Analysis and construction of pathogenicity island regulatory pathways in *Salmonella enterica* serovar Typhi

Su Yean Ong[1 *], Fui Ling Ng[1], Siti Suriawati Badai[1], Anton Yuryev[2], Maqsudul Alam[1]

[1] Centre for Chemical Biology, Universiti Sains Malaysia, 1st Floor, Block B, No. 10, Persiaran Bukit Jambul, 11900 Bayan Lepas, Pulau Pinang, Malaysia

[2] Ariadne Genomics Inc., 9430 Key West Avenue, Suite 113, Rockville, MD 20850, USA

suyeanccb@gmail.com, fuilingccb@gmail.com, suria.brilliant@gmail.com, ayuryev@ariadnegenomics.com, maqsudul@gmail.com

Summary

Signal transduction through protein-protein interactions and protein modifications are the main mechanisms controlling many biological processes. Here we described the implementation of MedScan information extraction technology and Pathway Studio software (Ariadne Genomics Inc.) to create a *Salmonella* specific molecular interaction database. Using the database, we have constructed several signal transduction pathways in *Salmonella enterica* serovar Typhi which causes Typhoid Fever, a major health threat especially in developing countries. *S.* Typhi has several pathogenicity islands that control rapid switching between different phenotypes including adhesion and colonization, invasion, intracellular survival, proliferation, and biofilm formation in response to environmental changes. Understanding of the detailed mechanism for *S.* Typhi survival in host cells is necessary for development of efficient detection and treatment of this pathogen. The constructed pathways were validated using publically available gene expression microarray data for *Salmonella*.

1 Introduction

S. Typhi is able to survive a variety of harsh conditions and defense mechanisms existing in the human gastrointestinal tract. Multiple survival strategies allow *S.* Typhi to cause epidemic outbreaks of typhoid fever in many developing countries. Therefore, *Salmonella* represents a major health risk according to the World Health Organization (WHO) [1]. Propagation of *S.* Typhi infection is due to its ability to enter a dormant state by forming biofilm in the human gallbladder (typhoid carriers), enabling it to evade the immune system [2]. Typhoid carriers do not show any symptoms, and are the only reservoir for *S.* Typhi which is transmitted via contaminated food or water. Existing diagnostic tools cannot detect *S.* Typhi in typhoid carriers.

Different bacterial species use similar infection strategies due to the acquisition of diverse pathogenicity islands. Similar pathogenicity islands are found in both Gram-positive and Gram-negative bacteria. They represent a distinct class of genomic regions which is acquired through horizontal gene transfer. To get classified as a pathogenicity island, a region should carry genes encoding one or more virulence factors such as adhesins, toxins, and invasins. Pathogenicity islands are located on the bacterial chromosome or on a plasmid and carry functional genes for DNA recombination such as integrase, transposase, or part of an

* To whom correspondence should be addressed.

insertion element. The G+C content of the pathogenicity island differs from the rest of the genome. They represent an unstable DNA region which may move from one tRNA locus to another or get deleted [3]. Most of the pathogenicity islands have pseudogenes that are defunct relatives of known genes which have lost their protein-coding ability or are no longer expressed in the cell. Nevertheless, most pseudogenes have recognizable gene-like features. Therefore, they share functional ancestry with a functional gene and contain biological and evolutionary histories within their sequences [4]. In this study, we analyzed the molecular interaction network enabling global transcriptional regulation of *Salmonella* pathogenicity genes using *in silico* approach and have constructed nine SPI pathways responsible for different stages of *S.* Typhi infection including host invasion, intracellular host survival, and drug resistance. Protein activity in *Salmonella* is regulated by various environmental factors. Comprehensive studies of this regulation can facilitate the discovery of key protein players in pathogenic bacteria. Reconstruction of *Salmonella* pathogenicity pathways also allows compiling the comprehensive list of candidate biomarkers expressed during the infection that can be further used for development of new typhoid diagnostics. Pathogenicity pathways can also be used for interpretation of new experimental data and for comparison of different *Salmonella* strains with respect to the infection mechanism. Pathway Studio software from Ariadne Genomics was used for network analysis and pathway construction as well as for analysis of gene expression microarray data. The resulting networks and pathways from this work are publicly available for download from http://www.ccbusm.com.

2 Methodology

2.1 Construction of Biological Associations Database for *Salmonella*

We used Pathway Studio software (Ariadne Genomics Inc.) to construct *S.* Typhi pathogenicity islands regulatory pathways. Pathway Studio software allows automatic extraction of regulatory and physical interactions from MEDLINE abstracts using natural language processing technology called MedScan [5]. Interactions extracted by MedScan which contain a formalized set of relationships are imported into the Pathway Studio database and analyzed further using data-mining tools for knowledge inference and pathway reconstruction available in Pathway Studio [6]. Since MedScan keeps the reference of the original article containing a statement about the extracted interaction, it also helps to perform quick assertion of extracted facts and identification of relevant publications. Thus, it facilitates our pathway reconstruction by selecting the appropriate interactions.

Using protein names dictionaries for *Salmonella,* we processed more than 70,000 PubMed abstracts and more than 15,000 full-length articles containing the keywords *Salmonella* including *S.* Typhi. This yielded the database of more than 10,000 relationships reported for *Salmonella* proteins that included information about physical and regulatory interactions between *Salmonella* proteins and metabolites as well as regulatory interactions between proteins and cell processes. All found interactions used for pathway construction in *S.* Typhi were manually curated and only validated interactions were included in the pathways.

2.2 Prediction of interactions for *Salmonella* from other bacterial species

To further facilitate pathway construction we used interactions from Pathway Studio Bacterial database described previously [7]. It allowed us to predict interactions between *Salmonella* proteins based on interactions reported in other bacterial species. The approach to predict interactions between orthologs in different species is called interolog annotation [8]. Orthologs for *Salmonella* proteins in other bacterial organisms were predicted using the best

reciprocal hit method from full length protein sequence similarities calculated from BLAST alignments as described previously [9]. The Bacterial database contains molecular interactions extracted by MedScan for all bacterial species from over 1,000,000 PubMed abstracts annotated with Medical Subject Headings (MeSH) term "Bacteria" and from more than 74,000 full-length articles from 22 microbiology journals. Proteins in the Bacterial database are annotated with Entrez Gene and GenBank identifiers from 32 bacterial genomes, including *S.* Typhi Ty2, *S.* Typhi CT18, and *S. typhimurium* LT2. Additional annotation from 716 partial genomes was obtained from the NCBI Protein Clusters database. The database allows quick identification of interactions reported for different bacterial organisms that can be relevant for pathway construction in *S.* Typhi. All interactions extracted for *Salmonella* orthologs were imported into the Pathway Studio *Salmonella* database for pathway building and network analysis of the gene expression data. All interologs used for pathway reconstruction in *S.* Typhi were manually curated. Only validated interactions were included in pathways.

2.3 Construction of pathways controlling expression of SPIs

The first step in pathway building was identification of proteins in the database encoded by each SPI in *S.* Typhi Ty2 or *S.* Typhi CT18. A simple search for the proteins with corresponding Entrez Gene ID was performed in Pathway Studio database. Entrez Gene IDs for SPI proteins were obtained by exploring the *S.* Typhi CT18 genome (GenBank accession number NC_003198) in NCBI sequence viewer. Once SPI proteins were identified, we connected them with either physical interactions or expression regulatory relations found in the Bacterial and *Salmonella* databases. We then expanded the pathways by adding all known transcriptional regulators for SPI proteins. We also added autophosphokinases that regulate the activity of transcriptional factors in two-component relay signaling system. Next, we added environmental signals that are sensed by two-component regulatory systems. Finally, we added human proteins that are known to interact with *S.* Typhi effectors.

We manually verified each interaction used for pathway construction by reading the original article and making substantiative assertion to validate the interaction. MedScan classifies extracted relations using only the information available in the sentence describing the extracted fact. Therefore, we manually converted all regulatory relations classified as *Expresssion* by MedScan into *PromoterBinding* if the regulation has been described as a direct interaction elsewhere in the text. Some indirect regulatory interactions were explained by connecting several intermediate proteins into a path consisting of consecutive direct physical interactions. Lastly, we excluded the redundant interactions that were extracted by MedScan from our pathways.

