the probability that a stretch of x consecutive nucleotides is unpaired, which is useful for predicting possible binding sites.

Algorithm 1 ED Estimation

1: **procedure** CALCULATE ED($spacer, gene$) ▷ Site-Accessibility of given gene
2: $i, k \leftarrow Align(spacer, ReverseComplement(gene))$
3: $u \leftarrow length(spacer)$
4: $i^{'} \leftarrow length(gene) - 1 - i$
5: $A \leftarrow RNAplfold(gene, u, i^{'}, length(gene))$
6: $Pr^{unpaired}(i..k) \leftarrow A[i^{'}, k - i]$
7: $ED \leftarrow -RTlog(Pr^{unpaired}(i..k))$
8: **return** ED ▷ The ED is accessibility

Algorithm for estimating ED can be understood from Algorithm 1.

We align the spacer and reverse complement of gene on line 2. This alignment is necessary to calculate i and k, the sequence identical to spacer. We then transform i to $i^{'}$ to find the base where spacer attaches to the target gene.

RNAplfold is then called with these parameters, which outputs the matrix A. A is a plain text matrix containing on each line a position i followed by the probability that i is unpaired, [i-1..i] is unpaired [i-2..i] is unpaired and so on to the probability that [i-x+1..i] is unpaired.

Finally ED is calculated using Eq 3.1

Chapter 4

Extended Hybridisation Energies

We now combine both energy contributions. The extended hybridization energy of a specific interaction of two target sites [a,b] and [c,d] is now defined by summing up the ED-values and the free energy. For calculating the ED-values, we must know the first and the last interacting base in both sequences.

A graphical description of the two features is provided

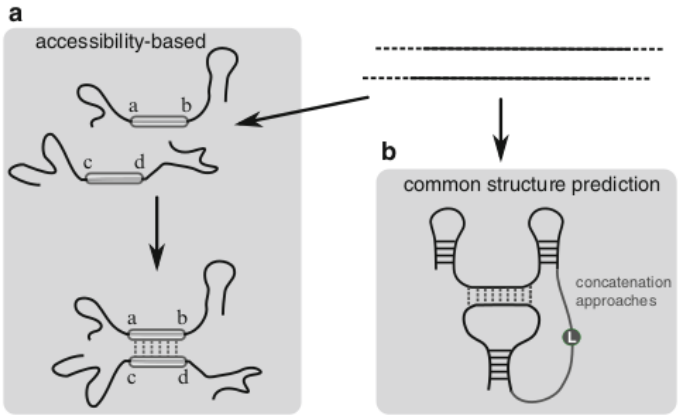

Figure 4.1: *a* depicts the site-accessibility and *b* depicts the Δ G. Extended Mean Free Energy is defined as the sum of *a* and *b*. It is to be noted that a tends to be positive as RNA rarely forms a single stranded structure in the cell's environment, while Δ G tends to be negative for putative spacer-target gene pair

Hence, the basic recursion for calculating the extended hybridization energy requires a 4D array $C(a, b, c, d)$. Note that the basic assumption in $C(a, b, c, d)$ is that both (a, b) and (c, d) form a base pair. C should be negative for a putative spacer-target gene pair

$$C(a, b, c, d) = \Delta G(a, b, c, d) + ED(a, b) \tag{4.1}$$

Chapter 5

Pathway Analysis

After shortlisting of genes on the basis of Extended Hybridisation Energies, we analyse the pathways these genes are involved in

5.1 Gene Set Enrichment Analysis

The general process of identifying a set of genes that are over represented in a given list of genes is known as Gene Set Enrichment Analysis. This process requires gene expression data (microarray data) and a control background expression data. Various tools compare the expression data and the background data to produce a set of genes that are over represented in expression data. This set of genes is further analysed to see the biological pathways that are regulated by these genes.

In our case, we already have a set of genes that we need to analyse pathways for.

5.2 KEGG

Kyoto Encyclopedia of Genes and Genomes is a database of hand drawn biological pathways. Kegg Mapper is a tool that takes in a list of genes Kegg ID and generates a list of pathways these genes affect. These pathways are then sorted in order of their gene size, to get the pathways which includes maximum amount of genes from our initial gene list.

