G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2018;16(4):e26
https://doi.org/10.5808/GI.2018.16.4.e26

ORIGINAL ARTICLE

Functional Prediction of Hypothetical Proteins from *Shigella flexneri* and Validation of the Predicted Models by Using ROC Curve Analysis

Md. Amran Gazi[1]*, Sultan Mahmud[2], Shah Mohammad Fahim[1], Mohammad Golam Kibria[2], Parag Palit[1], Md. Rezaul Islam[3], Humaira Rashid[2], Subhasish Das[1], Mustafa Mahfuz[1], Tahmeed Ahmeed[1]

[1]Nutrition and Clinical Services Division, International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b), Dhaka 1212, Bangladesh, [2]Infectious Diseases Division, International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b), Dhaka 1212, Bangladesh, [3]International Max Planck Research School, Grisebachstraße 5, 37077 Göttingen, Germany

Shigella spp. constitutes some of the key pathogens responsible for the global burden of diarrhoeal disease. With over 164 million reported cases per annum, shigellosis accounts for 1.1 million deaths each year. Majority of these cases occur among the children of the developing nations and the emergence of multi-drug resistance *Shigella* strains in clinical isolates demands the development of better/new drugs against this pathogen. The genome of *Shigella flexneri* was extensively analyzed and found 4,362 proteins among which the functions of 674 proteins, termed as hypothetical proteins (HPs) had not been previously elucidated. Amino acid sequences of all these 674 HPs were studied and the functions of a total of 39 HPs have been assigned with high level of confidence. Here we have utilized a combination of the latest versions of databases to assign the precise function of HPs for which no experimental information is available. These HPs were found to belong to various classes of proteins such as enzymes, binding proteins, signal transducers, lipoprotein, transporters, virulence and other proteins. Evaluation of the performance of the various computational tools conducted using receiver operating characteristic curve analysis and a resoundingly high average accuracy of 93.6% were obtained. Our comprehensive analysis will help to gain greater understanding for the development of many novel potential therapeutic interventions to defeat *Shigella* infection.

Keywords: hypothetical protein, *in silico*, NCBI, ROC curve, *Shigella*

Introduction

Shigella, refers to a genus of gram-negative facultative anaerobes that belongs to members of the family *Enterobacteriaceae* and is the causative agent of shigellosis, a severe enteric infection, one of the most common causes of morbidity and mortality among children in developing nations. The Global Burden of Disease (GBD) classified *Shigella* as the second leading cause of diarrheal deaths on a global scale in 2015 [1]. Shigellosis leads to the recurrent passing of small, bloody mucoidal stools with synchronous abdominal cramps and tenesmus caused by ulceration of the colonic epithelium [2]. In malnourished children, *Shigella* infection may lead to a vicious cycle of further impaired nutrition, frequent infection and growth retardation resulting from protein loss enteropathy [3].

The *Shigella* genus is divided into four species: *Shigella flexneri*, *Shigella boydii*, *Shigella sonnei*, and *Shigella dysenteriae*. These are further classified into serotypes based on biochemical differences and variations in their O-antigen [4]. A total of 19 different serotypes of *S. flexneri* have been reported so far by various research groups [5]. Among the four *Shigella* species, shigellosis is predominantly caused by *S. flexneri* in the developing world especially in Asia, and is responsible for approximately 10% of all diarrheal episodes among children of < 5 years [6]. Recent multicenter study in

Received July 2, 2018; Revised September 16, 2018; Accepted September 16, 2018; Published online December 28, 2018

***Corresponding author:** Tel: +880-1680731163, Fax: +880-29827075, E-mail: amran.gazi@icddrb.org

Asia revealed that the incidence of this disease might even exceed previous estimations, due to *Shigella* DNA being detectable in up to one third of the total culture negative specimens [7]. Currently, no effective vaccine with the ability to confer adequate protection against the many different serotypes of *Shigella* has been developed and made available. Existing antimicrobial treatments are becoming compromised in terms of efficacy due to increased antibiotic resistance, soaring cost of treatment, and persistence of poor hygiene and unsanitary conditions in the developing world.

A particular study conducted on numerous isolates of *Shigella* collected over a time span of 10 years, multi-drug resistance (MDR) were found to be exhibited by 78.5% of the isolates. 2% of the isolates were found to harbor genetic information capable of conferring resistance to azithromycin, a final resort antimicrobial agent for shigellosis [8]. On the other hand, a recent whole genome analysis of a particular strain of *S. flexneri* revealed 82 distinct chromosomal antibiotic resistance genes while successive re-sequencing platforms elucidated several distinct single nucleotide polymorphisms that contributed to eventual MDR [9]. Therefore, the development of new drugs has risen to become a subject of immense magnitude to not only shorten the medication period but also to treat MDR shigellosis. The genome sequence of *S. flexneri* serotype 2a strain 2457T, available in the NCBI database consists of 4,599,354 bp in a single circular chromosome containing 4,906 genes encoding 4,362 proteins and has G + C content of 50.9% [10]. Among these, the functions of 674 proteins have not been experimentally determined till date and are termed as hypothetical proteins (HPs). A HP is one that has been predicted to be encoded by an identified open reading frame, but for which there is a lack of experimental evidence [11]. Nearly half of the proteins in most genomes belong to the class of HPs and this class of proteins presumably have their own importance to complete genomic and proteomic platform of an organism [12, 13]. Precise annotation of the HPs of particular genome leads to the discovery of new structures as well as new functions, and elucidating a list of additional protein pathways and cascades, thus completing our incomplete understanding on the mosaic of proteins [13]. HPs may possibly play crucial roles in disease progression and survival of pathogen [11, 14]. Furthermore, novel HPs may also serve as markers and pharmacological targets for development of new drugs and therapies [15]. Functions of HPs from several pathogenic organisms have been already reported using a plethora of sequence and structure based methods [14, 16, 17].

Functional annotation of HPs utilizing advanced bioinformatics tools is a well-established platform in current proteomics [18]. Cost and time efficiency of these methods also favoring their preference over contemporary in vitro techniques [19]. In this study, we have used several well optimized and up to date bioinformatics tools to assign functions of a number of HPs from the genome of *S. flexneri* with high precision [20]. Functional domains were considered as the basis to infer the biological functions of HPs in this case. The receiver operating characteristic (ROC) analysis [21] was used for evaluating the performance of bioinformatics tools executed in our study. We also measured the confidence level of the functional predictions on the basis of bioinformatics tools employed during the course of the investigation [22]. We believe that this analysis will expand our knowledge regarding the functional roles of HPs of *Shigella* and provide an opportunity to unveil a number of potential novel drug targets [17].

Methods

The computational algorithm used for this study has been illustrated in Fig. 1. The entire work scheme has been divided into three phases namely, phase I, II and III. Phase I involves the characterization and sequence retrieval of the HPs, following the analysis of the *S. flexneri* genome. Phase II comprises of the annotation of various functional parameters using well optimized series of tools. The probable functions of the characterized HPs were predicted by the integration of various functional predictions. In phase III, an approach was made for systematic performance evaluation of various bioinformatics tools used in this study. In this case, *S. flexneri* protein sequences with known function were used as control. Finally, expert knowledge was applied for annotation of HPs at a considerable degree of confidence.

Phase I

Accession of genome and sequence retrieval

Complete genome sequence of *S. flexneri* 2a str. 2457T was retrieved from NCBI database (http://www.ncbi.nlm.nih.gov/genome/) and was found to code for a total of 4,362 proteins (accessed July 5, 2017). Fasta sequences of the complete coding sequence of 682 proteins, characterized as HPs were retrieved from UniProt (http://www.uniprot.org/). Finally, a total of 674 proteins were retained for downstream analysis following exclusion of duplicates.

Analysis of the conserved domains

Domains are often identified as recurring (sequence or structure) units, and can be thought of as distinct functional and/or structural units of a protein. During molecular evolution, it is assumed that domains may have been utilized as building blocks and have encountered recombination to modulate protein function [23]. A domain or fold might also

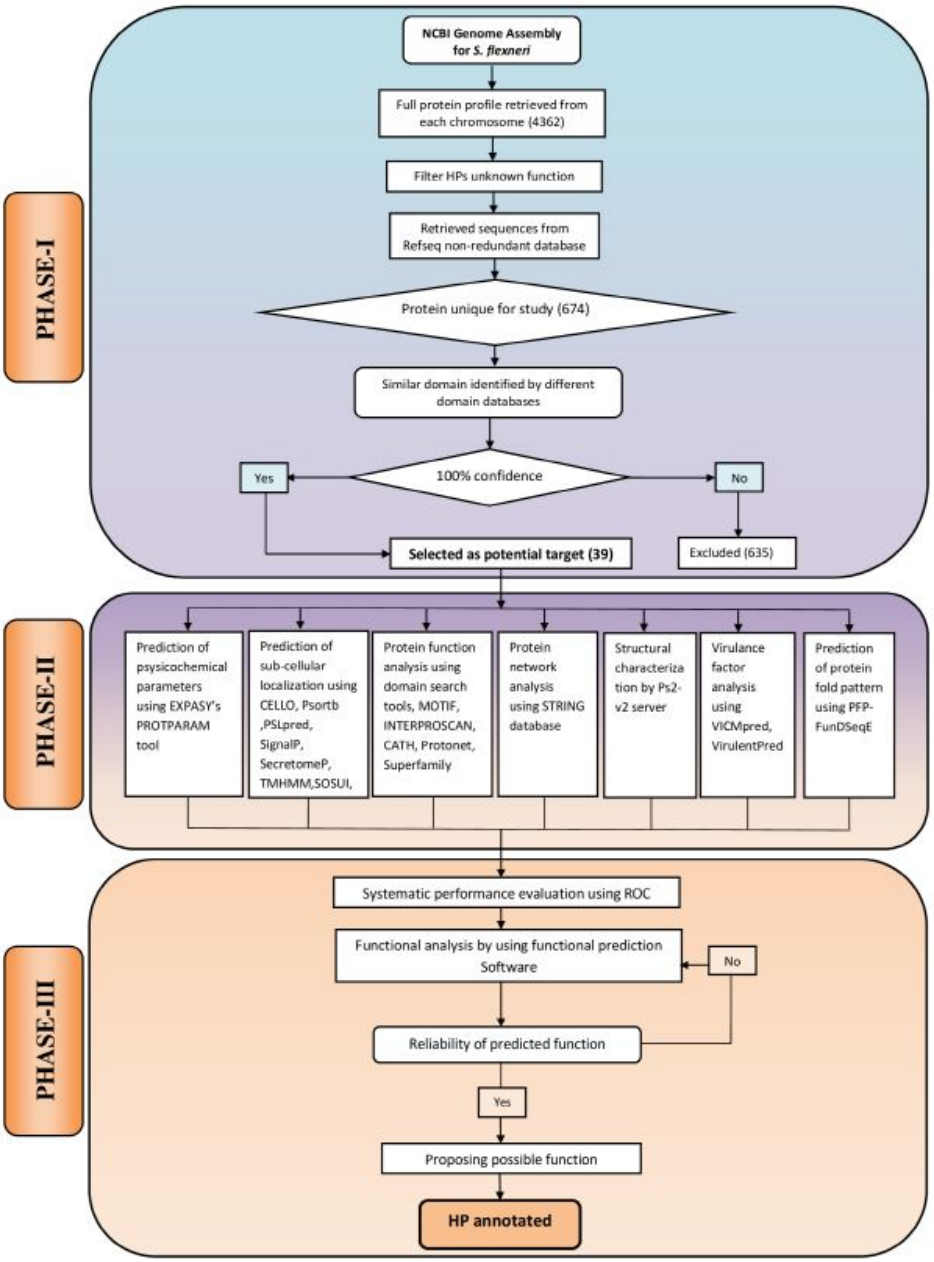

Fig. 1. Computational algorithm used for annotating function of 39 hypothetical proteins (HPs) from *Shigella flexneri*. The framework has been divided into three phases: PHASE I, sequence retrieval from online databases; PHASE II, the extensive analysis of sub-cellular localization, physicochemical parameters, virulence, function and domain present in HPs; PHASE III, assessment of the predicted functions using the protein with known function from *S. flexneri* and reliable prediction of possible functions of HPs.

exhibit a higher degree of conservancy when compared with the entire sequence [24].

In our study, five bioinformatics tools namely: CDD-BLAST (Conserved Domain Database-Basic Local Alignment Search Tool) [25-27], PFAM [28], HmmScan [29], SMART (Simple Modular Architecture Research Tool) [30], and SCANPROSITE [31] were used. These tools are able to search for the defined conserved domains in the targeted protein sequences and further assist in the classification of putative proteins in a particular protein family. HPs analyzed by five aforementioned function prediction web tools revealed the variable results when searched for the conserved domains in hypothetical sequences. Therefore, different confidence levels were assigned on the basis of collective results of these web-tools. One hundred percentage confidence level was considered upon obtaining the same results from the five distinct tools. Finally, we obtained 39 such proteins from 674 primary collected proteins, which were taken for further analysis (Supplementary Table 1).

Phase II

Physicochemical characterization

Theoretical physiochemical parameters such as molecular weight, isoelectric point, aliphatic index, instability index and grand average of hydropathicity (GRAVY) of these HPs were analyzed using ProtParam server of the Expasy tools

(http://web.expasy.org/protparam/). Results of this analysis have been listed in Supplementary Table 2.

Determination of sub-cellular localization

For the identification of a protein as a drug or vaccine candidate, determination of the sub-cellular localization of the protein becomes particularly important. Surface membrane protein can be served as a potential vaccine target while cytoplasmic proteins may act as promising drug targets [32]. We used CELLO [33], PSORTb [34], and PSLpred [35] for the denotation of sub-cellular localization of the query proteins. TMHMM, SOSUI, and HMMTOP were applied for the prediction of query proteins for being a membrane protein, based on Hidden Markov Model [36-38]. SingnalP 4.1 [39] was used to predict the signal peptide and SecretomeP 2.0 [40] were utilized for the identification of proteins involved in non-classical secretory pathway. Results of these predictions are summarized in Supplementary Table 3.

Functional prediction of the query proteins

Various tools were used for precise functional assignments of all 39 HPs from *S. flexneri* (described in Table 1) such as CDD-Blast, Pfam, HmmScan, SMART, Scanprosite, MOTIF [41], INTERPROSCAN [42], CATH [43], SUPERFAMILY

Table 1. List of annotated Hconf proteins from *Shigella flexneri*

No	Protein name	Protein function
1	WP_005053355.1	Peptidase, C92 family
2	WP_000092054.1	DUF1615/lipoprotein
3	WP_001382892.1	DUF3251/lipoprotein YajI/immunoglobulin like domain
4	WP_005053036.1	Lipoprotein_16/uncharacterized lipoprotein
5	WP_000779831.1	lipoprotein chaperone (YscW)
6	WP_011110552.1	YbfN-like lipoprotein
7	WP_001269672.1	LPS-assembly lipoprotein RlpB (LptE)
8	WP_001247854.1	Topoisomerases, DnaG-type primases, Hedgehog/intein domain
9	WP_000070107.1	ATP-binding cassette transporter
10	WP_000224274.1	MOSC beta barrel domain/2Fe-2S iron-sulfur cluster binding domain
11	WP_000749269.1	YceI-like domain
12	WP_001125713.1	YcgL domain
13	WP_001043881.1	GAF domain
14	WP_001295493.1	Endoribonuclease L-PSP/YjgFfamily
15	WP_000691930.1	Domain of unknown function (DUF333)
16	WP_000597196.1	Glycine zipper 2TM domain
17	WP_000248636.1	AI-2E family transporter/permease
18	WP_000755956.1	SPFH domain/Band 7 family
19	WP_001237866.1	YecR-like lipoprotein
20	WP_000454701.1	TerC family/Transporter associated domain/CBS domain
21	WP_000003197.1	von Willebrand factor type A domain
22	WP_005049020.1	Uncharacterized lipoprotein YehR
23	WP_048814497.1	Leucine rich repeat protein/NEL or novel E3 ligase domain
24	WP_000301054.1	Lipopolysaccharide kinase (Kdo/WaaP)
25	WP_000266171.1	Tetratricopeptide repeat (TPR)
26	WP_000589825.1	Outer membrane protein (ompA) like domain/membrane lipoprotein
27	WP_005051685.1	LysM (lysin-like motif)/peptidase family M23
28	WP_001387238.1	DNA repair protein RadC-like JAB domain
29	WP_000248097.1	Carrier protein (CP) domain and phosphopantetheine attachment site
30	WP_000848528.1	Lipoprotein leucine-zipper
31	WP_000189314.1	GIY-YIG nuclease domain
32	WP_001297375.1	DNA repair protein RadC-like JAB domain
33	WP_000858193.1	yiaA/B two helix domain
34	WP_001296791.1	Autotransporter beta-domain
35	WP_000778795.1	Acetyltransferase (GNAT) domain
36	WP_001205243.1	Xylose isomerase-like TIM barrel (AP_endonuc_2)
37	WP_001238362.1	Lipocalin-like domain
38	WP_000943980.1	Glutathionylspermidine synthase
39	WP_000132640.1	Toxin SymE/SpoVT-AbrB domain

Table 2. Different types of folds identified in *Shigella flexneri*

No.	Fold type	Accession number
1	Viral coat and capsid proteins	WP_005053355.1, WP_000691930.1
2	TIM-barrel	WP_000092054.1, WP_001247854.1, WP_001295493.1, WP_000266171.1, WP_001297375.1, WP_000943980.1, WP_000132640.1
3	Ferredoxin-like	WP_001382892.1, WP_000003197.1, WP_000301054.1
4	4-Helical up-and-down bundle	WP_005053036.1, WP_000779831.1, WP_001269672.1
5	DNA-binding 3-helical bundle	WP_011110552.1, WP_000589825.1, WP_001387238.1, WP_000848528.1, WP_000189314.1
6	Small inhibitors, toxins, lectins	WP_000070107.1, WP_000454701.1, WP_048814497.1, WP_001205243.1
7	Belta-grasp	WP_000224274.1
8	Cupredoxins	WP_000749269.1, WP_000248636.1, WP_005051685.1
9	Thioredoxin-like	WP_001125713.1
10	Flavodoxin-like	WP_001043881.1
11	Trypsin-like serine proteases	WP_000597196.1
12	OB-fold	WP_000755956.1, WP_001237866.1, WP_000858193.1, WP_000778795.1
13	Immunoglobulin-like	WP_005049020.1
14	EF-hand	WP_000248097.1
15	ConA-like lectin/glucanases	WP_001296791.1
16	Lipocalins	WP_001238362.1

[44], and Protonet [45]. Results of these analyses have been outlined in Supplementary Tables 4 and 5.