2.4 Network analysis of gene expression microarray data

Gene expression omnibus (GEO, NCBI) dataset GSE3096 was used for network analysis. GSE3096 measures *S.* Typhi gene expression during the infection of human macrophages (THP-1) [10]. We used Sub-network Enrichment Analysis (SNEA) algorithm [11] with option "Expression targets" available in Pathway Studio to identify significant transcription factors regulating most differentially expressed genes. If a gene was measured by multiple probes on the array only probe with best p-value was used for SNEA. All relationships used to identify major regulators were manually verified after the initial analysis, false positives were removed from the database and SNEA was run for second time to verify again the significance of transcription factors.

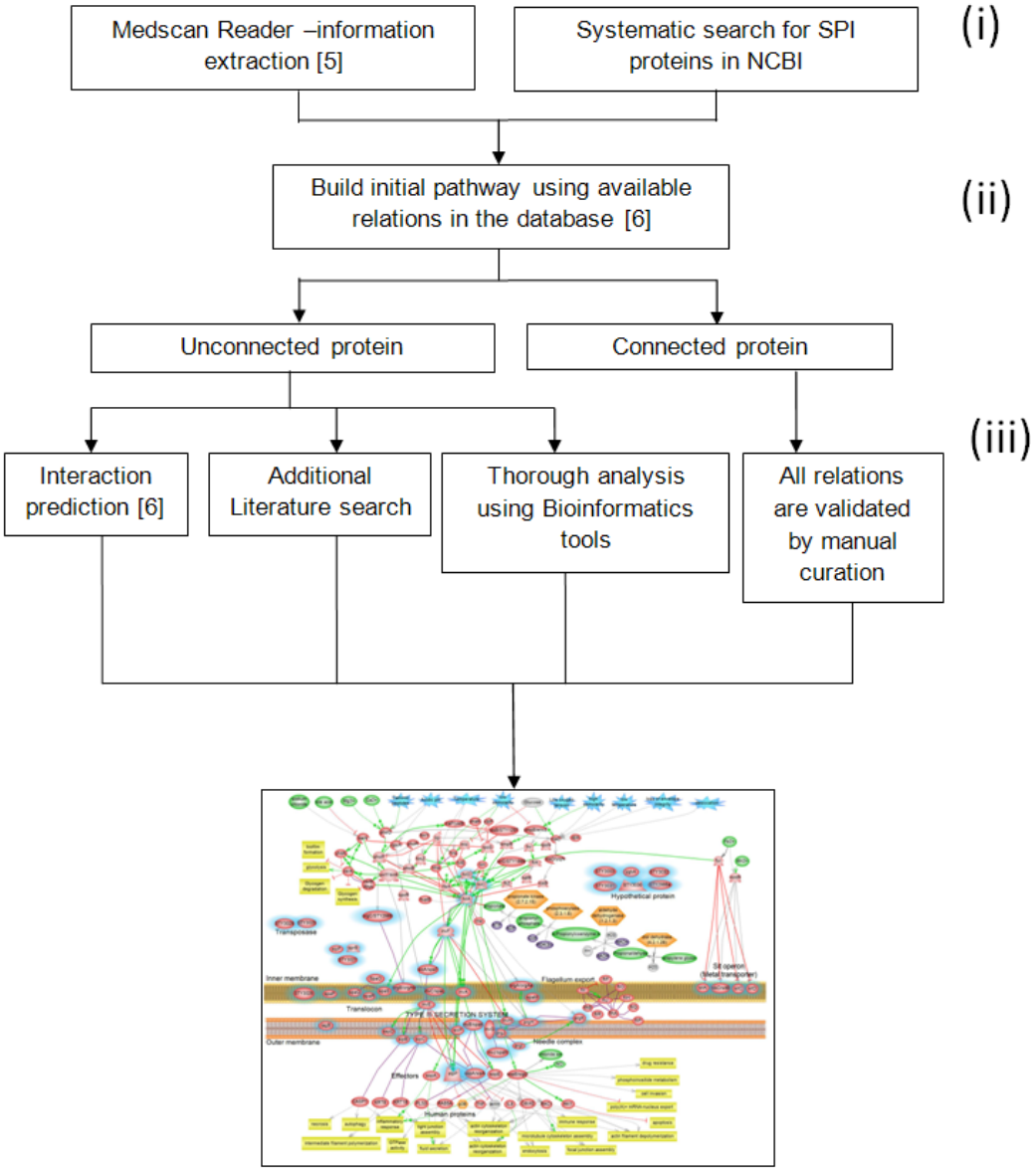


Figure 1: SPI proteins were searched in *Salmonella enterica subsp. enterica* serovar Typhi str. CT18 genome (GenBank accession NC_003198). Literature search was done through http://www.ncbi.nlm.nih.gov/pubmed. The bioinformatics tools that were used are accessible through http://blast.ncbi.nlm.nih.gov/, http://www.ncbi.nlm.nih.gov/projects/gorf, http://www.ebi.ac.uk/Tools/ClustalW2, http://www.ebi.ac.uk/Tools/Interproscan, http://www.ebi.ac.uk/Tools/emboss/align/index.html, http://pfam.sanger.ac.uk/.

2.5 Identification of gene expression clusters in SPI pathways

Genes from each SPI regulatory pathway were clustered by correlation network algorithm available in Pathway Studio under "Predict network from expression" menu using expression profiles from GSE3096. "Predict network from expression" command calculates Pearson correlation between each pair of genes and creates gene correlation network where correlation links are above user-defined threshold. We used the correlation threshold of 0.95 (95%) to identify gene clusters. Only genes with positive correlation were then selected for figures and for analysis of upstream transcription factors.

3 Results

3.1 Construction and validation of pathways controlling expression of pathogenicity islands

3.1.1 Salmonella Pathogenicity Island 1 (SPI-1)

SPI-1 encodes 48 genes including type III secretion system (T3SS-1) for invasion of epithelial cells (Figure 2). Most SPI-1 genes are regulated by several two-component systems which respond to different environmental signals. The reconstructed SPI-1 pathway supports previous suggestions that all environmental signals converge into the HilD-HilC-RtsA system and is then further transmitted by the HilA-InvF transcription factors to activate expression of effector genes encoded in SPI-1 by direct binding of their promoters. The signals dispersed by HilA and InvF towards the downstream effectors enable *S.* Typhi invasion of the host cell. SPI-1 also encodes the Fe^{2+} and Mn^{2+} uptake system (*sit* operon) that is required during the later stage of infection [12-15]. Among all environmental signals, only propionate indirectly represses HilA activity while other signals activate HilA.

Eleven proteins in SPI-1 are annotated as pseudogenes or as hypothetical proteins. We have reanalyzed their sequences using BLAST to reaffirm their function. We found that *sty3025* and *sty3029*, which are annotated as pseudogenes, have high similarity to transposase. Also, the major portion of *sty3027*, annotated as hypothetical protein, was found to be similar to the acetyltransferase (GNAT) family.