Chapter 6

Results and Validation

The following table shows the number of shortlisted genes in CRISPR 1 and CRISPR 2

CRISPR	Total Genes	Shortlisted Genes
1	652	92
2	1120	122

6.1 Common Genes

We try to find the genes that are common to both CRISPR 1 and CRISPR 2 systems and look at the pathways that these genes effect. Following genes are common, and affect a valid pathway:

Genes	Summary
acca	acetyl-CoA carboxylase carboxyltransferase subunit alpha
aroE_1	quinate/shikimate dehydrogenase
metC	cystathionine beta-lyase

List of genes specific to CRISPR 1 and CRISPR 2 can be found in appendix

6.2 Pathways

The major pathways that are affected by these common genes are shown in the following table

Pathway	Genes Involved
Metabolic pathways	accA,aroE_1,metC
Biosynthesis of secondary metabolites	accA,aroE_1,metC
Biosynthesis of antibiotics	accA,aroE_1
Propanoate metabolism	accA
Phenylalanine, tyrosine and tryptophan biosynthesis	aroE_1
Microbial metabolism in diverse environments	acca
Selenocompound metabolism	metC
Biosynthesis of amino acids	metC
Pyruvate metabolism	acca
Carbon metabolism	accA
Fatty acid metabolism	accA
Cysteine and methionine metabolism	metC

6.3 Validation

For Validation, we need a database of spacer sequence v/s gene sequence. As of now, there is no standard database for the same. An alternate way is to use a tool that generates a guideRNA for a given gene sequence.

Crisflash[6] can rapidly design CRISPR guides against any sequenced genome or genome sequences, and can be used to create a database for validation.

Appendix

*Appendix*1

6.3.1 CRISPR 1

Following table describe gene shortlisted for CRISPR 1

Gene	Description
ycaJ	recombination factor protein RarA
ychA	putative transcriptional regulator
yibP	hypothetical protein
fucO	L-1,2-propanediol oxidoreductase
ydcW	gamma-aminobutyraldehyde dehydrogenase
infC	translation initiation factor IF-3
sdhB	succinate dehydrogenase iron-sulfur subunit
cysN	sulfate adenylyltransferase subunit 1
yiiQ	putative periplasmic protein
accA	acetyl-CoA carboxylase carboxyltransferase subunit alpha
yadR	iron-sulfur cluster insertion protein ErpA
yhjJ	putative Zn-dependent peptidase
hslO	Hsp33-like chaperonin
polB	DNA polymerase II
yggM	putative periplasmic protein
ydcN	putative repressor
mig-14	putative transcriptional activator
ygjT	putative tellurite resistance protein
yccF	hypothetical protein
stjC	putative periplasmic chaperone protein
yeaS	leucine export protein LeuE

Gene	Description
ygcB	putative helicase
hycA	formate hydrogenlyase regulatory protein HycA
smf	hypothetical protein
yabI	hypothetical protein
perM	putative permease
ilvG	acetolactate synthase 2 catalytic subunit
slmA	nucleoid occlusion protein
cdsA	CDP-diglyceride synthase
yegU	putative glycohydrolase
mrdA	penicillin-binding protein 2
rnt	ribonuclease T
envR	DNA-binding transcriptional regulator EnvR
flhB	flagellar biosynthesis protein FlhB
dcuA	anaerobic C4-dicarboxylate transporter
metC	cystathionine beta-lyase
sopA	secreted effector protein
iacP	acyl carrier protein
tldD	protease TldD
trpR	Trp operon repressor
aroE_1	quinate/shikimate dehydrogenase
dniR	membrane-bound lytic murein transglycosylase D
pykA	pyruvate kinase
galM	aldose 1-epimerase
fpr	ferredoxin-NADP reductase
yjjP	hypothetical protein
pduF	propanediol diffusion facilitator
sitA	putative periplasmic binding protein
ytfE	cell morphogenesis/cell wall metabolism regulator
hypD	putative hydrogenase formation protein
nupG	-1
dcp	dipeptidyl carboxypeptidase II
mviM	putative virulence protein
pagO	integral membrane protein
yfeJ	glutamine amidotransferase
csgF	curli assembly protein CsgF
yaeL	zinc metallopeptidase