The computational prediction of the structure of a protein from its amino acid sequences greatly facilitates the subsequent prediction of its function [46]. An online server PS2-v2 (PS Square version 2) [47], a template based method were used to predict the structure of the HPs. The modeling of proteins using this online server further substantiated the function of HPs. Besides, PFP-FunDSeqE [48] has been used to elucidate the protein fold patterns based on a combination of functional domain information and evolutionary information (Table 2).

Virulence factors analysis

Virulence factors (VFs) are described as potent targets for developing drugs because it is essential for the severity of infection [49]. VICMpred [50] and Virulentpred [51] tools were employed to predict VFs from protein sequences with an accuracy of 70.75% and 81.8%, respectively.

Functional protein association networks

The function and activity of a protein are often modulated by other proteins with which it interacts. Therefore, understanding of protein-protein interactions serve as valuable leads for predicting the function of a protein. In this investigation, we had employed STRING (Search Tool for the Retrieval of Interacting Genes/Proteins, https://string-db.org/) [52] to predict protein interactions partners of HPs. To predict functional association, only highest confidence score partner proteins were chosen in this study.

Phase III

Performance assessment

The predicted functions of HPs from *S. flexneri* and the accuracy of associated tools were validated using the ROC curve analysis. In this analysis, the diagnostics efficacy is evaluated at six levels where 1 and 0 classified as true positive and true negative respectively as binary numerals. In addition, the integers (2, 3, 4, and 5) were used as confidence ratings for each case. The ROC curves were carried out using 25 *S. flexneri* proteins with known function as control and were compared with the results obtained for the 39 HPs (Supplementary Tables 6 and 7). The results were submitted to web-based calculator for the ROC curves [53] in "format 1" form and the program thereby calculated the ROC curves. The results were expressed in terms of accuracy (Ac), sensitivity (Se), specificity (Sp) and the area under the curve (AUC) [54]. The average accuracy of the employed pipeline was found 93.6% (Table 3, Fig. 2).

Results and Discussion

Sequence analysis

Sequences of all the 674 HPs were analyzed for identification of the functional domains using five bioinformatics tools namely CDD-BLAST, Pfam, HmmScan, SMART, and SCANPROSITE. If the given five tools indicated the same domains for a protein, we considered it as 100% confidence level. In our study, all the five tools mentioned above revealed 39 such proteins and hence were grouped together.

Table 3. List of accuracy, sensitivity, specificity, and ROC area of various bioinformatics tools used for predicting function of Hconf proteins from *Shigella flexneri* obtained after ROC analysis

No.	Software name	Accuracy of prediction (%)	Sensitivity (%)	Specificity (%)	ROC area
1	BLAST	100	100	n/a	n/a
2	Pfam	100	100	100	1
3	HmmScan	100	100	100	1
4	SMART	100	100	100	1
5	Scanprosite	72	100	12.50	0.662
6	MOTIF	100	100	100	1
7	Interproscan	100	100	100	1
8	CATH	80	100	16.70	0.539
9	SUPERFAMILY	84	100	20	0.54
10	ProtoNet	100	100	100	1
11	Average	93.6	100	64.35	0.774

ROC, receiver operating characteristic.

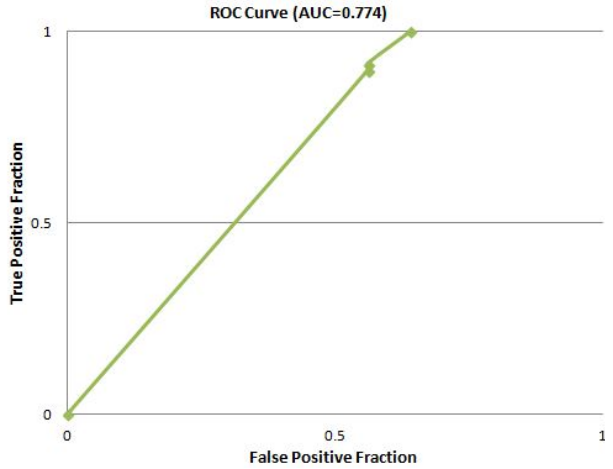

Fig. 2. Receiver operating characteristic (ROC) curve (area under the curve [AUC] = 0.774) for average accuracy of prediction.

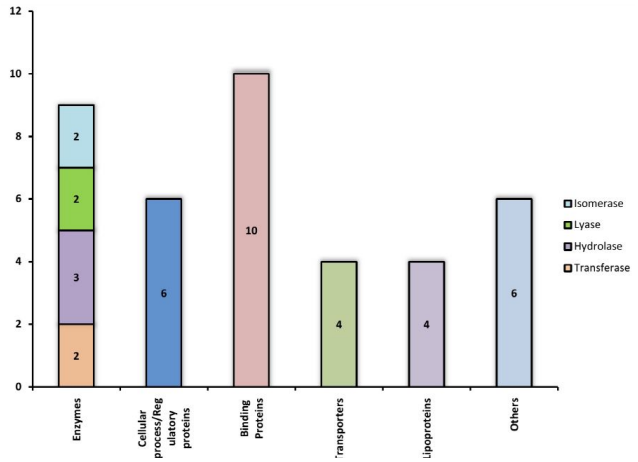

Fig. 3. Hypothetical proteins classified into different groups based on their functions.

Only these HPs having 100% confidence level were considered for further analyses and termed as highly confident (Hconf) proteins. From the rest of the 635 proteins, no specific conserved domains were found for a total of 257 proteins. For other HPs (n = 378), specific domains were identified using several of these tools. To know accurate function of these proteins further studies are required.

The function of each of these 39 Hconf were successfully assigned by using different online tools, listed in Table 1. All sequence analyses were compiled and categorized into various functional classes constituting 9 enzymes, 10 binding proteins, 4 transporters, 4 lipoproteins, 6 which are involved in various cellular processes, while 6 proteins were predicted to exhibit miscellaneous functions (Fig. 3). Various functional classes of these classified Hconf proteins are described below.

Enzymes

Enzymes are key players in many leading biochemical processes in the living system and may facilitate the survival of pathogens in the host and making it viable for the course of infection. A total of 9 proteins out of 39 (23%) of our annotated Hconfs were characterized as enzymes. Among these, two proteins were characterized as transferases, among which, WP_000301054.1 is a lipopolysaccharide kinase (Kdo/WaaP), involved in the formation of outer membrane (OM) of gram negative bacteria and is encoded by the Waap gene. The OM protects cells from toxic molecules and is important for survival during infection and is required for virulence of the pathogen [55]. According to reports made by Delucia [55], the depletion of WaaP gene was seen to halt the growth of the bacteria suggesting that WaaP is essential to produce the full-length lipopolysaccharide,

recognized by the OM [49]. Therefore, WaaP may result in a potent target for the development of novel antimicrobial agents. The other transferase, protein WP_000778795.1 was found to consist of an acetyltransferase (GNAT) domain that uses acetyl coenzyme A (CoA) to transfer an acetyl group to a substrate, a reaction implicated in various functions for the development of antibiotic resistance of bacteria [56].

Three enzymes were predicted to be hydrolases, which plays key role in the invasion of the host tissue and evading the host defense mechanism and are thus associated with various VFs [57]. For instance, WP_005051685.1 marks the lysin-like motif/peptidase family M23, is found in proteins from viruses, bacteria, fungi, plants and mammals. It is present in bacterial extracellular proteins including hydrolases, adhesins and VFs such as protein A from *Staphylococcus aureus*. We report WP_001295493.1 protein as the endoribonuclease/YjgF family active on single-stranded mRNA that inhibits protein synthesis by cleaving mRNA [58]. YjgF family members are enamine/imine deaminases that hydrolyze reactive intermediates released by pyridoxal phosphate-dependent enzymes, including threonine dehydratase [59]. It has also been reported in the inhibition of transaminase B in *Salmonella* [60].

Among the other enzymes predicted, there has been two isomerase and one lyase enzyme. WP_001247854.1 constitutes the toprim (topoisomerase-primase),a catalytic domain involved in breakage and rejoining of DNA strand [61]. WP_001205243.1 marks the Xylose isomerase-like TIM barrel involved in the myo-inositol catabolism pathway [62]. Lyases also play a key role in bacterial pathogenesis due to their involvements in various biosynthesis processes. WP_000943980.1 was found to demonstrate synthase activity that causes hydrolysis of ATP with the formation of an amide bond between spermidine and the glycine carboxylate of glutathione. In the pathogenic trypanosomatids, this reaction is the penultimate step in the biosynthesis of the antioxidant metabolite, and is a resounding target for target mediated drug design [63]. The WP_000454701.1 protein was found to be a cystathionine b-lyase, an enzyme which forms the cystathionine intermediate in cysteine biosynthesis and may be considered as the target for pyridiamine anti-microbial agents [64].

Binding proteins

Ten proteins annotated as binding proteins among which 1 RNA binding, 3 protein binding, 3 lipid binding, 1 metal binding, 1 peptidoglycan binding, and 1 adhesion protein have been predicted. WP_000132640.1 protein was predicted to be SymE (SOS-induced yjiW gene with similarity to MazE). It has been reported to involve in inhibiting cell growth, decrease protein synthesis and increase RNA degradation and thus exhibit a vital role in the survival and propagation of pathogen in the host [65, 66]. Despite not manifesting any functional homology with other type I toxin proteins, SymE belongs to the type I toxin-antitoxin system. Its function resembles that of type II toxins such as MazF, which is able to perform the cleavage of mRNA in a ribosome independent manner. However, SymE shares homology to the AbrB-fold superfamily proteins such as MazE, which acts as transcriptional factors and antitoxins in various type II TA modules [67]. It seems probable that SymE has evolved into an RNA cleavage protein with toxin-like properties from a transcription factor or antitoxin [66]. In our study, we reported WP_000003197.1 as von Willebrand factor with a type A domain which has been reported responsible for various blood disorders [68-70]. The association of type A domain makes it liable to be involved in various significant activities such as cell adhesion and immune defense [71]. On the other hand, WP_000755956.1 has been predicted to belong to the band-7 protein family that comprises of a diverse set of membrane-bound proteins characterized by the presence of a conserved domain [72]. The exact function of this domain is not known, but concurrent reports from animal and bacterial stomatin-type proteins demonstrate the ability of binding to lipids and in the assembly of membrane-bound oligomers that form putative scaffolds [73]. We have also predicted WP_001269672.1 and WP_000749269.1 as the lipid binding domain called lipopolysaccharide (LPS)-assembly lipoprotein LptE and the YceI-like domain respectively. The LPS transport machinery is composed of LptA, LptB, LptC, LptD, and LptE. LptE forms a complex with LptD, which is involved in the assembly of LPS in the outer leaflet of the OM [74]. This OM is an effective permeability barrier that protects the cells from toxic compounds, such as antibiotics and detergents, thus conferring the bacteria with the capability to adapt and consequently inhabit several different and often hostile environments. Among the binding prtoteins, WP_000266171.1 was found to be a tetratricopeptide repeat containing protein which is involved in protein-protein interactions and thus plays an important role in virulence [75].

Cellular processes/regulatory proteins

A total of 6 HPs have been predicted to be involved in various cellular and regulatory mechanisms, which are vital cognates in the pathogenesis of *S. flexneri* and thus can be treated as possible drug targets [76]. For example, WP_000189314.1 predicted to be a member of the GIY-YIG family involved in many cellular processes including DNA repair and recombination, transfer of mobile genetic elements, and restriction of incoming foreign DNA [77, 78].

WP_001387238.1 and WP_001297375.1 have been found to be RadC-like domain belonging to the JAB superfamily of metalloproteins [79]. In most instances, this domain shows fusions to an N-terminal Helix-hairpin-Helix (HhH) domain and may also be function as a nuclease [79]. WP_000848528.1 has been predicted to be a leucine-zipper found in the enterobacterial OM lipoprotein LPP [80]. It is likely that this domain is involved in protein-protein interaction via subsequent oligomerization. WP_000597196.1 and WP_048814497.1 have been respectively found to be a Glycine zipper 2TM domain found in the *Rickettsia* genus and leucine-rich repeat involved in a variety of biological processes, including signal transduction, cell adhesion, DNA repair, recombination, transcription, RNA processing, disease resistance, apoptosis, and the immune response [81].

Lipoprotein

Bacterial lipoproteins are a set of membrane proteins with many different functions. Due to this broad-ranging functionality, these proteins have a considerable significance in many phenomena, from cellular physiology through cell division and virulence [82]. Lipoprotein of gram-negative bacteria is essential for growth and division [83]. In our analysis, we report a total of 4 lipoproteins from the group of HPs predicted in this study. It has been also revealed that lipoproteins may function as vaccines [82]. The knowledge of these facts may be utilized for the generation of novel countermeasures to bacterial diseases [82].

Transport

In our findings, we report the prediction of HP WP_000070107.1 to be a member of the ATP-binding cassette superfamily,largest of all protein families with a diversity of physiological functions [84]. It has recently been identified that these proteins may be involved in virulence and are essential for intracellular survival of pathogens [85]. We have found protein WP_001296791.1 to be an auto-transporter of the YhjY type involved in DNA repair [86]. Protein WP_001238362.1 has been found to exhibit the function of transport of nutrients, control of cell regulation, pheromone transport, cryptic coloration and in the enzymatic synthesis of prostaglandins. An example a protein with such function is the retinol-binding protein 4, which transfers retinol from liver to peripheral tissues [87].

Other proteins

Six HPs have been predicted to exhibit miscellaneous functions where most of them are protein with unknown function. Among them, WP_001382892.1 and WP_000691930.1 have been preicted to be domains of unknown function and are found in a number of bacterial proteins. WP_001125713.1 has been found to be YcgL domain with conserved class of small proteins widespread in gammaproteo bacteria. This group of proteins contain a 85-residue domain of unknown function and two alpha-helices and four beta- strands in the sequential arrangement [88]. We have also predicted WP_001237866.1 and WP_005049020.1 as YecR and YehR like family of lipoproteins found in bacteria and viruses and are functionally uncharacterized.

Virulent proteins

Gram-negative bacteria undergo frequent genomic alterations and consequent evolutions thus increasing their virulence inside the host environment [89]. We have found 2 HPs that showed positive virulence scores servers among the Hconf proteins. These have been listed in Supplementary Table 8. It had already been hypothesized that targeting VF provides a better therapeutic intervention strategy against bacterial pathogenesis [89]. Predicted HPs having virulent characteristics thus provide powerful target-based therapies for the mitigation of an existing infection and are further considered as an adjunct therapy to existing antibiotics, or potentiators of the host immune response [90].

ROC curve

The average accuracy of the employed pipeline was identified 93.6% in our analysis which indicated that outcomes of the functional annotation of HPs were predicted with a high degree of confidence. We have also found sensitivity of 100% and specificity 64.3% for the tools used in this study. Finally, area under the curve was found to be 0.774. AUC is an effective way to summarize the overall diagnostic accuracy of the test. It takes values from 0 to 1, where 0.7 to 0.8 is considered acceptable.

Conclusion

Using an innovative *in silico* approach, all 674 HPs from *S. flexneri* were primarily analyzed and then using the ROC analysis and confidence level measurements of the predicted results the functions of the 39 HPs were precisely predicted with a reasonably high degree of confidence and thereby were successfully characterized. Following this, the validation of the functions of these proteins were carried out by using different approaches including structure based PS2-v2 server, sub-cellular localization and physicochemical parameters. These are important for distinguishing the HPs from the rest of the protein. The protein-protein interaction also gave insights in elucidation of the involvement of such proteins in various metabolic pathways. Moreover, some virulence proteins had also been detected which are essential

for the survival of this pathogen. This *in silico* approach for functional annotation of the HPs can be further utilized in drug discovery for characterizing putative drug targets for other clinically important pathogens. The outcomes of ROC analysis indicated high reliability of bioinformatics tools used in this study. Hence, the functional annotation of HPs is reliable and can be further utilized for other experimental research.

ORCID: Md. Amran Gazi: https://orcid.org/0000-0002-3286-7536; Sultan Mahmud: https://orcid.org/0000-0002-0392-9646; Shah Mohammad Fahim: https://orcid.org/0000-0002-3627-202X; Mohammad Golam Kibria: https://orcid.org/0000-0002-7821-2455; Parag Palit: https://orcid.org/0000-0001-7863-2639; Humaira Rashid: https://orcid.org/0000-0001-8607-573X; Subhasish Das: https://orcid.org/0000-0002-7852-6569

Author's contributions

Conceptualization: MAG
Data curation: MAG, SM, SMF, MGK
Formal analysis: MAG, SM
Funding acquisition: MAG, MRI, HR, SD
Methodology: SM, MGK, PP, MRI, HR, SD
Writing – original draft: MAG, MM, TA
Writing – review & editing: MAG, MM, TA

Conflicts of Interest

No potential conflicts of interest relevant to this article was reported.

Acknowledgments

The authors are grateful to core donors which provide unrestricted support to icddr,b for its operations and research. Current donors providing unrestricted support include: Government of the People's Republic of Bangladesh; Canadian International Development Agency (CIDA), Swedish International Development Cooperation Agency (Sida), and the Department for International Development, UK (DFID). We gratefully acknowledge these donors for their support and commitment to icddr,b's research efforts.

Supplementary material

Supplementary data including eight tables can be found with this article online at https://doi.org/10.5808/GI.2018.16.4.e26.