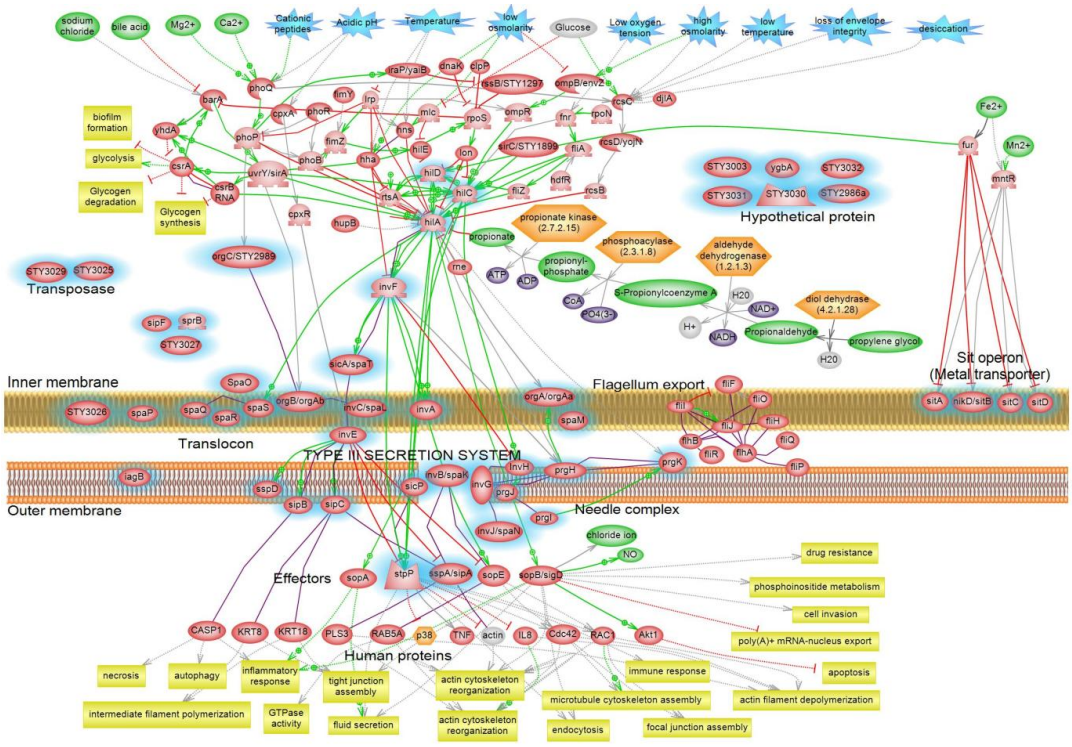


Figure 2: SPI-1 regulation pathway. Proteins encoded by SPI-1 are highlighted in blue. SPI-1 encodes for T3SS which is important for *Salmonella* invasion of the host cell. The central regulator of SPI-1 expression is HilA transcription factor. A detailed view of the SPI-1 pathway including supporting literature is available at http://www.ccbusm.com/publications/spi/SPI-1.html.

3.1.2 Salmonella Pathogenicity Island 2 (SPI-2)

SPI-2 consists of 45 genes that are required for survival of *S.* Typhi inside phagosomes (Figure 3). OmpR activates SPI-2 genes by binding to the promoter of the *ssrAB* operon to induce expression of SsrA and SsrB proteins [16]. The OmpB-OmpR two-component system is activated by low osmolarity while the PhoQ-PhoP system, which also regulates SPI-2 genes, senses the acidity of the environment inside the phagosomes. Expression of SPI-2 genes is mainly regulated through the SsrA-SsrB two-component system. Many secreted effector proteins are located at different *Salmonella* loci but are translocated via the T3SS system encoded by SPI-2 (eg: PipA and PipB from SPI-5). SPI-2 also contains the *ttrRSBCA* operon which encodes tetrathionate reductase. Although TtrB, TtrC, and TtrA are not involved in virulence, they are essential for anaerobic respiration [12, 17, 18]. According to [17], the ability to respire tetrathionate is likely to be significant within the life cycle of *Salmonella*. This ability is a characteristic of only certain genera of Enterobacteriacea including *Salmonella, Citrobacter, and Proteus* [19]. Further in the text we demonstrate that low oxygen serves as a main trigger for activation of SPI-1 invasion genes during macrophage invasion. Hence, expression of tetrathionate reductase during SPI-2 activation may be used to promote *Salmonella* survival inside the host cell.

Ten SPI-2 genes were reblasted to confirm their identity and function. Analysis of BLAST results shows that the major portion of pseudogene *sty1739* is highly similar to DeoR family transcriptional regulator and pseudogene *sty1742* is similar to proline iminopeptidase, suggesting that these genes are functional as both were expressed in the microarray experiment.

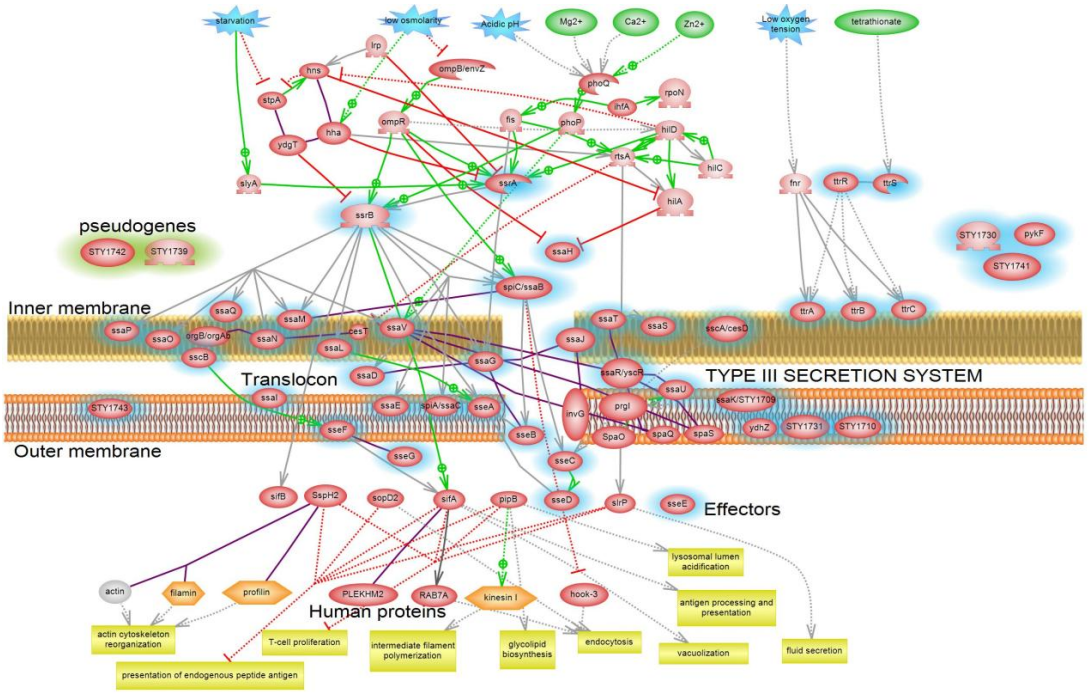


Figure 3: SPI-2 regulation pathway. SPI-2 encodes for T3SS and the expression of genes is governed by OmpB-OmpR and SsrA-SsrB. Proteins encoded by SPI-2 are highlighted in blue. A detailed view of the SPI-2 pathway including supporting literature is available at http://www.ccbusm.com/publications/spi/SPI-2.html.

3.1.3 Salmonella Pathogenicity Island 3 (SPI-3)

SPI-3 genes were shown to be important for *S.* Typhi survival inside the host cells. Existing literature indicates that SPI-3 consists of fourteen genes including six pseudogenes (Figure 4). We found only five proteins in SPI-3 annotated with known functions: RmbA, SlsA, MgtA, STY4022 (MgtB), and STY4023 (MgtC). Expression of MgtA is dependent on RpoS, RcsC-YojN-RcsB, and PhoP. MgtA, MgtC, and MgtB function in high-affinity Mg^{2+} uptake. The ability to survive in Mg^{2+} limitation is necessary for *S.* Typhi virulence [20]. Nine other proteins encoded by SPI-3 were reblasted to refine their functional annotation available in *S.* Typhi CT18 genome. We found that *sty4030* encodes a full length homolog of *S. typhimurium* MisL (an autotransporter) which serves as an intestinal colonization factor that binds to human fibronectin [21]. *sty4024* was similar to CigR from *S. typhimurium*, and *sty4027* was similar to *S. typhimurium* putative transcriptional regulator MarT. Surprisingly, *sty4030*, *sty4024*, and *sty4027* are annotated as pseudogenes in the *S.* Typhi CT18 genome [22] but the microarray data shows that these genes are expressed during macrophage infection.