Gene	Description
uraA	uracil transporter
dnaA	chromosomal replication initiation protein
pagO	integral membrane protein
yfeJ	glutamine amidotransferase
csgF	curli assembly protein CsgF
yaeL	zinc metallopeptidase
uraA	uracil transporter
dnaA	chromosomal replication initiation protein
crp	cAMP-regulatory protein
yaiY	putative inner membrane protein
ydhB	putative DNA-binding transcriptional regulator
tbpA	thiamine transporter substrate binding subunit
pflD	putative formate acetyltransferase 2
rpoH	RNA polymerase factor sigma-32
ydeA	sugar efflux transporter
imp	organic solvent tolerance protein
pitA	low-affinity phosphate transporter
rplC	50S ribosomal protein L3
yibL	hypothetical protein
rpsL	30S ribosomal protein S12
yqjI	putative transcriptional regulator
rfaY	lipopolysaccharide core biosynthesis protein
yabI	hypothetical protein
yajO	putative oxidoreductase
rumA	23S rRNA 5-methyluridine methyltransferase
yadQ	chloride channel protein
clpB	protein disaggregation chaperone
yghA	oxidoreductase
ynfB	hypothetical protein
sapD	peptide transport protein
flgJ	peptidoglycan hydrolase
atpA	F0F1 ATP synthase subunit alpha
yfhL	putative ferredoxin
yojI	multidrug transporter membrane component/ATP-binding component
rfaC	ADP-heptose:LPS heptosyl transferase I
sufS	selenocysteine lyase

Gene	Description
pgpB	phosphatidylglycerophosphatase B
entC	isochorismate synthase
ssrA	sensor kinase
melB	melibiose:sodium symporter
ulaA_1	ascorbate-specific PTS system enzyme IIC
yceH	hypothetical protein
aroE_1	quinate/shikimate dehydrogenase
ssaI	type III secretion system apparatus protein
ychH	hypothetical protein
yfhJ	hypothetical protein
yjjA	hypothetical protein
fhuA	ferrichrome outer membrane transporter
slrP	leucine-rich repeat-containing protein
ybhP	putative cytoplasmic protein
hpaD	4-hydroxyphenylacetate catabolism
ydeA	sugar efflux transporter
torC	trimethylamine N-oxide reductase cytochrome c-like subunit
allA	ureidoglycolate hydrolase
ybiR	putative transporter
copS	copper resistance protein
flgD	flagellar basal body rod modification protein
tppB	putative tripeptide transporter permease
phsA	thiosulfate reductase precursor
livM	leucine/isoleucine/valine transporter permease subunit
yiaN	hypothetical protein
gltS	glutamate transport protein
marT	putative transcriptional regulator
ubiC	chorismate pyruvate lyase
idnK	D-gluconate kinase

6.3.2 CRISPR 2

Following table describes genes shortlisted in CRISPR 2

Gene	Description
leuC	isopropylmalate isomerase large subunit

Gene	Description
yfjD	hypothetical protein
yjiN	putative inner membrane protein
accA	acetyl-CoA carboxylase carboxyltransferase subunit alpha
tqsA	putative transport protein
ycgM	hypothetical protein
cysM	cysteine synthase B
cadC	DNA-binding transcriptional activator CadC
rpsF	30S ribosomal protein S6
dpiA	two-component response regulator DpiA
hpaB	4-hydroxyphenylacetate catabolism
tppB	putative tripeptide transporter permease
rnb	exoribonuclease II
ompW	outer membrane protein W
treA	trehalase
yedJ	hypothetical protein
marT	putative transcriptional regulator
htrA	serine endoprotease
ybgH	POT family transport protein
ycfM	putative outer membrane lipoprotein
dcm	DNA cytosine methylase
gmd	GDP-D-mannose dehydratase
yehS	putative cytoplasmic protein
yfhD	putative transglycosylase
smpB	SsrA-binding protein
rumA	23S rRNA 5-methyluridine methyltransferase
hybD	hydrogenase 2 maturation endopeptidase
ygjO	putative methyltransferase
pstB	phosphate transporter subunit
adi	catabolic arginine decarboxylase
cbiC	precorrin-8X methylmutase
ybeR	putative cytoplasmic protein
ybgH	POT family transport protein
ygbI	putative regulatory protein
yghW	putative cytoplasmic protein
zntR	zinc-responsive transcriptional regulator
yiaG	putative transcriptional regulator