References

1. GBD Diarrhoeal Diseases Collaborators. Estimates of global, regional, and national morbidity, mortality, and aetiologies of diarrhoeal diseases: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Infect Dis* 2017;17:909-948.
2. Mathan MM, Mathan VI. Ultrastructural pathology of the rectal mucosa in Shigella dysentery. *Am J Pathol* 1986;123:25-38.
3. Keusch GT. *Shigella* infections. *Clin Gastroenterol* 1979;8:645-662.
4. Taneja N, Mewara A. Shigellosis: epidemiology in India. *Indian J Med Res* 2016;143:565-576.
5. Parajuli P, Adamski M, Verma NK. Bacteriophages are the major drivers of *Shigella flexneri* serotype 1c genome plasticity: a complete genome analysis. *BMC Genomics* 2017;18:722.
6. Ferreccio C, Prado V, Ojeda A, Cayyazo M, Abrego P, Guers L, *et al*. Epidemiologic patterns of acute diarrhea and endemic *Shigella* infections in children in a poor periurban setting in Santiago, Chile. *Am J Epidemiol* 1991;134:614-627.
7. von Seidlein L, Kim DR, Ali M, Lee H, Wang X, Thiem VD, *et al*. A multicentre study of *Shigella* diarrhoea in six Asian countries: disease burden, clinical manifestations, and microbiology. *PLoS Med* 2006;3:e353.
8. Nuesch-Inderbinen M, Heini N, Zurfluh K, Althaus D, Hachler H, Stephan R. *Shigella* antimicrobial drug resistance mechanisms, 2004-2014. *Emerg Infect Dis* 2016;22:1083-1085.
9. Zhu Z, Zhou X, Li B, Wang S, Cheng F, Zhang J. Genomic analysis and resistance mechanisms in *Shigella flexneri* 2a strain 301. *Microb Drug Resist* 2018;24:323-336.
10. Wei J, Goldberg MB, Burland V, Venkatesan MM, Deng W, Fournier G, *et al*. Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. *Infect Immun* 2003;71:2775-2786.
11. Desler C, Suravajhala P, Sanderhoff M, Rasmussen M, Rasmussen LJ. *In silico* screening for functional candidates amongst hypothetical proteins. *BMC Bioinformatics* 2009;10:289.
12. Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, *et al*. Protein function annotation by homology-based inference. *Genome Biol* 2009;10:207.
13. Nimrod G, Schushan M, Steinberg DM, Ben-Tal N. Detection of functionally important regions in "hypothetical proteins" of known structure. *Structure* 2008;16:1755-1763.
14. Kumar K, Prakash A, Tasleem M, Islam A, Ahmad F, Hassan MI. Functional annotation of putative hypothetical proteins from *Candida dubliniensis*. *Gene* 2014;543:93-100.
15. Lubec G, Afjehi-Sadat L, Yang JW, John JP. Searching for hypothetical proteins: theory and practice based upon original data and literature. *Prog Neurobiol* 2005;77:90-127.
16. Shahbaaz M, Ahmad F, Imtaiyaz Hassan M. Structure-based functional annotation of putative conserved proteins having lyase activity from *Haemophilus influenzae*. *3 Biotech* 2015;5:317-336.
17. Sinha A, Ahmad F, Hassan MI. Structure based functional annotation of putative conserved proteins from *Treponema pallidum*: search for a potential drug target. *Lett Drug Des Discov*

2015;12:46-59.
18. Adams MA, Suits MD, Zheng J, Jia Z. Piecing together the structure-function puzzle: experiences in structure-based functional annotation of hypothetical proteins. *Proteomics* 2007;7:2920-2932.
19. Doerks T, von Mering C, Bork P. Functional clues for hypothetical proteins based on genomic context analysis in prokaryotes. *Nucleic Acids Res* 2004;32:6321-6326.
20. Gazi MA, Kibria MG, Mahfuz M, Islam MR, Ghosh P, Afsar MN, *et al*. Functional, structural and epitopic prediction of hypothetical proteins of *Mycobacterium tuberculosis* H37Rv: an *in silico* approach for prioritizing the targets. *Gene* 2016; 591:442-455.
21. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med* 1978;8:283-298.
22. Anandakumar S, Shanmughavel P. Computational annotation for hypothetical proteins of *Mycobacterium tuberculosis*. *J Comput Sci Syst Biol* 2008;1:50-62.
23. Galperin MY, Koonin EV. 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Res* 2004;32:5452-5463.
24. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. *EMBO J* 1986;5:823-826.
25. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25: 3389-3402.
26. Eddy SR. Profile hidden Markov models. *Bioinformatics* 1998;14:755-763.
27. Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, *et al*. CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res* 2007;35:D237-D240.
28. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, *et al*. The Pfam protein families database. *Nucleic Acids Res* 2002;30:276-280.
29. Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, *et al*. HMMER web server: 2015 update. *Nucleic Acids Res* 2015;43:W30-W38.
30. Letunic I, Doerks T, Bork P. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res* 2012; 40:D302-D305.
31. de Castro E, Sigrist CJ, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, *et al*. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res* 2006;34:W362-W365.
32. Shanmugham B, Pan A. Identification and characterization of potential therapeutic candidates in emerging human pathogen *Mycobacterium abscessus*: a novel hierarchical *in silico* approach. *PLoS One* 2013;8:e59126.
33. Yu CS, Chen YC, Lu CH, Hwang JK. Prediction of protein subcellular localization. *Proteins* 2006;64:643-651.
34. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, *et al*. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 2010;26:1608-1615.
35. Bhasin M, Garg A, Raghava GP. PSLpred: prediction of subcellular localization of bacterial proteins. *Bioinformatics* 2005; 21:2522-2524.
36. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 2001;305:567-580.
37. Hirokawa T, Boon-Chieng S, Mitaku S. SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* 1998;14:378-379.
38. Tusnady GE, Simon I. The HMMTOP transmembrane topology prediction server. *Bioinformatics* 2001;17:849-850.
39. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* 2011;8:785-786.
40. Bendtsen JD, Kiemer L, Fausboll A, Brunak S. Non-classical protein secretion in bacteria. *BMC Microbiol* 2005;5:58.
41. Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. *Nucleic Acids Res* 2002;30:42-46.
42. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, *et al*. InterProScan: protein domains identifier. *Nucleic Acids Res* 2005;33:W116-W120.
43. Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, Lee D, *et al*. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res* 2013;41:D490-D498.
44. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* 2001;313:903-919.
45. Rappoport N, Karsenty S, Stern A, Linial N, Linial M. ProtoNet 6.0: organizing 10 million protein sequences in a compact hierarchical family tree. *Nucleic Acids Res* 2012;40:D313-D320.
46. Xu D, Xu Y, Uberbacher EC. Computational tools for protein modeling. *Curr Protein Pept Sci* 2000;1:1-21.
47. Chen CC, Hwang JK, Yang JM. (PS)2-v2: template-based protein structure prediction server. *BMC Bioinformatics* 2009; 10:366.
48. Shen HB, Chou KC. Predicting protein fold pattern with functional domain and sequential evolution information. *J Theor Biol* 2009;256:441-446.
49. Baron C, Coombes B. Targeting bacterial secretion systems: benefits of disarmament in the microcosm. *Infect Disord Drug Targets* 2007;7:19-27.
50. Saha S, Raghava GP. VICMpred: an SVM-based method for the prediction of functional proteins of Gram-negative bacteria using amino acid patterns and composition. *Genomics Proteomics Bioinformatics* 2006;4:42-47.
51. Garg A, Gupta D. VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. *BMC Bioinformatics* 2008;9:62.
52. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, *et al*. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* 2011;39:D561-D568.
53. Eng J. ROC analysis: web-based calculator for ROC curves.

Baltimore: Johns Hopkins University, 2006. Accessed 2018 Sep 1. Available from: http://www.jrocfit.org.
54. Shahbaaz M, Hassan MI, Ahmad F. Functional annotation of conserved hypothetical proteins from Haemophilus influenzae Rd KW20. *PLoS One* 2013;8:e84263.
55. Delucia AM, Six DA, Caughlan RE, Gee P, Hunt I, Lam JS, *et al*. Lipopolysaccharide (LPS) inner-core phosphates are required for complete LPS synthesis and transport to the outer membrane in *Pseudomonas aeruginosa* PAO1. *MBio* 2011;2: e00142-11.
56. Burk DL, Ghuman N, Wybenga-Groot LE, Berghuis AM. X-ray structure of the AAC(6')-Ii antibiotic resistance enzyme at 1.8 A resolution: examination of oligomeric arrangements in GNAT superfamily members. *Protein Sci* 2003;12:426-437.
57. Bjornson HS. Enzymes associated with the survival and virulence of gram-negative anaerobes. *Rev Infect Dis* 1984;6 Suppl 1:S21-S24.
58. Morishita R, Kawagoshi A, Sawasaki T, Madin K, Ogasawara T, Oka T, *et al*. Ribonuclease activity of rat liver perchloric acid-soluble protein, a potent inhibitor of protein synthesis. *J Biol Chem* 1999;274:20688-20692.
59. Lambrecht JA, Flynn JM, Downs DM. Conserved YjgF protein family deaminates reactive enamine/imine intermediates of pyridoxal 5'-phosphate (PLP)-dependent enzyme reactions. *J Biol Chem* 2012;287:3454-3461.
60. Schmitz G, Downs DM. Reduced transaminase B (IlvE) activity caused by the lack of yjgF is dependent on the status of threonine deaminase (IlvA) in *Salmonella enterica* serovar Typhimurium. *J Bacteriol* 2004;186:803-810.
61. Aravind L, Leipe DD, Koonin EV. Toprim: a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. *Nucleic Acids Res* 1998;26:4205-4213.
62. Fry J, Wood M, Poole PS. Investigation of myo-inositol catabolism in *Rhizobium* leguminosarum bv. viciae and its effect on nodulation competitiveness. *Mol Plant Microbe Interact* 2001;14:1016-1025.
63. Bollinger JM Jr, Kwon DS, Huisman GW, Kolter R, Walsh CT. Glutathionylspermidine metabolism in *Escherichia coli:* purification, cloning, overproduction, and characterization of a bifunctional glutathionylspermidine synthetase/amidase. *J Biol Chem* 1995;270:14031-14041.
64. Ejim LJ, D'Costa VM, Elowe NH, Loredo-Osti JC, Malo D, Wright GD. Cystathionine beta-lyase is important for virulence of *Salmonella enterica* serovar Typhimurium. *Infect Immun* 2004;72:3310-3314.
65. Gerdes K, Wagner EG. RNA antitoxins. *Curr Opin Microbiol* 2007;10:117-124.
66. Kawano M, Aravind L, Storz G. An antisense RNA controls synthesis of an SOS-induced toxin evolved from an antitoxin. *Mol Microbiol* 2007;64:738-754.
67. Kawano M. Divergently overlapping cis-encoded antisense RNA regulating toxin-antitoxin systems from *E. coli*: hok/sok, ldr/rdl, symE/symR. *RNA Biol* 2012;9:1520-1527.
68. Ruggeri ZM, Ware J. von Willebrand factor. *FASEB J* 1993; 7:308-316.
69. Ahmad F, Jan R, Kannan M, Obser T, Hassan MI, Oyen F, *et al*. Characterisation of mutations and molecular studies of type 2 von Willebrand disease. *Thromb Haemost* 2013;109:39-46.
70. Naqvi AA, Shahbaaz M, Ahmad F, Hassan MI. Identification of functional candidates amongst hypothetical proteins of *Treponema pallidum* ssp. *pallidum*. *PLoS One* 2015;10:e0124177.
71. Colombatti A, Bonaldo P, Doliana R. Type A modules: interacting domains found in several non-fibrillar collagens and in other extracellular matrix proteins. *Matrix* 1993;13:297-306.
72. Tavernarakis N, Driscoll M, Kyrpides NC. The SPFH domain: implicated in regulating targeted protein turnover in stomatins and other membrane-associated proteins. *Trends Biochem Sci* 1999;24:425-427.
73. Gehl B, Sweetlove LJ. Mitochondrial Band-7 family proteins: scaffolds for respiratory chain assembly? *Front Plant Sci* 2014;5:141.
74. Wu T, McCandlish AC, Gronenberg LS, Chng SS, Silhavy TJ, Kahne D. Identification of a protein complex that assembles lipopolysaccharide in the outer membrane of *Escherichia coli*. *Proc Natl Acad Sci U S A* 2006;103:11754-11759.
75. Cerveny L, Straskova A, Dankova V, Hartlova A, Ceckova M, Staud F, *et al*. Tetratricopeptide repeat motifs in the world of bacterial pathogens: role in virulence mechanisms. *Infect Immun* 2013;81:629-635.
76. Singer HM, Kuhne C, Deditius JA, Hughes KT, Erhardt M. The *Salmonell*a Spi1 virulence regulatory protein HilD directly activates transcription of the flagellar master operon flhDC. *J Bacteriol* 2014;196:1448-1457.
77. Kowalski JC, Belfort M, Stapleton MA, Holpert M, Dansereau JT, Pietrokovski S, *et al*. Configuration of the catalytic GIY-YIG domain of intron endonuclease I-TevI: coincidence of computational and molecular findings. *Nucleic Acids Res* 1999;27: 2115-2125.
78. Van Roey P, Meehan L, Kowalski JC, Belfort M, Derbyshire V. Catalytic domain structure and hypothesis for function of GIY-YIG intron endonuclease I-TevI. *Nat Struct Biol* 2002;9: 806-811.
79. Iyer LM, Zhang D, Rogozin IB, Aravind L. Evolution of the deaminase fold and multiple origins of eukaryotic editing and mutagenic nucleic acid deaminases from bacterial toxin systems. *Nucleic Acids Res* 2011;39:9473-9497.
80. Shu W, Liu J, Ji H, Lu M. Core structure of the outer membrane lipoprotein from *Escherichia coli* at 1.9 A resolution. *J Mol Biol* 2000;299:1101-1112.
81. Rothberg JM, Jacobs JR, Goodman CS, Artavanis-Tsakonas S. slit: an extracellular protein necessary for development of midline glia and commissural axon pathways contains both EGF and LRR domains. *Genes Dev* 1990;4:2169-2187.
82. Kovacs-Simon A, Titball RW, Michell SL. Lipoproteins of bacterial pathogens. *Infect Immun* 2011;79:548-561.
83. Torti SV, Park JT. Lipoprotein of gram-negative bacteria is essential for growth and division. *Nature* 1976;263:323-326.
84. Saurin W, Hofnung M, Dassa E. Getting in or out: early segregation between importers and exporters in the evolution of ATP-binding cassette (ABC) transporters. *J Mol Evol* 1999;48: 22-41.
85. Freeman ZN, Dorus S, Waterfield NR. The KdpD/KdpE two-component system: integrating K(+) homeostasis and

virulence. *PLoS Pathog* 2013;9:e1003201.
86. Ibanez-Ruiz M, Robbe-Saule V, Hermant D, Labrude S, Norel F. Identification of RpoS (sigma(S))-regulated genes in *Salmonella enterica* serovar Typhimurium. *J Bacteriol* 2000; 182:5749-5756.
87. Peterson PA, Rask L, Ostberg L, Andersson L, Kamwendo F, Pertoft H. Studies on the transport and cellular distribution of vitamin A in normal and vitamin A-deficient rats with special reference to the vitamin A-binding plasma protein. *J Biol Chem* 1973;248:4009-4022.
88. Minailiuc OM, Vavelyuk O, Gandhi S, Hung MN, Cygler M, Ekiel I. NMR structure of YcgL, a conserved protein from *Escherichia coli* representing the DUF709 family, with a novel alpha/beta/alpha sandwich fold. *Proteins* 2007;66:1004-1007.
89. Livorsi DJ, Stenehjem E, Stephens DS. Virulence factors of gram-negative bacteria in sepsis with a focus on Neisseria meningitidis. In: *Sepsis*: *Pro-Inflammatory and Anti-Inflammatory Responses* (Herwald H, Egesten A, eds.). Basel: Karger Publishers, 2011. pp. 31-47.
90. Marra A. Targeting virulence for antibacterial chemotherapy: identifying and characterising virulence factors for lead discovery. *Drugs R D* 2006;7:1-16.