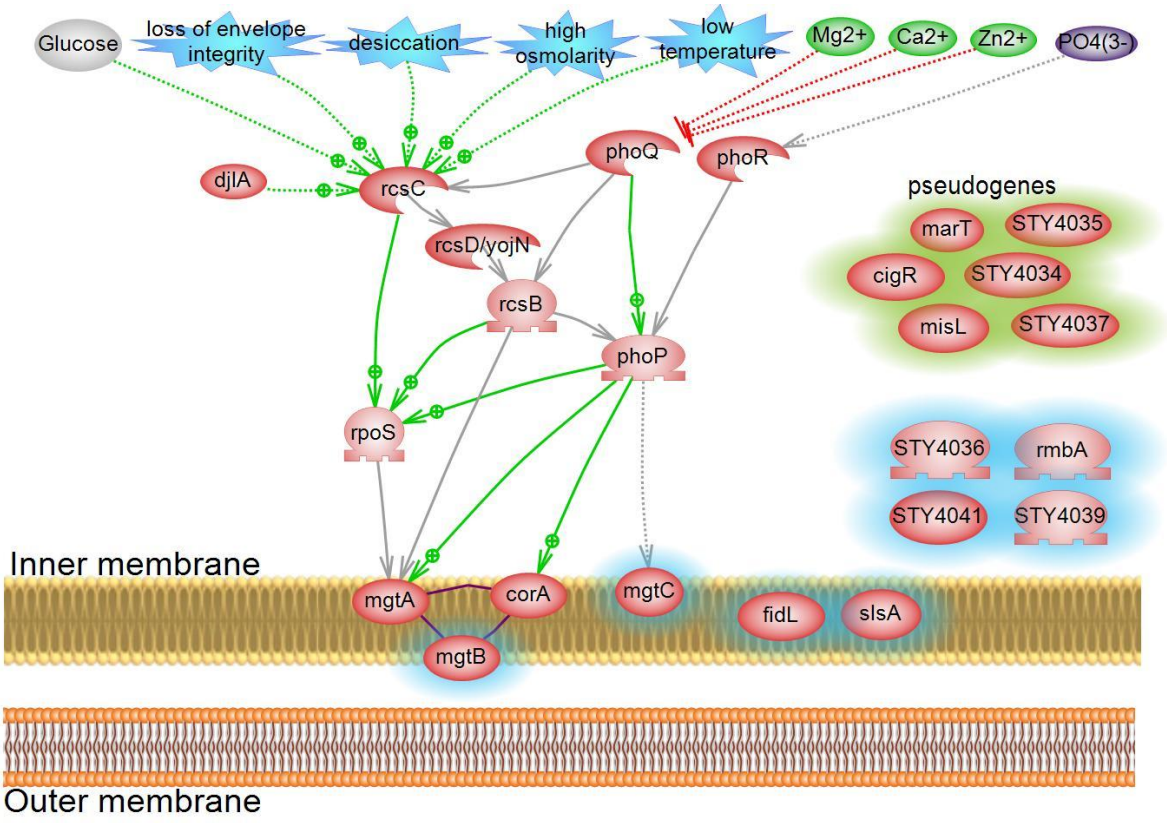


Figure 4: SPI-3 regulation pathway. SPI-3 encodes MgtB, and MgtC which are responsible for Mg^{2}+ uptake. Most SPI-3 proteins remain unconnected. Proteins encoded by SPI-3 are highlighted in blue. A detailed view of the SPI-3 pathway including supporting literature is available at http://www.ccbusm.com/publications/spi/SPI-3.html.

3.1.4 Salmonella Pathogenicity Island 4 (SPI-4)

SPI-4 has 7 genes, regulated by the same regulatory network as SPI-1 genes (Figure 5). SPI-1 was shown to be required for the activation of SPI-4 [23], which further supports our SPI-4 pathway. In addition to the SPI-1 regulators SirA, HilA, and H-NS, expression of SPI-4 genes is also regulated by RfaH. RfaH is an anti-termination factor preventing premature termination of transcription in SPI-4 [23]. The organization of SPI-4 genes in *S.* Typhi is similar to the s*iiABCDEF* operon in *S. typhimurium*. *sty4456* (*siiC*), *sty4457* (*siiD*), and *sty4460* (*siiF*) encode a type I secretion system (T1SS) necessary for the secretion of *siiE* [24]. SiiE is a large repetitive protein that functions as a nonfimbrial adhesin in binding to epithelial cell surfaces [25]. Unlike in *S. typhimurium*, *siiE* in *S.* Typhi is encoded by two orfs, *sty4458* and *sty4459*. Our sequence analysis suggested that *sty4458* and *sty4459* were not pseudogenes as reported previously [22]. Besides both genes being similar to *siiE* from *S. typhimurium*, microarray data also confirms that *siiE* is expressed in *S.* Typhi [10].

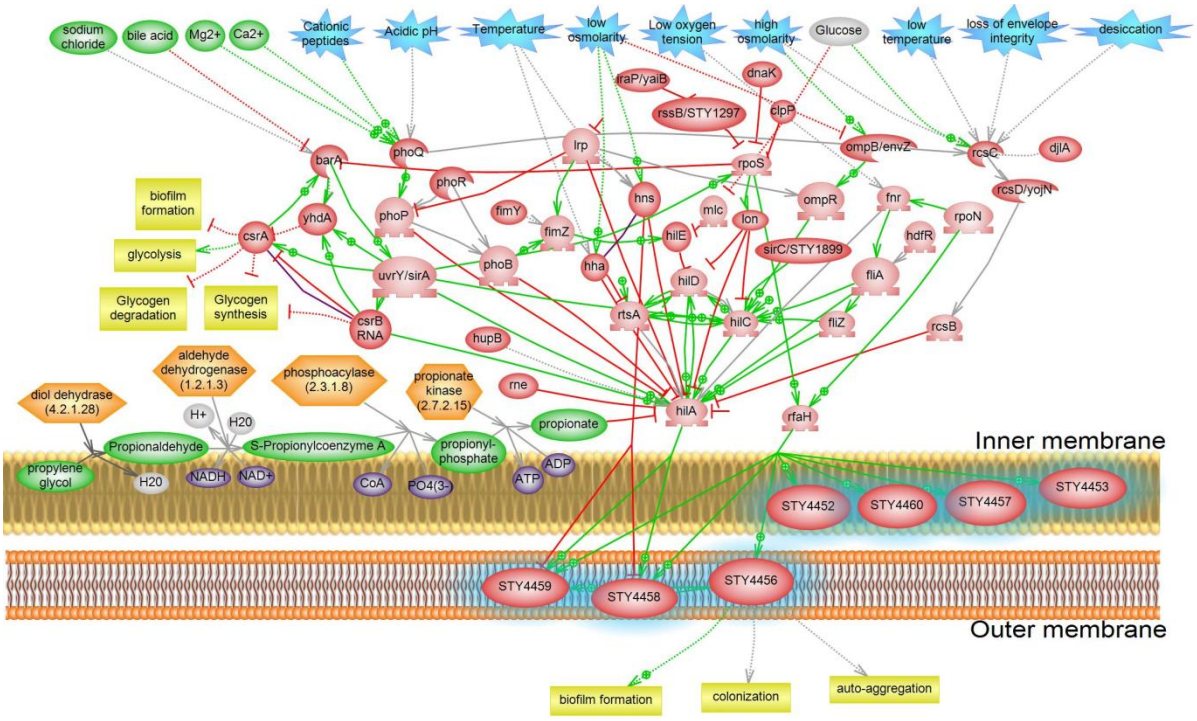


Figure 5: SPI-4 regulation pathway. SPI-4 encodes for T1SS and the proteins are mainly regulated by HilA and RfaH. Proteins encoded by SPI-4 are highlighted in blue. A detailed view of SPI-4 pathway including supporting literature is available at http://www.ccbusm.com/publications/spi/SPI-4.html.

3.1.5 Salmonella Pathogenicity Island 5 (SPI-5)

SPI-5 is a 7.6 kb region encoding 8 genes: *pipD*, *sigD/sopB*, *sigE*, *pipA*, *pipB*, and three transposases (*sty1124*, *tnpA*, and *sty1125*) (Figure 6). The genes are controlled by the SPI-1 and SPI-2 regulatory circuits and are known to contribute to *Salmonella* enteropathogenesis

[12, 26]. SopB is secreted through the T3SS encoded by SPI-1, while PipA and PipB are secreted through the T3SS encoded by SPI-2. Expression of PipA and PipB is regulated by the EnvZ/OmpR two-component regulatory system. SigE is a molecular chaperone which is important for the stabilization and secretion of SopB/SigD [27]. SigD/SopB is a secreted inositol phosphatase that triggers fluid secretion responsible for diarrhea [26]. It activates mammalian protooncogene Akt, a serine threonine kinase responsible for inhibition of

apoptosis in normal intestinal epithelial cells during the infection [28]. *pipD* encodes a cysteine protease homolog which is crucial in contributing to long-term systemic infection [29].