Gene	Description
sgaE	L-ribulose-5-phosphate 4-epimerase
yjcC	putative diguanylate cyclase/phosphodiesterase
cls	cardiolipin synthetase
aidB	isovaleryl CoA dehydrogenase
tgt	queuine tRNA-ribosyltransferase
metG	methionyl-tRNA synthetase
emrA	multidrug resistance secretion protein
srlR	DNA-binding transcriptional repressor SrlR
orgB	needle complex export protein
pgk	phosphoglycerate kinase
yccV	putative inner membrane protein
fabD	acyl carrier protein S-malonyltransferase
yrbC	putative transport protein
yifA	transcriptional regulator HdfR
flgJ	peptidoglycan hydrolase
sufC	cysteine desulfurase ATPase component
sbmC	DNA gyrase inhibitor
stcC	putative outer membrane protein
endA	DNA-specific endonuclease I
degS	serine endoprotease
nepI	ribonucleoside transporter
metJ	transcriptional repressor protein MetJ
yjiE	putative DNA-binding transcriptional regulator
nadR	nicotinamide-nucleotide adenylyltransferase
ybgJ	putative carboxylase
rpsC	30S ribosomal protein S3
yjfI	putative cytoplasmic protein
hcr	HCP oxidoreductase, NADH-dependent
ygaA	anaerobic nitric oxide reductase transcription regulator
yigN	DNA recombination protein RmuC
birA	biotin–protein ligase
pipA	pathogenicity island-encoded protein A
yeaR	putative cytoplasmic protein
metC	cystathionine beta-lyase
pepQ	proline dipeptidase
dnaT	primosomal protein I

Gene	Description
allB	allantoinase
gpmA	phosphoglyceromutase
pqiA	paraquat-inducible protein A
umuC	DNA polymerase V subunit UmuC
rfbC	dTDP-4,deoxyrhamnose 3,5 epimerase
dedD	hypothetical protein
nanT	putative sialic acid transporter
yhcO	putative cytoplasmic protein
aceB	malate synthase
hofC	type IV pilin biogenesis protein
glnL	nitrogen regulation protein NR(II)
yjjK	putative ABC transporter ATP-binding protein
pabC	4-amino-4-deoxychorismate lyase
ycaQ	putative cytoplasmic protein
yebC	hypothetical protein
dld	D-lactate dehydrogenase
ego	putative ABC-type aldose transport system ATPase component
copS	copper resistance protein
amn	AMP nucleosidase

References

1. Peterson SM, Thompson JA, Ufkin ML, Sathyanarayana P, Liaw L, Congdon CB. Common features of microRNA target prediction tools. Front Genet. 2014;5:23. Published 2014 Feb 18. doi:10.3389/fgene.2014.00023

2. Lorenz, Ronny et al. "ViennaRNA Package 2.0" Algorithms for molecular biology : AMB vol. 6 26. 24 Nov. 2011, doi:10.1186/1748-7188-6-26

3. Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R. Fast and effective prediction of microRNA/target duplexes. RNA. 2004;10(10):1507-17

4. Anke Busch, Andreas S. Richter, Rolf Backofen; IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions, Bioinformatics, Volume 24, Issue 24, 15 December 2008, Pages 2849–2856

5. Wright PR, Mann M, Backofen R. Structure and Interaction Prediction in Prokaryotic RNA Biology. Microbiol Spectr. 2018;6(2):10.1128/microbiolspec.RWR-0001-2017

6. Adrien L S Jacquin, Duncan T Odom, Margus Lukk, Crisflash: open-source software to generate CRISPR guide RNAs against genomes annotated with individual variation, Bioinformatics, , btz019, https://doi.org/10.1093/bioinformatics/btz019