SUPPLEMENTARY INFORMATION

Functional Prediction of Hypothetical Proteins from *Shigella flexneri* and Validation of the Predicted Models by Using ROC Curve Analysis

Md. Amran Gazi[1*], Sultan Mahmud[2], Shah Mohammad Fahim[1], Mohammad Golam Kibria[2], Parag Palit[1], Md. Rezaul Islam[3], Humaira Rashid[2], Subhasish Das[1], Mustafa Mahfuz[1], Tahmeed Ahmeed[1]

[1]Nutrition and Clinical Services Division, International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b), Dhaka 1212, Bangladesh, [2]Infectious Diseases Division, International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b), Dhaka 1212, Bangladesh, [3]International Max Planck Research School, Grisebachstraße 5, 37077 Göttingen, Germany

https://doi.org/10.5808/GI.2018.16.4.e26

Supplementary Table 1. Scores of conserved domain search for 674 hypothetical proteins of *Shigella flexneri* serotype 2a strain 2457T using CDD-Blast, Pfam, HmmScan, SMART, and Scanprosite tools

Sl No.	Nucleotide ID	Accession ID_Protein	CDD Blast	Pfam	HmmScan	SMART	Scanprosite	Percentage
1	NC_004741.1	WP_000414150.1	0	0	0	0	0	0
2	NC_004741.1	WP_000738723.1	1	1	1	1	0	80
3	NC_004741.1	WP_001102351.1	1	1	1	1	0	80
4	NC_004741.1	WP_000843568.1	1	1	1	1	0	80
5	NC_004741.1	WP_032155592.1	0	0	0	0	0	0
6	NC_004741.1	WP_011110533.1	1	1	1	1	0	80
7	NC_004741.1	WP_032155631.1	0	0	0	0	0	0
8	NC_004741.1	WP_000196533.1	0	0	0	0	0	0
9	NC_004741.1	WP_001255305.1	0	0	0	0	0	0
10	NC_004741.1	WP_001303790.1	0	0	0	0	0	0
11	NC_004741.1	WP_000464383.1	1	0	0	0	0	20
12	NC_004741.1	WP_032155552.1	0	0	0	0	0	0
13	NC_004741.1	WP_005053505.1	1	1	1	1	0	80
14	NC_004741.1	WP_005094211.1	1	1	1	1	0	80
15	NC_004741.1	WP_001347263.1	0	0	0	0	0	0
16	NC_004741.1	WP_000964241.1	1	1	1	1	0	80
17	NC_004741.1	WP_000272188.1	1	1	1	1	0	80
18	NC_004741.1	WP_000417058.1	1	1	1	1	0	80
19	NC_004741.1	WP_005053355.1	1	1	1	1	1	100
20	NC_004741.1	WP_000402248.1	1	1	1	1	0	80
21	NC_004741.1	WP_024259146.1	0	0	0	0	0	0
22	NC_004741.1	WP_001276640.1	1	1	1	1	0	80
23	NC_004741.1	WP_000183806.1	1	1	1	1	0	80
24	NC_004741.1	WP_000343116.1	0	0	0	0	0	0

25	NC_004741.1	WP_000192453.1	1	1	1	1	0	80
26	NC_004741.1	WP_005053303.1	0	0	0	0	0	0
27	NC_004741.1	WP_000627639.1	1	1	1	1	0	80
28	NC_004741.1	WP_032333103.1	0	0	0	0	0	0
29	NC_004741.1	WP_001191885.1	0	0	0	0	0	0
30	NC_004741.1	WP_000092054.1	1	1	1	1	1	100
31	NC_004741.1	WP_001110751.1	0	0	0	0	0	0
32	NC_004741.1	WP_005053269.1	0	0	0	0	0	0
33	NC_004741.1	WP_000556489.1	0	0	0	0	0	0
34	NC_004741.1	WP_000003122.1	0	0	0	1	0	20
35	NC_004741.1	WP_001102097.1	1	1	1	1	0	80
36	NC_004741.1	WP_001201712.1	0	0	0	0	0	0
37	NC_004741.1	WP_000667623.1	0	0	0	0	0	0
38	NC_004741.1	WP_000343515.1	1	0	0	0	0	20
39	NC_004741.1	WP_005100645.1	1	1	1	1	1	80
40	NC_004741.1	WP_000554647.1	0	0	0	0	0	0
41	NC_004741.1	WP_005060548.1	0	0	0	0	0	0
42	NC_004741.1	WP_001142439.1	1	1	1	1	0	80
43	NC_004741.1	WP_032155747.1	0	0	0	0	0	0
44	NC_004741.1	WP_000941942.1	1	1	1	1	0	80
45	NC_004741.1	WP_000942006.1	1	1	1	1	0	80
46	NC_004741.1	WP_000194195.1	1	0	0	0	0	20
47	NC_004741.1	WP_001326928.1	0	1	0	1	0	40
48	NC_004741.1	WP_001382892.1	1	1	1	1	1	100
49	NC_004741.1	WP_000645013.1	0	0	0	0	0	0
50	NC_004741.1	WP_005060811.1	1	1	1	1	0	80
51	NC_004741.1	WP_005053055.1	0	0	0	0	0	0
52	NC_004741.1	WP_005053036.1	1	1	1	1	1	100

53	NC_004741.1	WP_001177122.1	0	0	0	0	0	0
54	NC_004741.1	WP_000680312.1	1	1	1	1	0	80
55	NC_004741.1	WP_005053020.1	1	1	1	1	0	80
56	NC_004741.1	WP_000779831.1	1	1	1	1	1	100
57	NC_004741.1	WP_000136192.1	1	1	1	1	0	80
58	NC_004741.1	WP_032140245.1	0	1	0	1	0	40
59	NC_004741.1	WP_001188905.1	1	1	1	1	0	80
60	NC_004741.1	WP_000970323.1	1	1	1	1	0	80
61	NC_004741.1	WP_000701358.1	1	1	1	1	0	80
62	NC_004741.1	WP_001224555.1	1	1	1	1	0	80
63	NC_004741.1	WP_000752567.1	1	1	1	1	0	80
64	NC_004741.1	WP_000360957.1	1	1	1	1	0	80
65	NC_004741.1	WP_064193753.1	0	0	0	0	0	0
66	NC_004741.1	WP_000460431.1	1	1	1	1	0	80
67	NC_004741.1	WP_000283754.1	1	1	1	1	0	80
68	NC_004741.1	WP_000287805.1	1	0	0	0	0	20
69	NC_004741.1	WP_001301130.1	0	0	0	0	0	0
70	NC_004741.1	WP_005083189.1	0	0	0	0	0	0
71	NC_004741.1	WP_001303843.1	0	0	0	0	0	0
72	NC_004741.1	WP_005049395.1	0	0	0	0	0	0
73	NC_004741.1	WP_005083246.1	0	0	0	0	0	0
74	NC_004741.1	WP_001188346.1	1	1	1	1	0	80
75	NC_004741.1	WP_001053303.1	1	1	1	1	0	80
76	NC_004741.1	WP_000807562.1	1	1	1	1	0	80
77	NC_004741.1	WP_025715253.1	0	0	0	0	0	0
78	NC_004741.1	WP_000405563.1	1	1	1	1	0	80
79	NC_004741.1	WP_001325427.1	0	0	0	0	0	0
80	NC_004741.1	WP_011110552.1	1	1	1	1	1	100

81	NC_004741.1	WP_000153125.1	1	1	1	1	0	80
82	NC_004741.1	WP_000232643.1	1	1	1	1	0	80
83	NC_004741.1	WP_005049464.1	0	0	0	1	0	20
84	NC_004741.1	WP_001030938.1	1	1	1	1	0	80
85	NC_004741.1	WP_000367140.1	1	1	1	1	0	80
86	NC_004741.1	WP_000578172.1	1	1	1	1	0	80
87	NC_004741.1	WP_001044870.1	1	1	1	1	0	80
88	NC_004741.1	WP_001269672.1	1	1	1	1	1	100
89	NC_004741.1	WP_000850550.1	1	1	1	1	0	80
90	NC_004741.1	WP_000073523.1	0	0	0	0	0	0
91	NC_004741.1	WP_005049496.1	0	0	0	0	0	0
92	NC_004741.1	WP_000627468.1	1	1	1	1	0	80
93	NC_004741.1	WP_005060669.1	0	0	0	0	0	0
94	NC_004741.1	WP_000873153.1	1	1	1	1	0	80
95	NC_004741.1	WP_000113500.1	1	1	1	1	0	80
96	NC_004741.1	WP_005020049.1	0	0	0	0	0	0
97	NC_004741.1	WP_005098291.1	0	0	0	0	0	0
98	NC_004741.1	WP_001108106.1	0	0	0	0	0	0
99	NC_004741.1	WP_000454800.1	1	1	1	1	0	80
100	NC_004741.1	WP_047199943.1	1	1	1	1	0	80
101	NC_004741.1	WP_001247854.1	1	1	1	1	1	100
102	NC_004741.1	WP_001343960.1	0	0	0	0	0	0
103	NC_004741.1	WP_005053437.1	0	0	0	0	0	0
104	NC_004741.1	WP_000551132.1	1	1	1	1	0	80
105	NC_004741.1	WP_000266134.1	1	0	0	0	0	20
106	NC_004741.1	WP_005048534.1	1	1	1	1	0	80
107	NC_004741.1	WP_000446914.1	1	1	1	1	0	80
108	NC_004741.1	WP_000871982.1	1	1	1	1	0	80

109	NC_004741.1	WP_000070107.1	1	1	1	1	1	100
110	NC_004741.1	WP_001336078.1	0	1	0	1	0	40
111	NC_004741.1	WP_005048500.1	1	1	1	1	0	80
112	NC_004741.1	WP_000849301.1	1	1	1	1	0	80
113	NC_004741.1	WP_000710620.1	1	1	1	1	0	80
114	NC_004741.1	WP_000188784.1	0	0	0	0	0	0
115	NC_004741.1	WP_000168813.1	1	1	1	1	0	80
116	NC_004741.1	WP_001295900.1	0	0	0	0	0	0
117	NC_004741.1	WP_000681108.1	1	1	1	1	0	80
118	NC_004741.1	WP_001201557.1	1	1	1	1	0	80
119	NC_004741.1	WP_000389260.1	1	1	1	1	0	80
120	NC_004741.1	WP_001303862.1	0	0	0	0	0	0
121	NC_004741.1	WP_001160722.1	1	1	1	1	0	80
122	NC_004741.1	WP_032155760.1	0	0	0	0	0	0
123	NC_004741.1	WP_000687442.1	0	0	0	0	0	0
124	NC_004741.1	WP_029716636.1	1	1	1	1	0	80
125	NC_004741.1	WP_005048249.1	0	0	1	0	0	20
126	NC_004741.1	WP_001118167.1	1	1	1	1	0	80
127	NC_004741.1	WP_000702036.1	1	1	1	1	0	80
128	NC_004741.1	WP_001091985.1	1	1	1	1	0	80
129	NC_004741.1	WP_001005968.1	0	0	0	0	0	0
130	NC_004741.1	WP_005051132.1	1	0	0	0	0	20
131	NC_004741.1	WP_005061679.1	1	1	1	1	0	80
132	NC_004741.1	WP_001039888.1	1	0	1	0	0	40
133	NC_004741.1	WP_000723623.1	1	1	1	1	0	80
134	NC_004741.1	WP_000959226.1	0	0	0	0	0	0
135	NC_004741.1	WP_000350058.1	1	1	1	1	0	80
136	NC_004741.1	WP_000196607.1	0	0	0	0	0	0

137	NC_004741.1	WP_000235193.1	1	0	0	0	0	20
138	NC_004741.1	WP_000224274.1	1	1	1	1	1	100
139	NC_004741.1	WP_001261235.1	1	1	1	1	0	80
140	NC_004741.1	WP_000847791.1	1	1	1	1	0	80
141	NC_004741.1	WP_001301416.1	0	0	0	0	0	0
142	NC_004741.1	WP_001038092.1	1	1	1	1	0	80
143	NC_004741.1	WP_005083611.1	0	0	0		0	0
144	NC_004741.1	WP_000505101.1	1	0	0	0	0	20
145	NC_004741.1	WP_000535353.1	1	0	0	0	0	20
146	NC_004741.1	WP_001143120.1	1	1	1	1	0	80
147	NC_004741.1	WP_000124106.1	0	0	0	0	0	0
148	NC_004741.1	WP_000611853.1	0	0	0	0	0	0
149	NC_004741.1	WP_001297187.1	1	1	1	1	0	80
150	NC_004741.1	WP_032155907.1	0	0	0	0	0	0
151	NC_004741.1	WP_001111218.1	0	0	0	0	0	0
152	NC_004741.1	WP_005047366.1	0	0		0	0	0
153	NC_004741.1	WP_000818776.1	0	0	0	0	0	0
154	NC_004741.1	WP_000749269.1	1	1	1	1	1	100
155	NC_004741.1	WP_000877111.1	1	1	1	1	0	80
156	NC_004741.1	WP_001295962.1	1	1	1	1	0	80
157	NC_004741.1	WP_000587933.1	1	1	1	1	0	80
158	NC_004741.1	WP_001043459.1	1	1	1	1	0	80
159	NC_004741.1	WP_032155646.1	0	0	0	0	0	0
160	NC_004741.1	WP_000103622.1	0	0	0	0	0	0
161	NC_004741.1	WP_005005155.1	0	0	0	0	0	0
162	NC_004741.1	WP_001204964.1	0	0	0	0	1	20
163	NC_004741.1	WP_000770157.1	1	1	1	1	0	80
164	NC_004741.1	WP_000557473.1	0	0	0	0	0	0

165	NC_004741.1	WP_001294167.1	0	0	0	0	0	0
166	NC_004741.1	WP_001132078.1	0	0	0	0	0	0
167	NC_004741.1	WP_000267598.1	1	1	1	1	0	80
168	NC_004741.1	WP_000134107.1	1	1	1	1	0	80
169	NC_004741.1	WP_001005703.1	0	0	0	0	0	0
170	NC_004741.1	WP_029716858.1	1	1	1	1	0	80
171	NC_004741.1	WP_000133415.1	0	0	0	0	0	0
172	NC_004741.1	WP_005047957.1	0	0	0	0	0	0
173	NC_004741.1	WP_032155828.1	0	0	0	0	0	0
174	NC_004741.1	WP_005047951.1	0	0	0	0	0	0
175	NC_004741.1	WP_001295611.1	1	1	1	1	0	80
176	NC_004741.1	WP_000122462.1	0	0	0	0	0	0
177	NC_004741.1	WP_005061990.1	0	0	0	0	0	0
178	NC_004741.1	WP_001125713.1	1	1	1	1	1	100
179	NC_004741.1	WP_000807626.1	1	1	1	1	0	80
180	NC_004741.1	WP_011069340.1	0	0	0	0	0	0
181	NC_004741.1	WP_000280742.1	0	0	0	0	0	0
182	NC_004741.1	WP_001257042.1	1	1	1	1	0	80
183	NC_004741.1	WP_000950192.1	1	1	1	1	0	80
184	NC_004741.1	WP_001169669.1	1	1	1	1	0	80
185	NC_004741.1	WP_000069487.1	0	0	0	0	0	0
186	NC_004741.1	WP_005105319.1	1	1	1	1	0	80
187	NC_004741.1	WP_001303937.1	0	0	0	0	0	0
188	NC_004741.1	WP_000967595.1	1	1	1	1	0	80
189	NC_004741.1	WP_000028536.1	1	1	1	1	0	80
190	NC_004741.1	WP_000807659.1	1	1	1	1	0	80
191	NC_004741.1	WP_001303289.1	0	0	0	0	0	0
192	NC_004741.1	WP_001031530.1	1	1	1	1	0	80

193	NC_004741.1	WP_014640285.1	0	0	0	0	0	0
194	NC_004741.1	WP_000233043.1	0	0	0	0	0	0
195	NC_004741.1	WP_000616081.1	0	0	0	0	0	0
196	NC_004741.1	WP_001288369.1	1	0	0	0	0	0
197	NC_004741.1	WP_001331106.1	1	1	1	1	0	80
198	NC_004741.1	WP_023636694.1	0	0	0	0	1	20
199	NC_004741.1	WP_000825769.1	1	1	1	1	0	80
200	NC_004741.1	WP_032155686.1	1	0	0	0	0	20
201	NC_004741.1	WP_000124119.1	0	0	0	0	0	0
202	NC_004741.1	WP_001296046.1	0	0	0	0	0	0
203	NC_004741.1	WP_005047705.1	1	0	0	0	0	20
204	NC_004741.1	WP_005047713.1	1	1	1	1	0	80
205	NC_004741.1	WP_023517638.1	1	1	1	1	0	80
206	NC_004741.1	WP_011069401.1	1	1	1	1	0	80
207	NC_004741.1	WP_000431885.1	1	1	1	1	0	80
208	NC_004741.1	WP_011069399.1	0	0	0	0	0	0
209	NC_004741.1	WP_001296778.1	0	0	0	0	0	80
210	NC_004741.1	WP_001077956.1	0	0	0	0	0	0
211	NC_004741.1	WP_000554382.1	1	0	0	0	0	20
212	NC_004741.1	WP_005049838.1	1	1	1	1	0	80
213	NC_004741.1	WP_001295499.1	1	1	1	1	0	80
214	NC_004741.1	WP_001043881.1	1	1	1	1	1	100
215	NC_004741.1	WP_001006860.1	1	1	1	1	0	80
216	NC_004741.1	WP_032155836.1	0	0	0	0	0	0
217	NC_004741.1	WP_000156246.1	1	1	1	1	0	80
218	NC_004741.1	WP_001306763.1	0	0	0	0	0	0
219	NC_004741.1	WP_001295493.1	1	1	1	1	1	100
220	NC_004741.1	WP_032155854.1	0	0	0	0	1	20

221	NC_004741.1	WP_000691930.1	1	1	1	1	1	100
222	NC_004741.1	WP_005126892.1	0	0	0	0	0	0
223	NC_004741.1	WP_000138039.1	1	1	1	1	0	80
224	NC_004741.1	WP_001046790.1	1	1	1	1	0	80
225	NC_004741.1	WP_001453023.1	0	0	0	0	0	0
226	NC_004741.1	WP_012602004.1	0	0	0	0	0	0
227	NC_004741.1	WP_000756955.1	1	1	1	1	0	80
228	NC_004741.1	WP_000085238.1	1	1	1	1	0	80
229	NC_004741.1	WP_001215295.1	1	0	0	0	0	20
230	NC_004741.1	WP_000077934.1	0	0	0	0	1	20
231	NC_004741.1	WP_032145487.1	0	1	0	1	0	40
232	NC_004741.1	WP_001297653.1	1	1	1	1	0	80
233	NC_004741.1	WP_000146138.1	1	1	1	1	0	80
234	NC_004741.1	WP_001142445.1	0	0	0	0	0	0
235	NC_004741.1	WP_005050031.1	1	1	1	1	0	80
236	NC_004741.1	WP_042003723.1	0	0	0	0	0	0
237	NC_004741.1	WP_000398613.1	0	0	0	0	0	0
238	NC_004741.1	WP_005062520.1	0	0	0	0	0	0
239	NC_004741.1	WP_000726666.1	1	1	1	1	0	80
240	NC_004741.1	WP_000874243.1	1	1	1	1	0	80
241	NC_004741.1	WP_045177689.1	0	0	0	0	0	0
242	NC_004741.1	WP_001265249.1	1	1	1	1	0	80
243	NC_004741.1	WP_000980987.1	0	0	0	0	0	0
244	NC_004741.1	WP_000214712.1	1	1	1	1	0	80
245	NC_004741.1	WP_001024558.1	1	1	1	1	0	80
246	NC_004741.1	WP_000901367.1	0	0	0	0	0	0
247	NC_004741.1	WP_000258546.1	0	0	0	0	0	0
248	NC_004741.1	WP_000957853.1	0	0	0	0	0	0