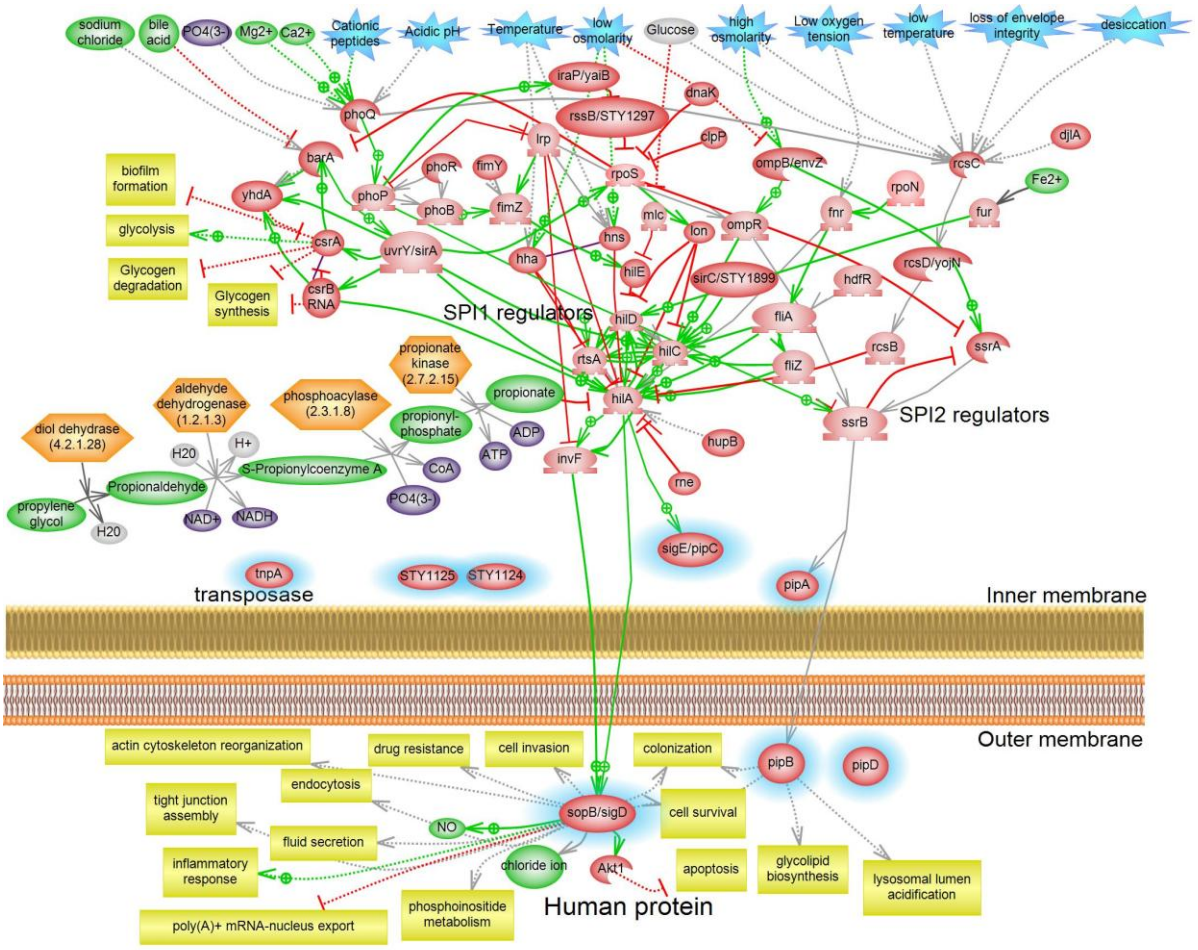


Figure 6: SPI-5 regulation pathway. SigD/SopB, PipA, and PipB contribute to enteropathogenesis, which triggers fluid secretion responsible for diarrhea. Proteins encoded by SPI-5 are highlighted in blue. A detailed view of the SPI-5 pathway including supporting literature is available at http://www.ccbusm.com/publications/spi/SPI-5.html.

3.1.6 Salmonella Pathogenicity Island 6 (SPI-6)

SPI-6 encodes 59 genes (Figure 7). The function and regulation of SPI-6 genes is still largely unknown and they are not annotated in GenBank. Therefore, we performed additional sequence analysis for SPI-6 genes. We found that SciN, SciP, SciS, SciK, VapD, VgrS, SciF/ImpF, and SciQ are homologous to the type VI secretion system (T6SS) machinery identified in *V. cholerae* [30]. The *Saf* operon (*safA*, *safB*, *safC*, and *safD*) and *tcf* operon (*tcfA*, *tcfB*, *tcfC*, and *tcfD*) are fimbrial usher proteins. Twenty proteins are identified as cytoplasmic proteins, two proteins as integral membrane proteins, two proteins as periplasmic proteins, and four proteins as transposases. After our sequence analysis there are still fifteen genes left as hypothetical with no homology to proteins with known function. The complete results of our analysis are shown in Table 1.

Table 1: List of proteins encoded by SPI-6. Most of the proteins are not connected and thorough bioinformatics analyses of these proteins were carried out.

Protein	Description
STY0286	SciA, ImpA-related N-family protein
STY0287	SciA, ImpA-related N-family protein
STY0288	SciB, type VI secretion protein
STY0289	SciC, type VI secretion protein
STY0290	SciD, type VI secretion protein lysozyme-related protein
STY0291	SciE, predicted virulence protein
STY0292	SciF, replication/virulence associated protein
STY0293	Tetratricopeptide repeat family protein
STY0294	ClpB protein
STY0295	Hypothetical protein
STY0296	Hypothetical protein
STY0297	SciH, type VI secretion protein
STY0298	SciI, type VI secretion protein
STY0300	Invasol SirA
STY0301	SciJ protein (Precursor)
STY0302	SciM, hemolysin-coregulated protein
STY0303	SciN, type VI secretion lipoprotein
STY0304	SciO, type VI secretion protein
STY0305	SciP, type VI secretion protein
STY0306	SciQ, putative membrane protein
STY0307	Hypothetical protein
STY0308	SciS, type VI secretion protein
STY0310	SciT, replication/virulence associated protein
STY0311	Mannosyl-glycoprotein endo-beta-N-acetylglucosamidase
STY0312	Hypothetical protein
STY0313	Hypothetical protein
STY0314	Hypothetical protein
STY0316	Hypothetical protein
STY0317	Putative cytoplasmic protein
STY0318	Hypothetical protein
STY0319	Rhs-family protein
STY0320	Putative cytoplasmic protein
STY0321	Rhs1 protein
STY0322	Hypothetical protein
STY0323	Hypothetical protein
STY0324	Rhs-family protein (cell envelope biogenesis, outer membrane)
STY0326	FhaB (filamentous hemagglutinin) protein
STY0327	Hypothetical protein
STY0328	yjiW; endoribonuclease SymE
STY0329	Transposase B
STY0338	Periplasmic binding protein, Ybe-J like protein
STY0339	Transposase
STY0342	Hypothetical protein
STY0343	Transposase
STY0344	IstB transposition protein
STY0350	TioA protein
STY0351	SapA-like protein
STY0352	VirG-like protein