249	NC_004741.1	WP_005050130.1	1	1	1	1	0	80
250	NC_004741.1	WP_001295395.1	1	1	1	1	0	80
251	NC_004741.1	WP_000705197.1	1	1	1	1	0	80
252	NC_004741.1	WP_000234660.1	0	0	0	0	0	0
253	NC_004741.1	WP_000520318.1	0	0	1	1	0	40
254	NC_004741.1	WP_000207512.1	0	0	0	0	0	0
255	NC_004741.1	WP_000971490.1	1	0	1	0	0	40
256	NC_004741.1	WP_001240758.1	0	0	0	0	0	0
257	NC_004741.1	WP_000199921.1	0	0	0	0	0	0
258	NC_004741.1	WP_001091024.1	0	0	0	0	0	0
259	NC_004741.1	WP_000113584.1	0	0	0	0	0	0
260	NC_004741.1	WP_000091718.1	0	0	0	0	0	0
261	NC_004741.1	WP_001249851.1	0	0	0	0	0	0
262	NC_004741.1	WP_000233090.1	1	0	0	0	0	20
263	NC_004741.1	WP_000769323.1	1	1	1	1	0	80
264	NC_004741.1	WP_000524868.1	1	1	1	1	0	80
265	NC_004741.1	WP_000597196.1	1	1	1	1	1	100
266	NC_004741.1	WP_032155892.1	1	0	0	0	0	20
267	NC_004741.1	WP_000534313.1	1	1	1	1	0	80
268	NC_004741.1	WP_000212657.1	1	1	1	1	0	80
269	NC_004741.1	WP_000587595.1	1	0	0	0	0	20
270	NC_004741.1	WP_001344535.1	1	0	0	0	0	20
271	NC_004741.1	WP_000528342.1	1	1	1	1	0	80
272	NC_004741.1	WP_001296104.1	1	1	1	1	0	80
273	NC_004741.1	WP_000248636.1	1	1	1	1	1	100
274	NC_004741.1	WP_001301287.1	1	1	1	1	0	80
275	NC_004741.1	WP_032155900.1	0	0	0	0	0	0
276	NC_004741.1	WP_000627104.1	1	0	0	0	1	40

277	NC_004741.1	WP_000124121.1	0	0	1	0	0	20
278	NC_004741.1	WP_000018633.1	1	1	1	1	0	80
279	NC_004741.1	WP_012817775.1	0	0	0	0	0	0
280	NC_004741.1	WP_032155659.1	0	0	0	0	0	0
281	NC_004741.1	WP_001380520.1	1	1	1	1	0	80
282	NC_004741.1	WP_000879272.1	1	0	0	0	0	20
283	NC_004741.1	WP_000168747.1	1	1	1	1	0	80
284	NC_004741.1	WP_000275187.1	0	0	0	0	0	0
285	NC_004741.1	WP_005047608.1	1	1	1	1	0	80
286	NC_004741.1	WP_000155622.1	0	0	0	0	0	0
287	NC_004741.1	WP_001024930.1	1	1	1	1	0	80
288	NC_004741.1	WP_005084198.1	0	0	0	0	0	0
289	NC_004741.1	WP_001039885.1	1	0	1	0	0	40
290	NC_004741.1	WP_000930145.1	1	1	1	1	0	80
291	NC_004741.1	WP_000009987.1	0	0	0	0	0	0
292	NC_004741.1	WP_000245528.1	1	1	1	1	0	80
293	NC_004741.1	WP_005049830.1	1	0	0	0	0	20
294	NC_004741.1	WP_001173294.1	1	1	1	0	0	60
295	NC_004741.1	WP_000930141.1	1	1	1	1	0	80
296	NC_004741.1	WP_001007942.1	0	0	0	0	0	0
297	NC_004741.1	WP_000082749.1	1	1	1	1	0	80
298	NC_004741.1	WP_005048633.1	0	0	0	0	0	0
299	NC_004741.1	WP_001028876.1	0	0	0	0	0	0
300	NC_004741.1	WP_000755956.1	1	1	1	1	1	100
301	NC_004741.1	WP_001099210.1	1	1	1	1	0	80
302	NC_004741.1	WP_000586688.1	1	0	0	0	0	20
303	NC_004741.1	WP_000457719.1	1	1	1	1	0	80
304	NC_004741.1	WP_001030133.1	0	0	0	0	0	0

305	NC_004741.1	WP_005063152.1	1	1	1	1	0	80
306	NC_004741.1	WP_000455174.1	1	1	1	1	0	80
307	NC_004741.1	WP_001103659.1	1	0	0	0	0	20
308	NC_004741.1	WP_042791229.1	0	0	0	0	0	0
309	NC_004741.1	WP_000082120.1	1	1	1	1	0	80
310	NC_004741.1	WP_001297814.1	0	0	0	0	0	0
311	NC_004741.1	WP_001237866.1	1	1	1	1	1	100
312	NC_004741.1	WP_000377229.1	1	1	1	1	0	80
313	NC_004741.1	WP_032155628.1	0	0	0	0	0	0
314	NC_004741.1	WP_000106474.1	1	1	1	1	0	80
315	NC_004741.1	WP_000118898.1	1	1	1	1	0	80
316	NC_004741.1	WP_000230645.1	0	0	0	0	0	0
317	NC_004741.1	WP_024259260.1	0	0	0	0	0	0
318	NC_004741.1	WP_005048789.1	1	1	1	1	0	80
319	NC_004741.1	WP_000431460.1	1	0	0	0	0	20
320	NC_004741.1	WP_032155863.1	1	0	0	0	0	20
321	NC_004741.1	WP_032155819.1	0	0	0	0	0	0
322	NC_004741.1	WP_001039899.1	0	0	0	0	0	0
323	NC_004741.1	WP_064716611.1	1	1	1	1	0	80
324	NC_004741.1	WP_000594909.1	0	0	0	0	1	20
325	NC_004741.1	WP_001062338.1	1	1	1	1	0	80
326	NC_004741.1	WP_001265248.1	1	1	1	1	0	80
327	NC_004741.1	WP_000152435.1	0	0	0	0	0	0
328	NC_004741.1	WP_032155691.1	1	0	0	0	0	20
329	NC_004741.1	WP_001343759.1	0	0	0	0	0	0
330	NC_004741.1	WP_000466572.1	1	0	0	0	0	20
331	NC_004741.1	WP_061440266.1	1	0	0	0	0	20
332	NC_004741.1	WP_001016348.1	1	1	1	1	0	80

333	NC_004741.1	WP_000450409.1	1	1	1	1	0	80
334	NC_004741.1	WP_005088730.1	0	0	1	0	0	20
335	NC_004741.1	WP_000282151.1	1	1	1	1	0	80
336	NC_004741.1	WP_001243860.1	1	1	1	1	0	80
337	NC_004741.1	WP_011110604.1	0	0	0	0	0	0
338	NC_004741.1	WP_011069433.1	0	0	0	0	0	0
339	NC_004741.1	WP_000055830.1	0	0	0	0	0	0
340	NC_004741.1	WP_011069434.1	1	0	1	0	0	40
341	NC_004741.1	WP_000454701.1	1	1	1	1	1	100
342	NC_004741.1	WP_000489605.1	1	0	0	0	0	20
343	NC_004741.1	WP_000003197.1	1	1	1	1	1	100
344	NC_004741.1	WP_000929408.1	1	1	1	1	0	80
345	NC_004741.1	WP_014532286.1	0	0	0	0	0	0
346	NC_004741.1	WP_001324860.1	0	0	0	0	1	20
347	NC_004741.1	WP_000830460.1	1	0	1	0	0	40
348	NC_004741.1	WP_011110605.1	1	1	1	1	0	80
349	NC_004741.1	WP_000261596.1	0	0	0	0	0	0
350	NC_004741.1	WP_005049040.1	0	0	0	0	0	0
351	NC_004741.1	WP_005098884.1	1	1	1	1	0	80
352	NC_004741.1	WP_005049034.1	1	1	1	1	0	80
353	NC_004741.1	WP_024259269.1	1	1	1	1	0	80
354	NC_004741.1	WP_000636931.1	0	0	0	0	0	0
355	NC_004741.1	WP_000380421.1	0	0	0	0	0	0
356	NC_004741.1	WP_001294399.1	0	0	0	0	0	0
357	NC_004741.1	WP_005049026.1	1	1	1	1	0	80
358	NC_004741.1	WP_001087240.1	1	0	1	0	1	60
359	NC_004741.1	WP_005049020.1	1	1	1	1	1	100
360	NC_004741.1	WP_048814497.1	1	1	1	1	1	100

361	NC_004741.1	WP_011069443.1	0	0	0	0	0	0
362	NC_004741.1	WP_000691708.1	0	0	0	0	0	0
363	NC_004741.1	WP_001295452.1	1	1	1	1	0	80
364	NC_004741.1	WP_032155550.1	0	0	0	0	0	0
365	NC_004741.1	WP_001308773.1	0	0	0	0	0	0
366	NC_004741.1	WP_000182053.1	1	1	1	1	0	80
367	NC_004741.1	WP_000202798.1	1	1	1	1	0	80
368	NC_004741.1	WP_001135673.1	1	1	1	1	0	80
369	NC_004741.1	WP_001303596.1	0	0	0	0	0	0
370	NC_004741.1	WP_001296837.1	0	0	0	0	0	0
371	NC_004741.1	WP_001225855.1	1	1	1	1	0	80
372	NC_004741.1	WP_001104543.1	1	1	1	1	0	80
373	NC_004741.1	WP_005046832.1	1	0	0	0	0	20
374	NC_004741.1	WP_001215763.1	1	1	1	1	0	80
375	NC_004741.1	WP_032083391.1	0	0	0	0	0	0
376	NC_004741.1	WP_000301054.1	1	1	1	1	1	100
377	NC_004741.1	WP_001009396.1	1	0	0	0	0	20
378	NC_004741.1	WP_000140529.1	1	1	1	1	0	80
379	NC_004741.1	WP_000070619.1	1	1	1	1	0	80
380	NC_004741.1	WP_001446945.1	1	1	1	1	0	80
381	NC_004741.1	WP_000525371.1	1	1	1	1	0	80
382	NC_004741.1	WP_000426116.1	1	1	1	1	0	80
383	NC_004741.1	WP_000106622.1	1	1	1	1	0	80
384	NC_004741.1	WP_001115612.1	0	0	0	0	0	0
385	NC_004741.1	WP_001274496.1	1	1	1	1	0	80
386	NC_004741.1	WP_000559763.1	1	1	1	1	0	80
387	NC_004741.1	WP_005070009.1	1	1	1	1	0	80
388	NC_004741.1	WP_000937783.1	0	0	0	0	0	0

389	NC_004741.1	WP_000937210.1	1	1	1	1	0	80
390	NC_004741.1	WP_000825597.1	1	1	1	1	0	80
391	NC_004741.1	WP_000867638.1	0	0	0	0	1	20
392	NC_004741.1	WP_009008053.1	0	0	0	0	0	0
393	NC_004741.1	WP_000639883.1	1	1	1	1	0	80
394	NC_004741.1	WP_000490072.1	1	1	1	1	0	80
395	NC_004741.1	WP_042188255.1	0	0	0	0	0	0
396	NC_004741.1	WP_000826512.1	1	1	1	1	0	80
397	NC_004741.1	WP_000201413.1	1	1	1	1	0	80
398	NC_004741.1	WP_000719924.1	1	0	1	0	0	40
399	NC_004741.1	WP_001107736.1	0	0	0	0	0	0
400	NC_004741.1	WP_000555795.1	0	0	0	0	0	0
401	NC_004741.1	WP_000338539.1	0	0	0	0	0	0
402	NC_004741.1	WP_001373377.1	0	0	0	0	0	0
403	NC_004741.1	WP_000806589.1	0	0	0	0	0	0
404	NC_004741.1	WP_001507728.1	1	1	1	1	0	80
405	NC_004741.1	WP_001308835.1	0	0	0	0	0	0
406	NC_004741.1	WP_001349976.1	1	0	0	0	0	20
407	NC_004741.1	WP_000339447.1	1	1	1	1	0	80
408	NC_004741.1	WP_001244758.1	1	0	0	0	0	20
409	NC_004741.1	WP_000017552.1	0	0	0	0	0	0
410	NC_004741.1	WP_000076001.1	1	1	1	1	0	80
411	NC_004741.1	WP_000755178.1	0	0	1	0	0	20
412	NC_004741.1	WP_000743213.1	0	0	0	0	0	0
413	NC_004741.1	WP_000131871.1	0	0	0	0	0	0
414	NC_004741.1	WP_000211355.1	1	1	1	1	0	80
415	NC_004741.1	WP_000213809.1	0	0	0	0	0	0
416	NC_004741.1	WP_001162384.1	1	1	1	1	0	80

417	NC_004741.1	WP_005047279.1	1	1	1	1	0	80
418	NC_004741.1	WP_001094726.1	1	1	1	1	0	80
419	NC_004741.1	WP_000266171.1	1	1	1	1	1	100
420	NC_004741.1	WP_000951754.1	1	0	0	0	0	20
421	NC_004741.1	WP_000647601.1	1	0	0	0	0	20
422	NC_004741.1	WP_012135949.1	0	0	0	0	0	0
423	NC_004741.1	WP_001303621.1	0	0	0	0	0	0
424	NC_004741.1	WP_032155618.1	0	0	0	0	0	0
425	NC_004741.1	WP_001330697.1	0	0	0	0	0	0
426	NC_004741.1	WP_001212392.1	1	1	1	1	0	80
427	NC_004741.1	WP_000589825.1	1	1	1	1	1	100
428	NC_004741.1	WP_000284119.1	0	0	0	0	0	0
429	NC_004741.1	WP_000491410.1	1	1	1	1	0	80
430	NC_004741.1	WP_000281320.1	1	1	1	1	0	80
431	NC_004741.1	WP_000483311.1	1	0	0	0	0	20
432	NC_004741.1	WP_001287454.1	1	1	1	1	0	80
433	NC_004741.1	WP_001307965.1	0	0	0	0	0	0
434	NC_004741.1	WP_000493764.1	1	0	0	0	0	20
435	NC_004741.1	WP_032155822.1	0	0	0	0	0	0
436	NC_004741.1	WP_000611930.1	0	0	0	0	0	0
437	NC_004741.1	WP_001224024.1	1	0	0	0	0	20
438	NC_004741.1	WP_001288227.1	0	0	0	0	0	0
439	NC_004741.1	WP_000444999.1	1	1	1	1	0	80
440	NC_004741.1	WP_000206987.1	1	1	1	1	0	80
441	NC_004741.1	WP_001393510.1	0	0	0	0	0	0
442	NC_004741.1	WP_000203905.1	1	1	1	1	0	80
443	NC_004741.1	WP_000184250.1	1	1	1	1	0	80
444	NC_004741.1	WP_001078387.1	1	1	1	1	0	80

445	NC_004741.1	WP_032155588.1	0	0	0	0	0	0
446	NC_004741.1	WP_000860229.1	0	0	0	0	0	0
447	NC_004741.1	WP_000242461.1	1	0	0	0	0	20
448	NC_004741.1	WP_000379402.1	0	0	0	0	0	0
449	NC_004741.1	WP_005099217.1	0	0	0	0	0	0
450	NC_004741.1	WP_005099219.1	0	0	0	0	0	0
451	NC_004741.1	WP_000971492.1	1	0	0	0	0	20
452	NC_004741.1	WP_005051685.1	1	1	1	1	1	100
453	NC_004741.1	WP_000528349.1	0	0	0	0	0	0
454	NC_004741.1	WP_001010156.1	0	0	0	0	0	0
455	NC_004741.1	WP_005051767.1	1	1	1	1	0	80
456	NC_004741.1	WP_011110620.1	0	1	1	0	0	40
457	NC_004741.1	WP_001094817.1	1	1	1	1	0	80
458	NC_004741.1	WP_000745204.1	1	1	1	1	0	80
459	NC_004741.1	WP_000984792.1	1	1	1	1	0	80
460	NC_004741.1	WP_005064025.1	0	1	0	1	0	40
461	NC_004741.1	WP_000338035.1	0	0	0	0	0	0
462	NC_004741.1	WP_000291751.1	0	0	0	0	0	0
463	NC_004741.1	WP_001323220.1	0	0	0	0	0	0
464	NC_004741.1	WP_032142224.1	0	0	0	0	0	0
465	NC_004741.1	WP_001128940.1	1	1	1	1	0	80
466	NC_004741.1	WP_001013320.1	1	1	1	1	0	80
467	NC_004741.1	WP_000271035.1	1	1	1	1	0	80
468	NC_004741.1	WP_001195464.1	1	1	1	1	0	80
469	NC_004741.1	WP_011069510.1	0	0	0	0	0	0
470	NC_004741.1	WP_005093820.1	0	0	0	0	1	20
471	NC_004741.1	WP_000261147.1	0	0	0	0	0	0
472	NC_004741.1	WP_005051842.1	0	0	0	0	0	0

473	NC_004741.1	WP_001069724.1	1	1	1	1	0	80
474	NC_004741.1	WP_005093816.1	0	0	0	0	0	0
475	NC_004741.1	WP_005051844.1	0	0	0	0	0	0
476	NC_004741.1	WP_001387238.1	1	1	1	1	1	100
477	NC_004741.1	WP_000692350.1	1	1	1	1	0	80
478	NC_004741.1	WP_000761715.1	0	0	0	0	0	0
479	NC_004741.1	WP_000772029.1	1	1	1	1	0	80
480	NC_004741.1	WP_000340141.1	0	0	0	0	0	0
481	NC_004741.1	WP_000853257.1	1	1	1	1	0	80
482	NC_004741.1	WP_000248097.1	1	1	1	1	1	100
483	NC_004741.1	WP_000984979.1	0	0	0	0	0	0
484	NC_004741.1	WP_000339534.1	1	0	0	0	0	20
485	NC_004741.1	WP_005051896.1	0	0	0	0	0	0
486	NC_004741.1	WP_001059136.1	1	1	1	1	0	80
487	NC_004741.1	WP_024167679.1	0	0	0	0	0	0
488	NC_004741.1	WP_001298764.1	0	0	0	0	0	0
489	NC_004741.1	WP_000691640.1	1	1	1	1	0	80
490	NC_004741.1	WP_000848528.1	1	1	1	1	1	100
491	NC_004741.1	WP_000527661.1	0	0	0	0	0	0
492	NC_004741.1	WP_001701108.1	0	0	0	0	0	0
493	NC_004741.1	WP_000442868.1	1	1	1	1	0	80
494	NC_004741.1	WP_001406537.1	0	0	0	0	0	0
495	NC_004741.1	WP_024259304.1	0	0	0	0	0	0
496	NC_004741.1	WP_001298386.1	0	0	0	0	0	0
497	NC_004741.1	WP_032140301.1	0	0	0	0	0	0
498	NC_004741.1	WP_000942538.1	1	1	1	1	0	80
499	NC_004741.1	WP_000016819.1	1	1	1	1	0	80
500	NC_004741.1	WP_000422149.1	1	1	1	1	0	80

501	NC_004741.1	WP_005050960.1	1	1	1	1	0	80
502	NC_004741.1	WP_000031415.1	1	1	1	1	0	80
503	NC_004741.1	WP_000785722.1	1	1	1	1	0	80
504	NC_004741.1	WP_000096080.1	1	1	1	1	0	80
505	NC_004741.1	WP_000732225.1	1	1	1	1	0	80
506	NC_004741.1	WP_005089560.1	0	0	0	0	0	0
507	NC_004741.1	WP_005050890.1	0	0	0	0	0	0
508	NC_004741.1	WP_001343556.1	1	1	1	1	0	80
509	NC_004741.1	WP_000449030.1	1	0	0	0	0	20
510	NC_004741.1	WP_000189314.1	1	1	1	1	1	100
511	NC_004741.1	WP_001346700.1	0	0	0	0	0	0
512	NC_004741.1	WP_000620405.1	1	1	1	1	0	80
513	NC_004741.1	WP_005077180.1	0	0	0	0	0	0
514	NC_004741.1	WP_005050708.1	1	1	1	1	0	80
515	NC_004741.1	WP_001028769.1	1	1	1	1	0	80
516	NC_004741.1	WP_001303690.1	0	0	0	0	0	0
517	NC_004741.1	WP_011110633.1	0	0	0	0	0	0
518	NC_004741.1	WP_001061203.1	0	0	0	0	0	0
519	NC_004741.1	WP_001326891.1	1	0	0	0	0	20
520	NC_004741.1	WP_024259309.1	1	1	1	1	0	80
521	NC_004741.1	WP_005050690.1	0	0	0	0	0	0
522	NC_004741.1	WP_000460680.1	1	1	1	1	0	80
523	NC_004741.1	WP_000757326.1	0	1	1	1	0	60
524	NC_004741.1	WP_000595564.1	1	1	1	1	0	80
525	NC_004741.1	WP_000487766.1	0	0	0	0	0	0
526	NC_004741.1	WP_024166609.1	0	0	0	0	0	0
527	NC_004741.1	WP_005050602.1	0	0	0	0	0	0
528	NC_004741.1	WP_011069598.1	1	1	1	1	0	80

529	NC_004741.1	WP_000155673.1	1	1	1	1	0	80
530	NC_004741.1	WP_001014565.1	1	1	1	1	0	80
531	NC_004741.1	WP_001324833.1	0	0	0	0	0	0
532	NC_004741.1	WP_000719886.1	1	1	1	1	0	80
533	NC_004741.1	WP_001112357.1	1	0	0	0	0	20
534	NC_004741.1	WP_000169147.1	1	1	1	1	0	80
535	NC_004741.1	WP_032155643.1	0	0	0	0	0	0
536	NC_004741.1	WP_000627171.1	1	1	1	1	0	80
537	NC_004741.1	WP_023517643.1	0	1	0	1	0	40
538	NC_004741.1	WP_046201574.1	1	0	0	0	0	20
539	NC_004741.1	WP_000940102.1	1	1	1	1	0	80
540	NC_004741.1	WP_011110634.1	0	0	0	0	0	0
541	NC_004741.1	WP_001442985.1	0	1	0	1	0	40
542	NC_004741.1	WP_000660586.1	0	0	0	0	0	0
543	NC_004741.1	WP_000797352.1	1	1	1	1	0	80
544	NC_004741.1	WP_000655986.1	1	1	1	1	0	80
545	NC_004741.1	WP_000802226.1	1	1	1	1	0	80
546	NC_004741.1	WP_000591073.1	0	0	0	0	0	0
547	NC_004741.1	WP_000510376.1	0	0	0	0	0	0
548	NC_004741.1	WP_014334093.1	0	1	0	1	0	40
549	NC_004741.1	WP_001086388.1	0	0	0	0	0	0
550	NC_004741.1	WP_001295676.1	1	1	1	1	0	80
551	NC_004741.1	WP_032155602.1	0	0	0	0	0	0
552	NC_004741.1	WP_024259324.1	0	0	1	0	0	20
553	NC_004741.1	WP_000256409.1	1	1	1	1	0	80
554	NC_004741.1	WP_032142137.1	0	0	0	0	0	0
555	NC_004741.1	WP_000893994.1	0	0	0	0	0	0
556	NC_004741.1	WP_000115988.1	0	0	0	0	0	0

557	NC_004741.1	WP_001295264.1	1	1	1	1	0	80
558	NC_004741.1	WP_005052859.1	1	1	1	1	0	80
559	NC_004741.1	WP_001277142.1	1	1	1	1	0	80
560	NC_004741.1	WP_024259323.1	1	0	0	0	0	20
561	NC_004741.1	WP_000032578.1	1	1	1	1	0	80
562	NC_004741.1	WP_001127088.1	0	0	0	0	0	0
563	NC_004741.1	WP_000841001.1	1	1	1	1	0	80
564	NC_004741.1	WP_000336276.1	0	0	0	0	0	0
565	NC_004741.1	WP_000456043.1	0	0	0	0	0	0
566	NC_004741.1	WP_005051995.1	1	1	1	1	0	80
567	NC_004741.1	WP_000454294.1	0	0	0	0	0	0
568	NC_004741.1	WP_032155619.1	0	0	0	0	0	0
569	NC_004741.1	WP_001113432.1	1	1	1	1	0	80
570	NC_004741.1	WP_000703959.1	1	1	1	1	0	80
571	NC_004741.1	WP_000511292.1	0	0	0	0	0	0
572	NC_004741.1	WP_005052029.1	1	1	1	1	0	80
573	NC_004741.1	WP_000772934.1	1	0	0	0	1	40
574	NC_004741.1	WP_005052034.1	1	1	1	1	0	80
575	NC_004741.1	WP_032155621.1	0	1	0	1	0	40
576	NC_004741.1	WP_000190670.1	1	0	0	0	0	20
577	NC_004741.1	WP_045178171.1	1	1	1	1	0	80
578	NC_004741.1	WP_005052068.1	1	1	1	1	0	80
579	NC_004741.1	WP_000542440.1	1	1	1	1	0	80
580	NC_004741.1	WP_032155816.1	0	0	0	0	0	0
581	NC_004741.1	WP_011069564.1	0	0	0	0	0	0
582	NC_004741.1	WP_001304210.1	0	0	0	0	0	0
583	NC_004741.1	WP_001004881.1	0	0	0	0	0	0
584	NC_004741.1	WP_045178164.1	0	0	0	0	0	0

585	NC_004741.1	WP_001390447.1	0	0	0	0	0	0
586	NC_004741.1	WP_005052132.1	0	0	0	0	0	0
587	NC_004741.1	WP_011069562.1	1	0	0	0	0	20
588	NC_004741.1	WP_000344113.1	0	0	0	0	0	0
589	NC_004741.1	WP_000999840.1	1	0	0	0	0	20
590	NC_004741.1	WP_000924289.1	1	1	1	1	0	80
591	NC_004741.1	WP_000621323.1	1	1	1	1	0	80
592	NC_004741.1	WP_001297375.1	1	1	1	1	1	100
593	NC_004741.1	WP_000483856.1	1	1	1	1	0	80
594	NC_004741.1	WP_000665677.1	1	1	1	1	0	80
595	NC_004741.1	WP_000517100.1	0	0	0	0	0	0
596	NC_004741.1	WP_000479627.1	1	1	1	1	0	80
597	NC_004741.1	WP_000332751.1	0	0	0	0	0	0
598	NC_004741.1	WP_001331222.1	0	1	0	1	0	40
599	NC_004741.1	WP_011069558.1	1	1	1	1	0	80
600	NC_004741.1	WP_000858193.1	1	1	1	1	1	100
601	NC_004741.1	WP_001296808.1	1	1	1	1	0	80
602	NC_004741.1	WP_000576411.1	1	1	1	1	0	80
603	NC_004741.1	WP_001296791.1	1	1	1	1	1	100
604	NC_004741.1	WP_000198578.1	0	0	0	0	0	0
605	NC_004741.1	WP_032155594.1	0	0	0	0	0	0
606	NC_004741.1	WP_001063318.1	1	1	1	1	0	80
607	NC_004741.1	WP_001328969.1	0	0	0	0	0	0
608	NC_004741.1	WP_001205330.1	0	0	0	0	1	20
609	NC_004741.1	WP_000020617.1	0	0	0	0	0	0
610	NC_004741.1	WP_005064932.1	0	0	0	0	0	0
611	NC_004741.1	WP_000555608.1	0	0	0	0	0	0
612	NC_004741.1	WP_000751953.1	0	0	0	0	0	0

613	NC_004741.1	WP_011069548.1	0	0	0	0	0	0
614	NC_004741.1	WP_032155607.1	0	0	0	0	0	0
615	NC_004741.1	WP_000643692.1	1	1	1	1	0	80
616	NC_004741.1	WP_005052620.1	1	1	1	1	0	80
617	NC_004741.1	WP_000042900.1	1	1	1	1	0	80
618	NC_004741.1	WP_000778795.1	1	1	1	1	1	100
619	NC_004741.1	WP_005093467.1	0	1	0	1	0	40
620	NC_004741.1	WP_001181212.1	0	0	0	0	0	0
621	NC_004741.1	WP_000907005.1	1	0	0	1	0	40
622	NC_004741.1	WP_005097678.1	1	0	0	0	0	20
623	NC_004741.1	WP_011069541.1	1	0	0	1	0	40
624	NC_004741.1	WP_001303701.1	1	0	0	0	0	20
625	NC_004741.1	WP_005052731.1	0	0	0	0	0	0
626	NC_004741.1	WP_001055752.1	0	0	0	0	0	0
627	NC_004741.1	WP_001254807.1	1	1	1	1	0	80
628	NC_004741.1	WP_005052744.1	1	0	0	0	1	40
629	NC_004741.1	WP_000497332.1	1	1	1	1	0	80
630	NC_004741.1	WP_000847163.1	1	1	1	1	0	80
631	NC_004741.1	WP_001303699.1	0	0	0	0	0	0
632	NC_004741.1	WP_001148908.1	1	1	1	1	0	80
633	NC_004741.1	WP_000907085.1	1	1	1	1	0	80
634	NC_004741.1	WP_001007729.1	1	1	1	1	0	80
635	NC_004741.1	WP_000786137.1	0	0	0	0	0	0
636	NC_004741.1	WP_005065417.1	1	1	1	1	0	80
637	NC_004741.1	WP_001295738.1	0	0	0	0	0	0
638	NC_004741.1	WP_000062539.1	1	1	1	1	0	80
639	NC_004741.1	WP_000242065.1	1	1	0	1	0	60
640	NC_004741.1	WP_032155811.1	0	0	0	0	0	0

641	NC_004741.1	WP_000121001.1	1	1	1	1	0	80
642	NC_004741.1	WP_024259336.1	0	0	0	0	0	0
643	NC_004741.1	WP_032155640.1	0	0	0	0	0	0
644	NC_004741.1	WP_001243676.1	1	1	1	1	0	80
645	NC_004741.1	WP_001351186.1	1	0	0	0	0	20
646	NC_004741.1	WP_011110644.1	1	1	1	1	0	80
647	NC_004741.1	WP_000079652.1	1	1	1	1	0	80
648	NC_004741.1	WP_001243871.1	1	1	1	1	0	80
649	NC_004741.1	WP_011069606.1	0	0	0	0	0	0
650	NC_004741.1	WP_000166281.1	1	1	1	1	0	80
651	NC_004741.1	WP_001296688.1	0	0	0	0	0	0
652	NC_004741.1	WP_005053984.1	1	1	1	1	0	80
653	NC_004741.1	WP_000937635.1	0	0	0	0	0	0
654	NC_004741.1	WP_001119485.1	1	1	1	1	0	80
655	NC_004741.1	WP_001205243.1	1	1	1	1	1	100
656	NC_004741.1	WP_000044756.1	0	0	0	0	0	0
657	NC_004741.1	WP_001008046.1	1	1	1	1	0	80
658	NC_004741.1	WP_001238362.1	1	1	1	1	1	100
659	NC_004741.1	WP_005134385.1	1	0	0	0	0	20
660	NC_004741.1	WP_001243705.1	1	1	1	1	0	80
661	NC_004741.1	WP_000943980.1	1	1	1	1	1	100
662	NC_004741.1	WP_005053837.1	1	0	0	0	0	20
663	NC_004741.1	WP_000492914.1	1	1	1	1	0	80
664	NC_004741.1	WP_032155818.1	0	0	0	0	0	0
665	NC_004741.1	WP_000895690.1	1	0	0	0	0	20
666	NC_004741.1	WP_000132640.1	1	1	1	1	1	100
667	NC_004741.1	WP_000467859.1	1	1	1	1	0	80
668	NC_004741.1	WP_005053796.1	0	0	0	0	0	0

669	NC_004741.1	WP_000538192.1	1	1	1	1	0	80
670	NC_004741.1	WP_001338213.1	1	1	1	1	0	80
671	NC_004741.1	WP_000604352.1	1	1	1	1	0	80
672	NC_004741.1	WP_000494556.1	0	0	0	0	0	0
673	NC_004741.1	WP_000007444.1	0	0	0	0	0	0
674	NC_004741.1	WP_001303782.1	0	0	0	0	0	0

Note: 0 = 0%, 1 = 25%.

Supplementary Table 2. List of predicted physicochemical parameters of 39 hypothetical proteins

Sl. No.	Accession ID_Protein	No. of amino acids	MW	PI	Extinction coefficient	Instability index	Classification	Alphabatic index	Grand average of hydropathicity (GRAVY)
1	WP_005053355.1	274	29970.4	7.62	24325	28.39	Stable	84.01	-0.016
2	WP_000092054.1	364	40443.3	9.61	51005	47.89	Unstable	79.67	-0.384
3	WP_001382892.1	179	19590.3	5.28	2980	35.06	Stable	101.23	-0.143
4	WP_005053036.1	192	20906	9.04	7450	35.73	Stable	95.05	-0.062
5	WP_000779831.1	190	19441.2	7.87	6990	45.39	Unstable	96.58	0.172
6	WP_011110552.1	108	12039.7	7.61	8730	66.94	Unstable	77.78	-0.544
7	WP_001269672.1	193	21386.6	8.73	11460	31.1	Stable	90.98	-0.238
8	WP_001247854.1	619	69683.8	5.5	107425	33.35	Stable	78.24	-0.458
9	WP_000070107.1	377	42056.8	7.71	52035	31.62	Stable	124.14	0.611
10	WP_000224274.1	369	40593.3	7.03	40950	37.16	Stable	82.22	-0.191
11	WP_000749269.1	191	20942.5	5.57	16960	11.14	Stable	75.6	-0.436
12	WP_001125713.1	108	12371.5	9.16	7450	53.58	Unstable	82.13	-0.624
13	WP_001043881.1	165	18093.6	4.66	11585	35.37	Stable	101.09	0.133
14	WP_001295493.1	114	12493.2	4.96	20970	30.77	Stable	97.63	0.024
15	WP_000691930.1	84	8942.38	7.66	12740	41.42	Unstable	81.31	-0.167
16	WP_000597196.1	155	15601.7	9.36	2980	24.38	Stable	94.77	0.114
17	WP_000248636.1	370	39841.2	8.49	79075	32.75	Stable	143.05	1.029
18	WP_000755956.1	275	30284.6	5.68	30035	24.35	Stable	84.04	-0.331
19	WP_001237866.1	107	11755.5	6.56	6210	30.08	Stable	97.57	0.277
20	WP_000454701.1	527	59450.2	5.15	37930	32.47	Stable	116.7	0.177
21	WP_000003197.1	219	24222.7	5.17	26595	42.24	Unstable	85.98	-0.055
22	WP_005049020.1	153	16568.1	7.7	7450	23.39	Stable	82.94	-0.176
23	WP_048814497.1	243	27279.8	4.72	22585	43.49	Unstable	104.32	-0.255

24	WP_000301054.1	216	25324.2	9.31	51005	44.31	Unstable	77.64	-0.482
25	WP_000266171.1	1033	117109	5.58	244955	42.42	Unstable	83.71	-0.343
26	WP_000589825.1	160	17240.6	8.71	18450	39.1	Stable	82.44	-0.271
27	WP_005051685.1	251	26572.5	10.11	34505	23.04	Stable	76.18	-0.279
28	WP_001387238.1	158	17740.4	6.76	8480	28.12	Stable	108.04	-0.073
29	WP_000248097.1	82	9417.1	4.02	6990	44.78	Unstable	135.37	0.266
30	WP_000848528.1	85	9513.04	8.8	4720	42.38	Unstable	94.24	-0.107
31	WP_000189314.1	100	11241.9	9.84	11460	36.12	Stable	84	-0.544
32	WP_001297375.1	222	25258.4	7.84	13075	32.89	Stable	101.49	-0.086
33	WP_000858193.1	113	12552.2	9.33	14565	22.75	Stable	129.38	1.041
34	WP_001296791.1	232	25960.7	4.58	65890	36.83	Stable	74.87	-0.38
35	WP_000778795.1	127	14544.5	6.59	26470	47.26	Unstable	85.98	-0.412
36	WP_001205243.1	294	32718.5	4.96	50795	48.91	Unstable	98.91	-0.096
37	WP_001238362.1	177	19911.7	8.88	35535	27.45	Stable	78.25	-0.194
38	WP_000943980.1	387	45038.7	4.64	94685	46.38	Unstable	83.7	-0.434
39	WP_000132640.1	113	12294.1	8.64	10095	36.84	Stable	87.26	-0.168

MW, molecular weight; GRAVY, grand average of hydropathicity.