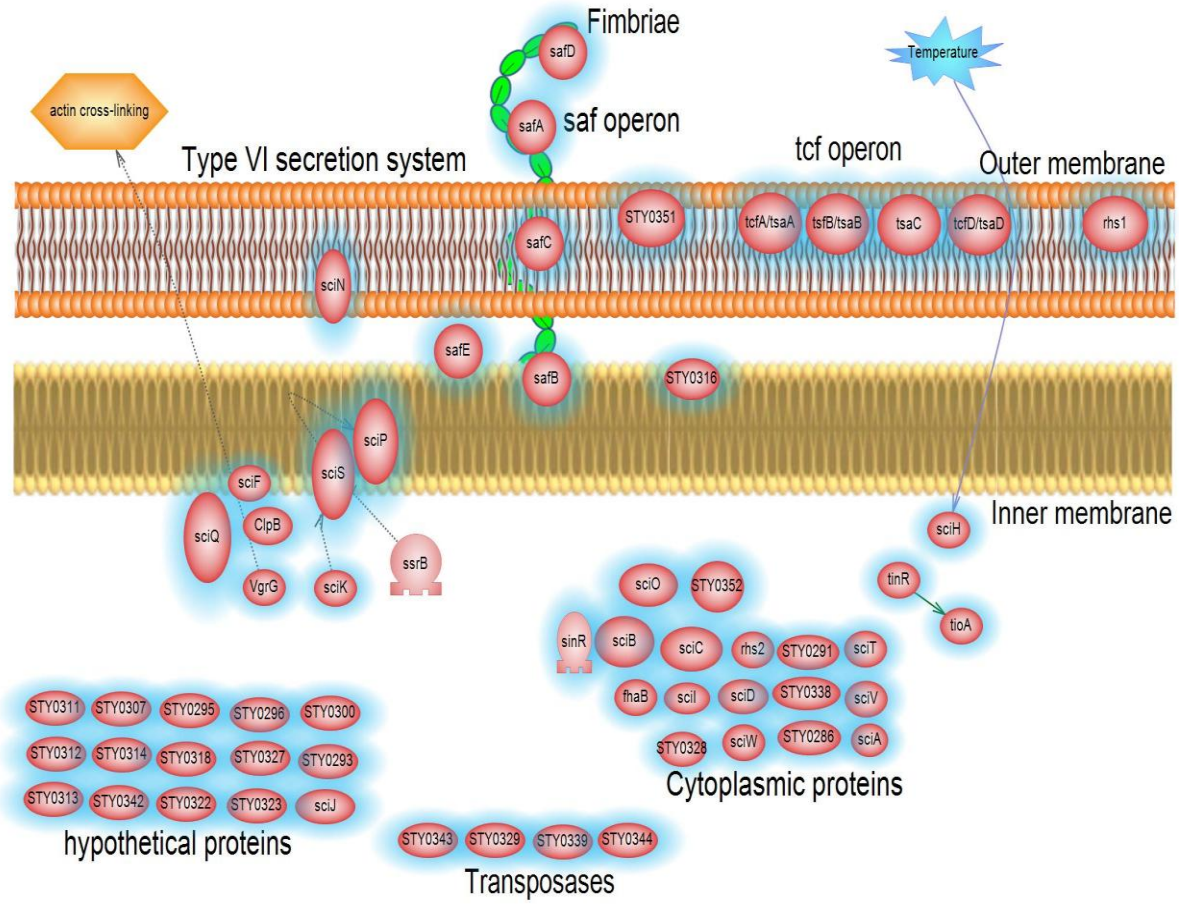


Figure 7: SPI-6 proteins shown in pathway diagram form. The function and regulation of the genes encoded in SPI-6 are still mainly unknown. Here we show only sub-cellular localization and function predicted for 44 genes from SPI-6 revealed by our sequence analysis. A detailed view of the SPI-5 pathway including supporting literature is available at http://www.ccbusm.com/publications/spi/SPI-6.html.

3.1.7 Salmonella Pathogenicity Island 7 (SPI-7)

The **SPI-7** region is unique to *S.* Typhi. It consists of 148 genes (Figure 8), encoding a prophage and genes for virulence factors such as Vi antigen (ten genes), SopE effector, and type IV pili (fifteen genes) [31]. The production of Vi antigen is governed by the two-component systems EnvZ-OmpR and RcsC-RcsB (Figure 8). The TviA regulator encoded by SPI-7 interacts with transcription factor RcsB to promote transcription of Vi antigen genes [32]. Interestingly, the same system also controls the *pil* operon (type IV pili) [32]. Meanwhile, effector protein SopE is translocated through the T3SS of SPI-1. 80 out of 148 proteins were classified as either hypothetical proteins or proteins with unknown function. We performed an extensive sequence analysis using bioinformatics tools to assign predicted functions to these proteins. We found that thirteen are related to prophage, another thirteen are related to DNA recombination, and three are similar to transporters. The remaining proteins are assigned with different functions associated with prophage biology (Table 2).

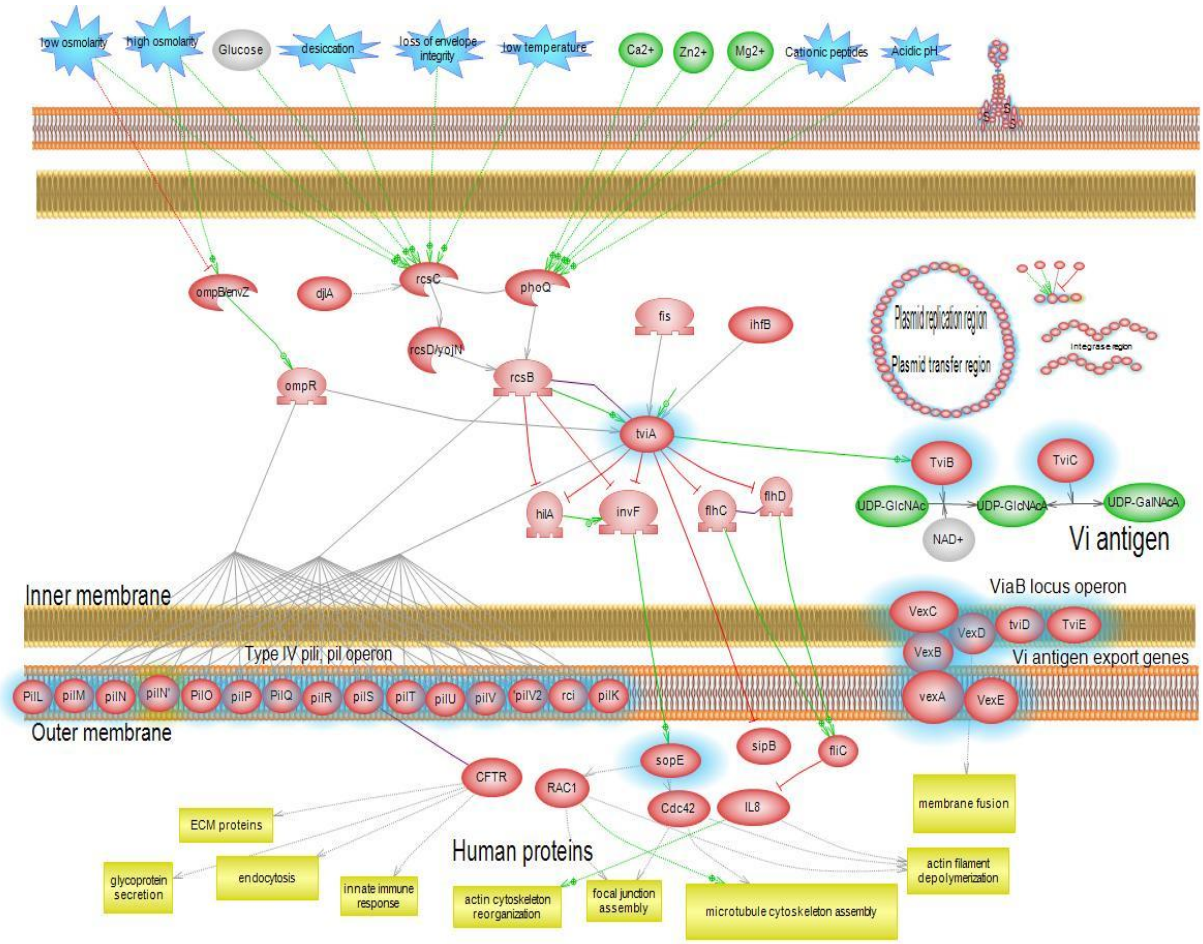


Figure 8: SPI-7 regulation pathway. SPI-7 carries genes for potential virulence factors such as Vi antigen, SopE, and type IV pili. Proteins encoded by SPI-7 are highlighted in blue. A detailed view of the SPI-7 pathway including supporting literature is available at A detailed view of the SPI-5 pathway including supporting literature is available at http://www.ccbusm.com/publications/spi/SPI-7.html.

Table 2: List of proteins in SPI-7. 80 out of 148 proteins were analyzed using bioinformatics tools in order to assign predicted functions to these proteins which are largely unconnected to one another.

Protein	Description
STY4523	ParB
STY4524	Transcriptional regulator, CdaR
STY4525	Putative phage associated protein
STY4526	Type I restriction enzyme restriction subunit
STY4528	Two component CheB methylesterase
STY4529	Exodeoxyribonuclease V, 135 kDa subunit
STY4534	DNA polymerase III, epsilon subunit
STY4535	Hypothetical protein