Supplementary Table 3. List of predicted sub-cellular localization of 39 hypothetical proteins

S. No.	Accession No.	Sub-cellular localization			Signal peptide (Signal P)	Secretory protein (Secretome P)	Trans membrane helices prediction		
		CELLO	PSORT B	PSLpred			HMMTOP	TMHMM	SOSUI
1	WP_005053355.1	Periplasmic	Unknown	Periplasmic	Yes	Yes	No	No	Soluble
2	WP_000092054.1	Periplasmic	Unknown	Periplasmic	Yes	Yes	1 TM Helices	No	Soluble
3	WP_001382892.1	Periplasmic/ Extracellular	Unknown	Outer membrane	Yes	Yes	No	No	Soluble
4	WP_005053036.1	Periplasmic	Unknown	Periplasmic	Yes	Yes	1 TM Helices	No	Membrane, 1 TM helix
5	WP_000779831.1	Periplasmic	Periplasmic	Periplasmic	Yes	Yes	No	No	Soluble
6	WP_011110552.1	Periplasmic	Unknown	Periplasmic	Yes	Yes	No	No	Membrane, 1 TM helix
7	WP_001269672.1	Periplasmic	Outer membrane	Periplasmic	No	Yes	2 TM Helices	No	Membrane, 1 TM helix
8	WP_001247854.1	Cytoplasmic	Cytoplasmic	Cytoplasmic	No	No	No	No	Soluble
9	WP_000070107.1	Inner membrane	Cytoplasmic membrane	Inner-membrane	No	No	6 TM Helices	6 TM Helices	membrane, 6 TM helix
10	WP_000224274.1	Periplasmic	Unknown	Cytoplasmic	No	No	No	No	Soluble
11	WP_000749269.1	Periplasmic	Unknown	Periplasmic	Yes	No	1 TM Helices	No	Soluble

12	WP_001125713.1	Cytoplasmic	Cytoplasmic	Periplasmic	No	Yes	No	No	Soluble
13	WP_001043881.1	Cytoplasmic	Unknown	Cytoplasmic	No	No	No	No	Soluble
14	WP_001295493.1	Cytoplasmic	Cytoplasmic	Cytoplasmic	No	No	No	No	Soluble
15	WP_000691930.1	Periplasmic	Unknown	Periplasmic	Yes	Yes	No	No	Membrane, 1 TM helix
16	WP_000597196.1	Extracellular	Outer membrane	Extracellular	No	Yes	No	No	Membrane, 1 TM helix
17	WP_000248636.1	Inner membrane	Cytoplasmic membrane	Inner membrane	No	No	9 TM Helices	10 TM Helices	8 TM Helices
18	WP_000755956.1	Periplasmic	Unknown	Periplasmic	No	Yes	No	No	Soluble
19	WP_001237866.1	Cytoplasmic	Unknown	Cytoplasmic	No	No	No	No	Membrane, 1 TM helix
20	WP_000454701.1	Inner membrane	Cytoplasmic membrane	InnerMembrane	No	No	7 TM Helices	7 TM Helices	Membrane, 7 TM helix
21	WP_000003197.1	Cytoplasmic	Unknown	Cytoplasmic	No	No	No	No	Soluble
22	WP_005049020.1	Periplasmic	Unknown	Periplasmic	Yes	Yes	No	No	Membrane, 1 TM helix
23	WP_048814497.1	Cytoplasmic/ Outer membrane	Extracellular	Extracellular	No	No	No	No	Soluble
24	WP_000301054.1	Cytoplasmic	Cytoplasmic	Cytoplasmic	No	No	No	No	Soluble
25	WP_000266171.1	Cytoplasmic	Unknown	Cytoplasmic	No	No	No	No	Soluble

26	WP_000589825.1	Periplasmic	Outer membrane	Outer membrane	No	Yes	No	No	Membrane, 1 TM helix
27	WP_005051685.1	Extracellular	Outer membrane	Extracellular	No	yes	No	1 TM Helices	Soluble
28	WP_001387238.1	Cytoplasmic	Cytoplasmic	Cytoplasmic	No	No	No	No	Soluble
29	WP_000248097.1	Cytoplasmic	Unknown	Cytoplasmic	No	No	No	No	Soluble
30	WP_000848528.1	Periplasmic	Outer membrane	Cytoplasmic	No	No	1 TM Helices	No	Membrane, 1 TM helix
31	WP_000189314.1	Cytoplasmic/Periplasmic	Unknown	Outer membrane	No	No	No	No	Soluble
32	WP_001297375.1	Cytoplasmic	Cytoplasmic	Cytoplasmic	No	No	No	No	Soluble
33	WP_000858193.1	Inner membrane	Cytoplasmic membrane	Inner membrane	No	No	4 TM Helices	4 TM Helices	Membrane, 3 TM helix
34	WP_001296791.1	Extracellular	Extracellular	Extracellular	No	Yes	1 TM Helices	1 TM Helices	Soluble
35	WP_000778795.1	Cytoplasmic	Unknown	Cytoplasmic	No	No	No	No	Soluble
36	WP_001205243.1	Cytoplasmic	Extracellular	Cytoplasmic	No	No	No	1 TM Helices	Soluble
37	WP_001238362.1	Periplasmic	Outer membrane	Cytoplasmic	Yes	Yes	No	1 TM Helices	Membrane, 1 TM helix
38	WP_000943980.1	Cytoplasmic	Cytoplasmic	Cytoplasmic	No	No	No	No	Soluble
39	WP_000132640.1	Cytoplasmic	Unknown	Cytoplasmic	No	yes	No	No	Soluble

Supplementary Table 4. List of annotated functions of 39 hypothetical proteins from *Shigella flexneri* using CDD-BLAST, Pfam, HmmScan, SMART, Scanprosite, PS2-v2, and STRING

Sl. No	Acc ID	Functional domain (BLAST,Pfam, HmmScan, SMART,Scanprosite)	Templates	Domain in (PS)2-v2	Predicted functional partner (STRING)
1	WP_005053355.1	Peptidase, C92 family	No template	Error	Minor fimbrial subunit, D-mannose specific adhesin
2	WP_000092054.1	DUF1615/Lipoprotein	1m9iA	Same	Microcin B17 transporter
3	WP_001382892.1	DUF3251/lipoprotein YajI/immunoglobulin like domain	2jwyA	Same	Hypothetical protein SF0234/ATP synthase
4	WP_005053036.1	Lipoprotein_16/Uncharacterized lipoprotein	2iqiF	Same	Regulatory protein AmpE
5	WP_000779831.1	Lipoprotein chaperone (YscW)	No template	Error	Universal stress protein UspB
6	WP_011110552.1	YbfN-like lipoprotein	No template	Error	Hypothetical protein ybfM
7	WP_001269672.1	LPS-assembly lipoprotein RlpB (LptE)	2r76A	Same	LPS assembly outer membrane complex protein LptD
8	WP_001247854.1	Topoisomerases, DnaG-type primases, Hedgehog/Intein domain	No template	Error	DNA-directed RNA polymerase subunit beta
9	WP_000070107.1	ATP-binding cassette transporter	2dyrA	**OXIDOREDUCTASE**	ATP-binding protein ybhF_2
10	WP_000224274.1	MOSC beta barrel domain/2Fe-2S iron-sulfur cluster binding domain	2piaA	Same	Fe/S biogenesis protein NfuA
11	WP_000749269.1	YceI-like domain	1y0gA	Same	yceJ Cytochrome
12	WP_001125713.1	YcgL domain	2h7aA	Same	Hypothetical protein ycgN

13	WP_001043881.1	GAF domain	1vhmB	Same	Hypothetical protein; RNA chaperone proQ
14	WP_001295493.1	Endoribonuclease L-PSP/YjgFfamily	1qd9A	Same	D-amino acid dehydrogenase small subunit
15	WP_000691930.1	Domain of unknown function (DUF333)	2pqcA	**TRANSFERASE**	Hypothetical protein yeaP
16	WP_000597196.1	Glycine zipper 2TM domain	No template	Error	Flagellar fliJ protein
17	WP_000248636.1	AI-2E family transporter/permease	2jlnA	Same	Glutamine amidotransferase/anthranilate phosphoribosyltransferase
18	WP_000755956.1	SPFH domain / Band 7 family	3bk6A	Same	Integrase
19	WP_001237866.1	YecR-like lipoprotein	No template	Error	Glycosyl transferase
20	WP_000454701.1	TerC family/Transporter associated domain/CBS domain	2yvyA	Same	Glutamate synthase
21	WP_000003197.1	von Willebrand factor type A domain	1atzB	Same	Chaperonin
22	WP_005049020.1	Uncharacterized lipoprotein YehR	2joeA	Same	Transporter
23	WP_048814497.1	Leucine rich repeat protein/NEL or novel E3 ligase domain	3cvrA	Same	Aerobic respiration control sensor protein ArcB
24	WP_000301054.1	Lipopolysaccharide kinase (Kdo/WaaP)	1blxA	Same	Lipopolysaccharide core heptose(I) kinase RfaP
25	WP_000266171.1	Tetratricopeptide repeat (TPR)	No template	Error	NAGC-like transcriptional regulator
26	WP_000589825.1	Outer membrane protein (ompA) like domain/membrane lipoprotein	2k1sA	Same	Hypothetical protein SF2663

27	WP_005051685.1	LysM (lysin-like motif)/ Peptidase family M23	2gu1A	Same	Beta-hexosaminidase
28	WP_001387238.1	DNA repair protein RadC-like JAB domain	2qlcA	Same	Hypothetical protein SF2995
29	WP_000248097.1	Carrier protein (CP) domain and phosphopantetheine attachment site	1x3oA	Same	Class II aminotransferase
30	WP_000848528.1	Lipoprotein leucine-zipper	1jcdB	Same	Porin
31	WP_000189314.1	GIY-YIG nuclease domain	1zg2A	Same	Hypothetical protein yhbP
32	WP_001297375.1	DNA repair protein RadC-like JAB domain	No template	Error	DNA mismatch repair protein MutS;
33	WP_000858193.1	yiaA/B two helix domain	1oedA	Same	Hypothetical protein yiaA
34	WP_001296791.1	Autotransporter beta-domain	No template	Error	Biofilm formation regulatory protein BssR
35	WP_000778795.1	Acetyltransferase (GNAT) domain	2k5tA	Same	Aspartate alpha-decarboxylase
36	WP_001205243.1	Xylose isomerase-like TIM barrel (AP_endonuc_2)	1k77A	Same	3-ketoacyl-ACP reductase
37	WP_001238362.1	Lipocalin-like domain	1qwdA	Same	Sugar nucleotide epimerase
38	WP_000943980.1	Glutathionylspermidine synthase	No template	Error	Nicotinate phosphoribosyltransferase
39	WP_000132640.1	Toxin SymE/SpoVT-AbrB domain	1ve0A	Same	Hypothetical protein SF1670

Note: Proteins with discrepant results are shown in bold.

Supplementary Table 5. List of annotated functions of 39 hypothetical proteins from *Shigella flexneri* using MOTIF, Interproscan, CATH, SUPERFAMILY, and ProtoNet

Sl No.	Acc ID	MOTIF	INTERPROSCAN	CATH	SUPERFAMILY	ProtoNet
1	WP_005053355.1	Papain-like amidase enzyme, YaeF/YiiX, C92 family	Papain-like amidase enzyme, YaeF/YiiX, C92 family	Lipoprotein/Uncharacterized protein	Cysteine proteinases YiiX-like	Cluster 3674930 Proteobacteria
2	WP_000092054.1	Protein of unknown function (DUF1615)	Protein of unknown function DUF1615	No hit	GFP-like	Cluster 4109548 Protein of unknown function DUF1615
3	WP_001382892.1	Protein of unknown function (DUF3251)	Domain of unknown function DUF3251	Hypothetical lipoprotein yajI	Phase 1 flagellin	Cluster 3711586 2JWY
4	WP_005053036.1	Uncharacterized lipoprotein	Uncharacterised protein family, YajG	No hit	Phase 1 flagellin	Cluster 4028813 Uncharacterized lipoprotein
5	WP_000779831.1	lipoprotein chaperone (YscW)	No result	No hit	No result	Cluster 4131069 Proteobacteria
6	WP_011110552.1	YbfN-like lipoprotein	YbfN-like lipoprotein	No hit	No result	Cluster 4085534 Lipoprotein
7	WP_001269672.1	Lipopolysaccharide-assembly LptE	LPS-assembly lipoprotein LptE	LPS-assembly lipoprotein LptE	No result	Cluster 3965977 Rare lipoprotein B
8	WP_001247854.1	Toprim-like	DNA primase/Toprim domain	DNA primase/helicase	DNA primase/helicasecore	Cluster 3410389 DNA helicase, DnaB-like
9	WP_00007010	ABC-2 family	ABC-2 transporter	membrane transport	MFS general	Cluster 4114591

	7.1	transporter protein		permease YbhS/ATP-binding protein	substrate transporter	ABC-2
10	WP_000224274.1	MOSC domain/ 2Fe-2S iron-sulfur cluster	MOSC, N-terminal beta barrel	MOSC domain/ 2Fe-2S iron-sulfur cluster	MOSC N-terminal domain-like	Cluster 4155424 MOSC, N-terminal beta barrel
11	WP_000749269.1	YceI-like domain	YceI-like domain	YceI-like domain	YceI-like domain	Cluster 4314345 YceI-like
12	WP_001125713.1	YcgL domain	YcgL domain	No hit	YcgL-like	Cluster 4083593 YcgL domain
13	WP_001043881.1	GAF domain	GAF domain	GAF domain	GAF domain	Cluster 4085038 GAF
14	WP_001295493.1	Endoribonuclease L-PSP	YjgF/L-PSP	Endoribonuclease L-PSP family	YjgF/L-PSP	Cluster 4054994 YjgF/L-PSP
15	WP_000691930.1	Domain of unknown function (DUF333)	Domain of unknown function (DUF333)	No hit	No result	Cluster 4079210 Protein of unknown function DUF333
16	WP_000597196.1	Glycine zipper 2TM domain	Glycine zipper 2TM domain	No hit	No result	Cluster 4073223 Glycine zipper 2TM domain
17	WP_000248636.1	AI-2E family transporter	AI-2E family transporter	No hit	No result	Cluster 4074174 AI-2E family transporter
18	WP_000755956.1	SPFH domain / Band 7 family	SPFH domain / Band 7 family	No hit	Band 7/SPFH domain	Cluster 3793474 Band 7/SPFH domain
19	WP_001237866.1	YecR-like lipoprotein	YecR-like lipoprotein	No hit	SRCR-like	Cluster 3743846 Enterobacteriales
20	WP_000454701.1	TerC family/Transporter	TerC family/Transporter	TerC family/Transporter	CBS-domain pair/transporter-	Cluster 4019065 membrane protein

		associated domain/CBS domain	associated domain/CBS domain	associated domain/CBS domain	associated domain	TerC
21	WP_000003197.1	von Willebrand factor type A domain	TerY/vWA-like	von Willebrand factor type A domain	vWA-like	Cluster 4002958 TerY/vWA-like
22	WP_005049020.1	Protein of unknown function (DUF1307)/YehR-like	Protein of unknown function (DUF1307)/YehR-like	Putative lipoprotein YehR	YehR-like	Cluster 4046970 Protein of unknown function (DUF1307)/YehR-like
23	WP_048814497.1	Leucine Rich repeats	LRR-containing bacterial E3 ligase	leucine rich repeat protein/E3 ligase domain	Leucine Rich repeats	Cluster 4154032 Protein binding
24	WP_000301054.1	Lipopolysaccharide kinase (Kdo/WaaP)	Lipopolysaccharide kinase	Lipopolysaccharide kinase (Kdo/WaaP)	Lipopolysaccharide kinase (Kdo/WaaP)	Cluster 3990101 Lipopolysaccharide kinase
25	WP_000266171.1	Tetratricopeptide repeat	Tetratricopeptide-like domain	TPR repeat-containing protein	TPR-like	Cluster 4040666 Tetratricopeptide region
26	WP_000589825.1	OmpA family	OmpA-like domain	outer membrane lipoprotein	OmpA-like	Cluster 4198784 Outer membrane protein
27	WP_005051685.1	Peptidase family M23/LysM domain	Peptidase M23/LysM domain	Peptidase M23	Peptidoglycan hydrolase LytM	Cluster 4141397 Peptidase M23B
28	WP_001387238.1	RadC-like JAB domain	RadC protein	DNA repair protein RadC	JAB1/MPN domain	Cluster 4114260 RadC protein
29	WP_000248097.1	Phosphopantetheine attachment site	Phosphopantetheine binding ACP	Acyl carrier protein	Acyl-carrier protein (ACP)	Cluster 4146821 Phosphopantetheine