STY4537	ISNCY family transposase
STY4539	PilL protein
STY4541	PilN
STY4546	PilR protein
STY4553	Polyribonucleotide nucleotidyltransferase
STY4554	TraE
STY4557	RND family efflux transporter MFP subunit
STY4558	Plasma-membrane proton-efflux P-type ATPase
STY4560	50S ribosomal protein L25/general stress protein Ctc
STY4563	TraD
STY4564	Type III effector Hop protein
STY4565	Phage integrase family site specific recombinase
STY4566	Membrane protein
STY4568	DDE superfamily endonuclease containing protein
STY4569	Type II and III secretion system protein
STY4570	TraB pilus assembly family protein
STY4572	Type IV secretory pathway, VirB4 component
STY4574	Capsular polysaccharide biosynthesis glycosyl transferase
STY4575	Multi-sensor hybrid histidine kinase
STY4576	Ribonuclease E (rne)
STY4577	COG2805: Tfp pilus assembly protein, pilus retraction ATPase PilT
STY4578	DNA repair and recombination protein RAD26
STY4579	Membrane protein
STY4580	Multidrug resistance protein 2
STY4582	Phage tail tape measure protein, TP901 family
STY4584	Transcriptional regulator IbrB
STY4585	4-hydroxybenzoate decarboxylase, subunit D
STY4587	Aminotransferase, class V
STY4588	Acetate--CoA ligase
STY4589	Sensor histidine kinase
STY4590	Retrotransposon hot spot (RHS) protein
STY4591	Type I site-specific restriction-modification system, R subunit
STY4593	Pseudouridine synthase
STY4594	Carboxyl-terminal protease
STY4595	D-alanyl-D-alanine carboxypeptidase/D-alanyl-D-alanine-endopeptidase
STY4596	ABC-2 type transporter (Precursor)
STY4599	Major facilitator superfamily protein
STY4602	Phage P2 GpU family protein
STY4605	Phage tail protein E
STY4608	DNA-invertase
STY4611	Phage tail fibre protein
STY4612	Phage tail protein I
STY4613	Phage baseplate assembly protein
STY4614	Phage baseplate assembly protein

STY4615	Phage baseplate assembly protein V
STY4616	Phage virion morphogenesis protein
STY4617	P2 phage tail completion protein R
STY4618	Phage lysis regulatory protein, LysB family
STY4619	LysA protein
STY4622	Phage Tail Protein X
STY4624	Terminase, endonuclease subunit
STY4629	Transcription-repair coupling factor
STY4630	3'-5' exoribonuclease, RNase R/RNase II family
STY4631	Hypothetical protein
STY4632	COG4226: Uncharacterized protein encoded in hypervariable junctions of pilus gene clusters
STY4633	DinI family protein
STY4636	DNA adenine methylase
STY4637	Exonuclease
STY4641	Conserved hypothetical protein fil of phage origin
STY4643	Phage regulatory protein
STY4647	Autoinducer 2-binding protein lsrB (AI-2-binding protein lsrB) (Precursor)
STY4648	Protein YjhX 2
STY4663	Cupin 2 domain-containing protein
STY4666	Phage integrase
STY4667	CopG-like DNA-binding
STY4669	MutT-like protein
STY4670	Glucosamine-6-phosphate deaminase-like protein
STY4671	PhiRv2 prophage protein
STY4672	Glutamate decarboxylase
STY4674	Hypothetical protein
STY4675	Short chain dehydrogenase/reductase family oxidoreductase
STY4677	Hypothetical protein STY4677
STY4679	SH3, type 3

3.1.8 Salmonella Pathogenicity Island 8 (SPI-8)

SPI-8 encodes 16 genes. No interactions among proteins encoded by SPI-8 are described in the published literature. We found by sequence analysis that *sty3280-sty3283* encode colicin/pyocin, and *sty3274* and *sty3277* encode for type VI secretion system (T6SS). The functions of the remaining ten proteins remain unknown. At the early stage of infection, *S.* Typhi may use pyocin to kill other bacteria in the intestine in order to compete for nutrients. T6SS is used by *S.* Typhi as a secretion machine to deliver proteins and toxins into the eukaryotic target cell. This is crucial for virulence and survival within the host cells [30].

3.1.9 Salmonella Pathogenicity Island 9 (SPI-9)

SPI-9 has 4 genes, *oprJ*, *prtC*, *prtB*, and *amyH* (Figure 9), which are involved in type I secretion systems (T1SS) [22]. Our sequence alignment analysis found that OprJ (STY2876) has high similarity with TolC, a component of AcrAB, which pumps out bile acids, antibiotics, dyes, and disinfectants [33]. PrtC (STY2878) and PrtB (STY2877) have high similarity with HlyD and HlyB respectively, which are exporters for repeats in toxin (RTX

toxin) proteins [34]. AmyH is a homolog of the BapA protein necessary to mediate bacterial recruitment into the biofilm pellicle [35]. BaeR regulates multidrug and metal efflux resistance systems [36] and is a component of the SPI-9 pathway. We show below that BaeR is the major regulator of gene expression in *S.* Typhi after 8 hours of macrophage infection according to the network analysis of microarray data. It can also synergize with the PhoR/PhoP signaling in *E. coli* [37]. Our results suggest that in *Salmonella*, BaeR may synergize with PhoP in response to the acidification of the environment in phagosomes during the infection.

Table 3: List of proteins encoded by SPI-8. At the early stage of infection, *S.* Typhi may use bacteriocin (pyocin) to kill other bacteria in the intestine in order to compete for nutrients. Remarks: Hypothetical protein refers to the predicted protein but without any putative function. Putative is a protein that has function predicted based on sequence similarity.

Protein	Description
STY3273	Putative prophage P4 integrase
STY3274	Hcp
STY3277	Vgr-like protein
STY3278	Hypothetical protein
STY3279	Hypothetical protein
STY3280	S-type Pyocin
STY3281	Colicin immunity protein / pyocin immunity protein
STY3282	Colicin immunity protein / pyocin immunity protein
STY3283	Colicin immunity protein / pyocin immunity protein
STY3285	Hypothetical protein
STY3287	Hypothetical protein
STY3288	Enterobacterial putative membrane protein (DUF943)
STY3289	Hypothetical protein
STY3290	Hypothetical protein
STY3291	Putative membrane protein
STY3292	Putative membrane protein

3.1.10 Salmonella Pathogenicity Island 10 (SPI-10)

SPI-10 has 29 genes that encode a Sef/Pef fimbrial islet, transposases, helicases, IS element, and P4 like-phage proteins [38]. The overview of SPI-10 is illustrated in Figure 10. Three genes of the *sef* operon (*sefA*, *sefD*, and *sefR*) contain multiple frame-shift mutations. Indeed, microarray data showed that the *sef* genes are not expressed in *S.* Typhi [39]. SPI-10 has a truncated *pefI* gene and lacks *pefA*, *pefB*, *pefC*, and *pefD* in comparison to the *pef* operon of *S. typhimurium* [38]. The presence of P4-like phage, transposase , helicases, IS element, and integrase suggest that this is a hot spot for the insertion of transposable elements which played a major role in driving the variability of this region [38].

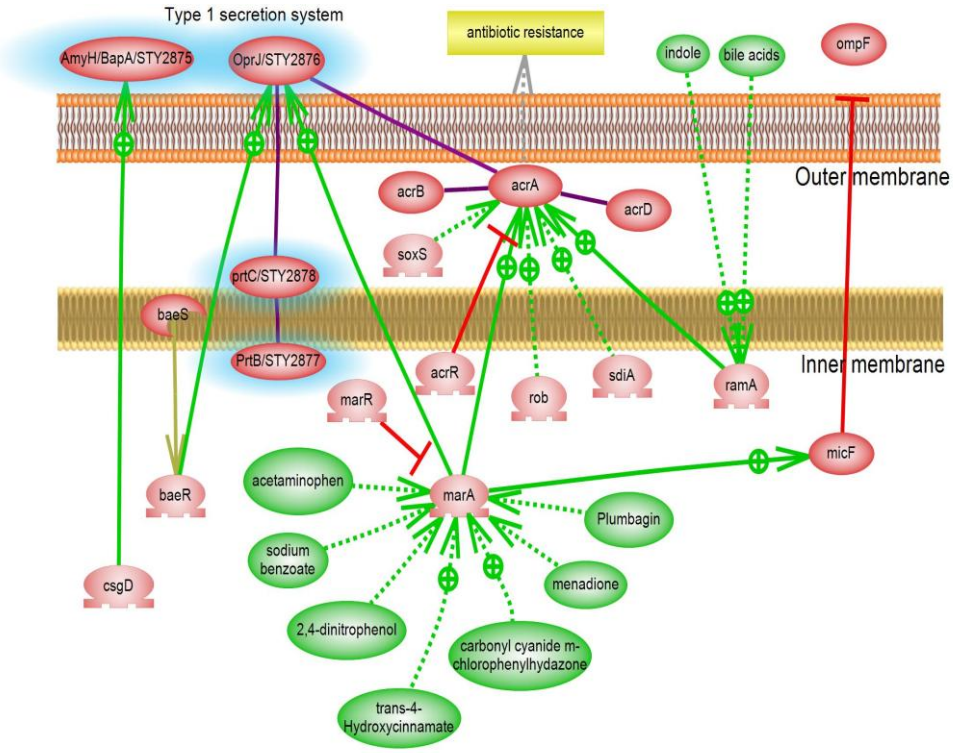


Figure 9: Multidrug resistance efflux pumps encoded by SPI-9. TolC, AcrAB, and BaeR regulate multidrug resistance to pump out bile acids, antibiotics, dyes, and disinfectants. Proteins encoded by SPI-9 are highlighted in blue. A detailed view of the SPI-9 pathway including supporting literature is available at http://www.ccbusm.com/publications/spi/SPI-9.html.