						-binding
30	WP_000848528.1	Lipoprotein leucine-zipper	Murein-lipoprotein	Major outer membrane lipoprotein	Outer membrane lipoprotein	Cluster 4066376 Murein-lipoprotein
31	WP_000189314.1	GIY-YIG catalytic domain	GIY-YIG endonuclease	No hit	GIY-YIG endonuclease	Cluster 4157077 GIY-YIG endonuclease
32	WP_001297375.1	RadC-like JAB domain	RadC-like JAB domain	DNA repair protein RadC	RuvA domain 2-like/JAB1/MPN domain	Cluster 4403730 RadC-like JAB domain
33	WP_000858193.1	yiaA/B two helix domain	YiaAB two helix	No hit	No result	Cluster 4440457 YiaAB two helix
34	WP_001296791.1	Autotransporter beta-domain	Autotransporter, YhjY	No hit	Autotransporter	Cluster 3853611 Autotransporter, YhjY
35	WP_000778795.1	Acetyltransferase (GNAT) domain	Acyl-CoA N-acyltransferase	Putative N-acetyltransferase	Acyl-CoA N-acyltransferases (Nat)	Cluster 4355736 N-acetyltransferase activity
36	WP_001205243.1	Xylose isomerase-like TIM barrel	Xylose isomerase-like TIM barrel	Putative hydroxypyruvate isomerase	Xylose isomerase-like	Cluster 4100779 Xylose isomerase-type TIM barrel
37	WP_001238362.1	Lipocalin-like domain	Lipocalin, ApoD type	Outer membrane lipoprotein Blc	Lipocalins	Cluster 4145424 Lipocalin
38	WP_000943980.1	Glutathionylspermidine synthase preATP-grasp	Glutathionylspermidine synthase, pre-ATP-grasp	No hit	Glutathione synthetase ATP-binding domain-like	Cluster 4243753 Glutathionylspermidine synthase
39	WP_000132640.1	Toxin SymE, type I toxin-antitoxin system	Type I toxin-antitoxin system, SymE toxin	No hit	No result	Cluster 4040297 Toxin SymE-like

Supplementary Table 6. List of annotated functions of 25 proteins with known function from *Shigella flexneri* using BLAST, Pfam, HmmScan, SMART, and Scanprosite for receiver operating characteristic analysis

Sl no	Acc ID protein	Protein name	BLAST	Pfam	HmmScan	SMART	Scanprosite
1	WP_000241642.1	Homoserine kinase	Homoserine kinase 1 (5)	Homoserine kinase 1 (5)	Homoserine kinase 1 (5)	Homoserine kinase 1 (5)	Homoserine kinase 1 (5)
2	WP_000809168.1	Protein hokC	Protein hokC 1 (5)	Protein hokC 1 (5)	Protein hokC 1 (5)	Protein hokC 1 (5)	Protein hokC 1 (5)
3	WP_001286897.1	Isoleucine--tRNA ligase	Isoleucine--tRNA ligase 1 (5)	Isoleucine--tRNA ligase 1 (5)	Isoleucine--tRNA ligase 1 (5)	Isoleucine--tRNA ligase 1 (5)	Isoleucine--tRNA ligase 1 (5)
4	WP_000124415.1	Ferrichrome porin FhuA	Ferrichrome outer membrane transporter 1 (4)	TonB dependent receptor 0 (2)	TonB dependent receptor 0 (2)	TonB dependent receptor 0 (2)	TonB dependent receptor 0 (2)
5	WP_001183183.1	MFS transporter	MFS transporter 1 (5)	MFS transporter 1 (5)	MFS transporter 1 (5)	MFS transporter 1 (5)	MFS transporter 1 (5)
6	WP_001230481.1	Ail/Lom family protein	OM_channels super family 1 (3)	Ail/Lom protein 1 (5)	Ail/Lom protein 1 (5)	Ail/Lom protein 1 (5)	Virulence outer membrane protein 1 (3)
7	WP_001287126.1	Glutamine--tRNA ligase	Glutaminyl-tRNA synthetase 1 (5)	tRNA synthetases 1 (5)	tRNA synthetases 1 (5)	tRNA synthetases 1 (5)	Aminoacyl-transfer RNA synthetases 1 (5)
8	WP_001295442.1	Flagellar L-ring protein	Flagellar basal body L-	Flagellar L-ring protein 1	Flagellar L-ring protein 1	Flagellar L-ring protein 1	PROKAR_LIPOPROTEIN 0 (3)

			ring protein 1 (4)	(5)	(5)	(5)	
9	WP_000130034.1	D-alanine--D-alanine ligase	D-alanine--D-alanine ligase 1 (5)	D-alanine--D-alanine ligase 1 (5)	D-alanine--D-alanine ligase 1 (5)	D-alanine--D-alanine ligase 1 (5)	D-alanine--D-alanine ligase 1 (5)
10	WP_000197853.1	Alanine racemase	Alanine racemase 1 (5)	Alanine racemase 1 (5)	Alanine racemase 1 (5)	Alanine racemase 1 (5)	Alanine racemase 1 (5)
11	WP_000569431.1	Ribonuclease HII	Ribonuclease HII 1 (5)	Ribonuclease HII 1 (5)	Ribonuclease HII 1 (5)	Ribonuclease HII 1 (5)	No hit 0 (5)
12	WP_000901098.1	VOC family protein	VOC family protein 1 (5)	VOC family protein 1 (5)	VOC family protein 1 (5)	VOC family protein 1 (5)	VOC family protein 1 (5)
13	WP_001260712.1	Proline--tRNA ligase	Proline--tRNA ligase 1 (5)	Proline--tRNA ligase 1 (5)	Proline--tRNA ligase 1 (5)	Proline--tRNA ligase 1 (5)	Proline--tRNA ligase 1 (5)
14	WP_000051892.1	Integrase	Integrase 1 (5)	Integrase 1 (5)	Integrase 1 (5)	Integrase 1 (5)	No hit 0 (5)
15	WP_001120449.1	Oxidoreductase	Oxidoreductase 1 (5)	Oxidoreductase 1 (5)	Oxidoreductase 1 (5)	Oxidoreductase 1 (5)	Oxidoreductase 1 (5)
16	WP_000460136.1	LysR family transcriptional regulator	LysR family transcriptional regulator 1 (5)	LysR family transcriptional regulator 1 (5)	LysR family transcriptional regulator 1 (5)	LysR family transcriptional regulator 1 (5)	LysR family transcriptional regulator 1 (5)
17	WP_001018618.1	Flavodoxin-1	Flavodoxin-1 1 (5)	Flavodoxin-1 1 (5)	Flavodoxin-1 1 (5)	Flavodoxin-1 1 (5)	Flavodoxin-1 1 (5)
18	WP_000773301.1	Acyl-CoA esterase	Acyl-CoA esterase 1 (5)	Acyl-CoA esterase 1 (5)	Acyl-CoA esterase 1 (5)	Acyl-CoA esterase 1 (5)	No hit 0 (5)
19	WP_000201488.1	DNA-packaging	DNA-packaging	DNA-packaging	DNA-packaging	DNA-packaging	No hit 0 (5)

		protein FI	protein FI 1 (5)	protein FI 1 (5)	protein FI 1 (5)	protein FI 1 (5)	
20	WP_005049594.1	Terminase	Terminase 1 (5)	Terminase 1 (5)	Terminase 1 (5)	Terminase 1 (5)	No hit 0 (5)
21	WP_000537402.1	Thioredoxin-disulfide reductase	Thioredoxin-disulfide reductase 1 (5)	Thioredoxin-disulfide reductase 1 (5)	Thioredoxin-disulfide reductase 1 (5)	Thioredoxin-disulfide reductase 1 (5)	Thioredoxin-disulfide reductase 1 (5)
22	WP_000109301.1	MFS transporter	MFS transporter 1 (5)	MFS transporter 1 (5)	MFS transporter 1 (5)	MFS transporter 1 (5)	MFS transporter 1 (5)
23	WP_005047463.1	Porin OmpA	Porin OmpA 1 (5)	Porin OmpA 1 (5)	Porin OmpA 1 (5)	Porin OmpA 1 (5)	Porin OmpA 1 (5)
24	WP_001247604.1	YjbF family lipoprotein	YjbF family lipoprotein 1 (5)	YjbF family lipoprotein 1 (5)	YjbF family lipoprotein 1 (5)	YjbF family lipoprotein 1 (5)	YjbF family lipoprotein 1 (5)
25	WP_014532269.1	DUF333 domain-containing protein	Domain of unknown function (DUF333) 1 (5)	Domain of unknown function (DUF333) 1 (5)	Domain of unknown function (DUF333) 1 (5)	Domain of unknown function (DUF333) 1 (5)	PROKAR_LIPOPROTEIN 0 (3)

True positive and true negative are denoted by “1” and “0”.

Integers in parentheses denote the confidence level.

Supplementary Table 7. List of annotated functions of 25 proteins with known function from *Shigella flexneri* using MOTIF, Interproscan, CATH, SUPERFAMILY, and ProtoNet for receiver operating characteristic analysis

Sl No.	Acc ID protein	Protein name	MOTIF	INTERPROSCAN	CATH	SUPERFAMILY	ProtoNet
1	WP_000241642.1	Homoserine kinase	Homoserine kinase 1 (5)	Homoserine kinase 1 (5)	Homoserine kinase 1 (5)	Homoserine kinase 1 (5)	Homoserine kinase 1 (5)
2	WP_000809168.1	Protein hokC	hok_gef 1 (3)	hok_gef 1 (3)	0 (5)	hok_gef 0 (5)	hok_gef 1 (3)
3	WP_001286897.1	Isoleucine--tRNA ligase	Isoleucine--tRNA ligase 1 (5)	Isoleucine--tRNA ligase 1 (5)	Isoleucine--tRNA ligase 1 (5)	Isoleucine--tRNA ligase 1 (5)	Isoleucine--tRNA ligase 1 (5)
4	WP_000124415.1	Ferrichrome porin FhuA	TonB-dependent Receptor 0 (2)	TonB-dependent siderophore receptor 0 (2)	TonB-dependent siderophore receptor 0 (2)	TonB-dependent siderophore receptor 0 (2)	TonB-dependent siderophore receptor 0 (2)
5	WP_001183183.1	MFS transporter	MFS transporter 1 (5)	MFS transporter 1 (5)	MFS transporter 1 (5)	MFS transporter 1 (5)	MFS transporter 1 (5)
6	WP_001230481.1	Ail/Lom family protein	Ail/Lom family protein 1 (5)	Ail/Lom family protein 1 (5)	Ail/Lom family protein 1 (5)	Ail/Lom family protein 1 (5)	Ail/Lom family protein 1 (5)
7	WP_001287126.1	Glutamine--tRNA ligase	Glutamine--tRNA ligase 1 (5)	Glutamine--tRNA ligase 1 (5)	Glutamine--tRNA ligase 1 (5)	Glutamine--tRNA ligase 1 (5)	Glutamine--tRNA ligase 1 (5)
8	WP_001295442.1	Flagellar L-ring protein	Flagellar L-ring protein 1 (5)	Flagellar L-ring protein 1 (5)	0 (5)	Flagellar L-ring protein 0 (5)	Flagellar L-ring protein 1 (5)

9	WP_000130034.1	D-alanine--D-alanine ligase	D-alanine--D-alanine ligase 1 (5)	D-alanine--D-alanine ligase 1 (5)	D-alanine--D-alanine ligase 1 (5)	D-alanine--D-alanine ligase 1 (5)	D-alanine--D-alanine ligase 1 (5)
10	WP_000197853.1	Alanine racemase	Alanine racemase 1 (5)	Alanine racemase 1 (5)	Alanine racemase 1 (5)	Alanine racemase 1 (5)	Alanine racemase 1 (5)
11	WP_000569431.1	Ribonuclease HII	Ribonuclease HII 1 (5)	Ribonuclease HII 1 (5)	Ribonuclease HII 1 (5)	Ribonuclease HII 1 (5	Ribonuclease HII 1 (5)
12	WP_000901098.1	VOC family protein	VOC family protein 1 (5)	VOC family protein 1 (5)	VOC family protein 1 (5)	VOC family protein 1 (5)	VOC family protein 1 (5)
13	WP_001260712.1	Proline--tRNA ligase	Proline--tRNA ligase 1 (5)	Proline--tRNA ligase 1 (5)	Proline--tRNA ligase 1 (5)	Proline--tRNA ligase 1 (5)	Proline--tRNA ligase 1 (5)
14	WP_000051892.1	Integrase	Integrase 1 (5)	Integrase 1 (5)	Integrase 1 (5)	Integrase 1 (5)	Integrase 1 (5)
15	WP_001120449.1	Oxidoreductase	Oxidoreductase 1 (5)	Oxidoreductase 1 (5)	Oxidoreductase 1 (5)	Oxidoreductase 1 (5)	Oxidoreductase 1 (5)
16	WP_000460136.1	LysR family transcriptional regulator	LysR family transcriptional regulator 1 (5)	LysR family transcriptional regulator 1 (5)	LysR family transcriptional regulator 1 (5)	LysR family transcriptional regulator 1 (5)	LysR family transcriptional regulator 1 (5)
17	WP_001018618.1	Flavodoxin-1	Flavodoxin_1, 3, 4, 5 1 (3)	Flavodoxin, long chain 1 (3)	Short-chain flavodoxin YkuP 1 (3)	Flavodoxin, long chain 1 (3)	Flavodoxin, long chain 1 (3)
18	WP_000773301.1	Acyl-CoA esterase	Alpha/beta hydrolase fold 1 (3)	Alpha/beta hydrolase fold, alpha/beta hydrolase fold, 1 1 (3)	Esterase YbfF 1 (5)	Alpha/beta hydrolase fold, alpha/beta hydrolase fold, 1 (3)	Alpha/beta hydrolase fold-1 1 (3)
19	WP_000201488.1	DNA-packaging	DNA-packaging	DNA-packaging protein FI 1 (5)	0 (5)	DNA-packaging protein FI 0 (5)	DNA-packaging

		protein FI	protein FI 1 (5)				protein FI 1 (5)
20	WP_005049594.1	Terminase	Phage terminase large subunit (GpA) 1 (3)	Bacteriophage lambda, GpA 1 (3)	0 (5)	Bacteriophage lambda, GpA 0 (5)	Phage terminase GpA 1 (3)
21	WP_000537402.1	Thioredoxin-disulfide reductase	Pyridine nucleotide-disulphide oxidoreductase 1 (3)	Pyridine nucleotide-disulphide oxidoreductase, class-II 1 (4)	Thioredoxin reductase 1 (4)	Pyridine nucleotide-disulphide oxidoreductase, class-II 1 (3)	Thioredoxin-disulfide reductase 1 (5)
22	WP_000109301.1	MFS transporter	MFS transporter 1 (5)	MFS transporter 1 (5)	MFS transporter 1 (5)	MFS transporter 1 (5)	MFS transporter 1 (5)
23	WP_005047463.1	Porin OmpA	Porin OmpA 1 (5)	Porin OmpA 1 (5)	Porin OmpA 1 (5)	Porin OmpA 1 (5)	Porin OmpA 1 (5)
24	WP_001247604.1	YjbF family lipoprotein	YjbF family lipoprotein 1 (5)	YjbF family lipoprotein 1 (5)	YjbF family lipoprotein 1 (5)	YjbF family lipoprotein 1 (5) YjbF family lipoprotein 1 (5)	YjbF family lipoprotein 1 (5)
25	WP_014532269.1	DUF333 domain-containing protein	DUF333 domain-containing protein 1 (5)	DUF333 domain-containing protein 1 (5)	0 (5)	DUF333 domain-containing protein 1 (5)	DUF333 domain-containing protein 1 (5)

True positive and true negative are denoted by “1” and “0.”

Integers in parentheses denote the confidence level.

Supplementary Table 8. List of predicted virulence factors of 39 hypothetical proteins by using VICMPred and Virulentpred

Sl. No.	Acc ID_Protein	VICMpred	Virulentpred
1	WP_005053355.1	Cellular process	Virulent
2	WP_000092054.1	Cellular process	Virulent
3	WP_001382892.1	Information and storage	Virulent
4	WP_005053036.1	Cellular process	Virulent
5	WP_000779831.1	Cellular process	Virulent
6	WP_011110552.1	Information and storage	Virulent
7	WP_001269672.1	Metabolism Molecule	Virulent
8	WP_001247854.1	Virulence factors	Non-Virulent
9	WP_000070107.1	Metabolism Molecule	Non-Virulent
10	WP_000224274.1	Cellular process	Non-Virulent
11	WP_000749269.1	Virulence factors	Virulent
12	WP_001125713.1	Cellular process	Virulent
13	WP_001043881.1	Cellular process	Non-Virulent
14	WP_001295493.1	Metabolism Molecule	Non-Virulent
15	WP_000691930.1	Cellular process	Virulent
16	WP_000597196.1	Metabolism Molecule	Virulent
17	WP_000248636.1	Metabolism Molecule	Non-Virulent
18	WP_000755956.1	Metabolism Molecule	Non-Virulent
19	WP_001237866.1	Cellular process	Virulent
20	WP_000454701.1	Metabolism Molecule	Non-Virulent
21	WP_000003197.1	Cellular process	Virulent
22	WP_005049020.1	Cellular process	Non-Virulent
23	WP_048814497.1	Cellular process	Virulent
24	WP_000301054.1	Metabolism Molecule	Non-Virulent
25	WP_000266171.1	Metabolism Molecule	Non-Virulent
26	WP_000589825.1	Cellular process	Virulent
27	WP_005051685.1	Cellular process	Virulent
28	WP_001387238.1	Cellular process	Virulent
29	WP_000248097.1	Cellular process	Virulent
30	WP_000848528.1	Cellular process	Virulent
31	WP_000189314.1	Cellular process	Virulent
32	WP_001297375.1	Metabolism Molecule	Non-Virulent
33	WP_000858193.1	Cellular process	Non-Virulent
34	WP_001296791.1	Cellular process	Virulent
35	WP_000778795.1	Cellular process	Virulent

36	WP_001205243.1	Cellular process	Non-Virulent
37	WP_001238362.1	Cellular process	Non-Virulent
38	WP_000943980.1	Cellular process	Non-Virulent
39	WP_000132640.1	Cellular process	Non-Virulent