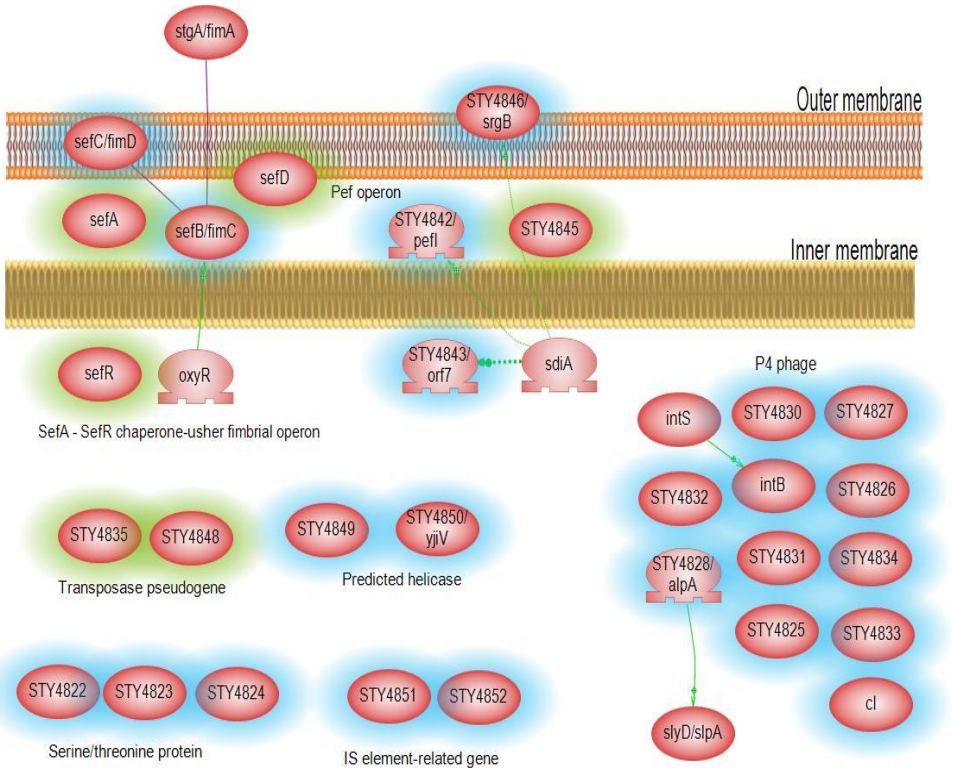
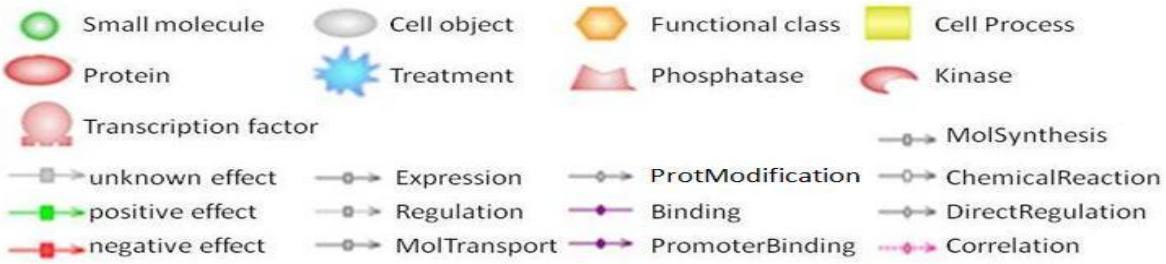


Figure 10 - SPI-10 proteins shown in pathway diagram form. SPI-10 is a hot spot for the insertion of transposable elements which played a major role in driving the variability of this

region. Proteins encoded by SPI-10 are highlighted in blue. A detailed view of the SPI-10 pathway including supporting literature is available at http://www.ccbusm.com/publications/spi/SPI-10.html.



Legend for Figure 2 to Figure 10.

3.2 Validation of SPI regulatory pathway by network analysis of gene expression during macrophage infection by *Salmonella*

3.2.1 SPI pathway validation by network enrichment analysis of Salmonella gene expression time-course during macrophage infection

Table 4 shows the major transcriptional regulators identified by sub-network enrichment analysis (SNEA) at 2, 8, and 24 hours of macrophage invasion. SNEA is described in the Methods section. We only report and discuss transcriptional factors with p-values smaller than 0.05 as calculated by SNEA. We found that the PhoP transcription factor is active in the beginning of invasion. PhoP is a component of our SPI-1/2/3/4/5 pathways. Interestingly, the period of PhoP activity coincides with the down-regulation of Lrp targets. Lrp is a component of our SPI-1 pathway and is a major expression regulator in Table 1. SNEA identifies major regulators that are either activated or inhibited according to the expression data. The analysis of expression changes for Lrp targets revealed that this global regulator is repressed in the beginning of infection because most of its targets are down-regulated (data not shown). Genes inhibited by Lrp apparently become de-repressed during infection because Lrp is no longer significant after 8 hours. After Lrp targets are de-repressed, PhoP is no longer active. Thus, SNEA results suggest that PhoP appears to initiate the transcriptional program necessary for survival inside macrophage phagosomes together with SlyA (STY1678) transcription factor. SlyA is a component of our SPI-2 pathway.

SsrB transcriptional factor is encoded by SPI-2 and remains significant during the entire infection time-course. SNEA results also suggest that integration host factor (IHF) and BaeR transcriptional factor appear to drive up the expression of most differentially expressed genes after 8 hours of invasion (Table 1). IHF is a component of our SPI-7 regulation pathway and BaeR is a component of SPI-9 pathway. RpoN (sigma 54) targets are significantly down-regulated throughout the entire time-course. RpoN is a component of our SPI-1/4/5 pathways. We further explain these results in the Discussion section.

Table 4: Most significant transcription factors identified by sub-network enrichment analysis (SNEA) from the time course of Salmonella invasion of human macrophages. pValue of SNEA indicates statistical significance of differential expressed downstream genes targeted by the transcription factors. This in turn signifies the activity of the transcription factor in the experiment.

Expression data was obtained from Gene Expression Omnibus at NCBI (GEO accession number GSE3096). Expression conformity shows how many targets are up- or down-regulated in the right direction relative to the reported activity of the transcription factor (which can be activator or repressor) towards the target.

Regulator	Regulator expression, log2	# of measured targets	SNEA p-value	Expression conformity %
2 hours after invasion				
phoP	0.99	36	0.000346406	69.4
ssrB	2.6	4	0.0165948	75
slyA	1.69	6	0.0272351	100
rpoN	-0.37	38	0.0440864	71.1
lrp	1.04	32	0.0493375	75
8 hours after invasion				
ihfA	1.05	14	0.0184476	57.1
ssrB	-0.05	4	0.0215288	75
rpoN	-0.41	37	0.0275744	64.9
baeR	-0.05	14	0.0295999	78.6
24 hours after invasion				
ihfA	1.15	14	0.0184476	64.3
ssrB	1.62	5	0.0215288	75
rpoN	-0.55	37	0.0275744	75.7
baeR	0.25	14	0.0295999	64.3

3.2.2 Validation of pathways by clustering analysis of Salmonella SPI genes during macrophage infection

Co-expressed genes tend to participate in common biological processes [40,41]. Therefore, to further validate our SPI regulatory pathways we have investigated the correlation among expression profiles of genes in our SPI pathways. We have identified a significant number of genes in each SPI pathway with expression correlated during the time-course of *Salmonella* invasion of macrophages. SPI-1 genes form two distinct gene expression clusters during the time-course of *Salmonella* infection of macrophages. Expression profile of the biggest cluster SPI-1 is shown in Figure 11a. Gene clusters for other SPI pathways are reported in Figure 12 and 13 respectively. In the figure legend we show how the combination of gene expression clustering and pathway analysis allows the identification of principal transcriptional factors controlling expression of genes co-regulated in the cluster.

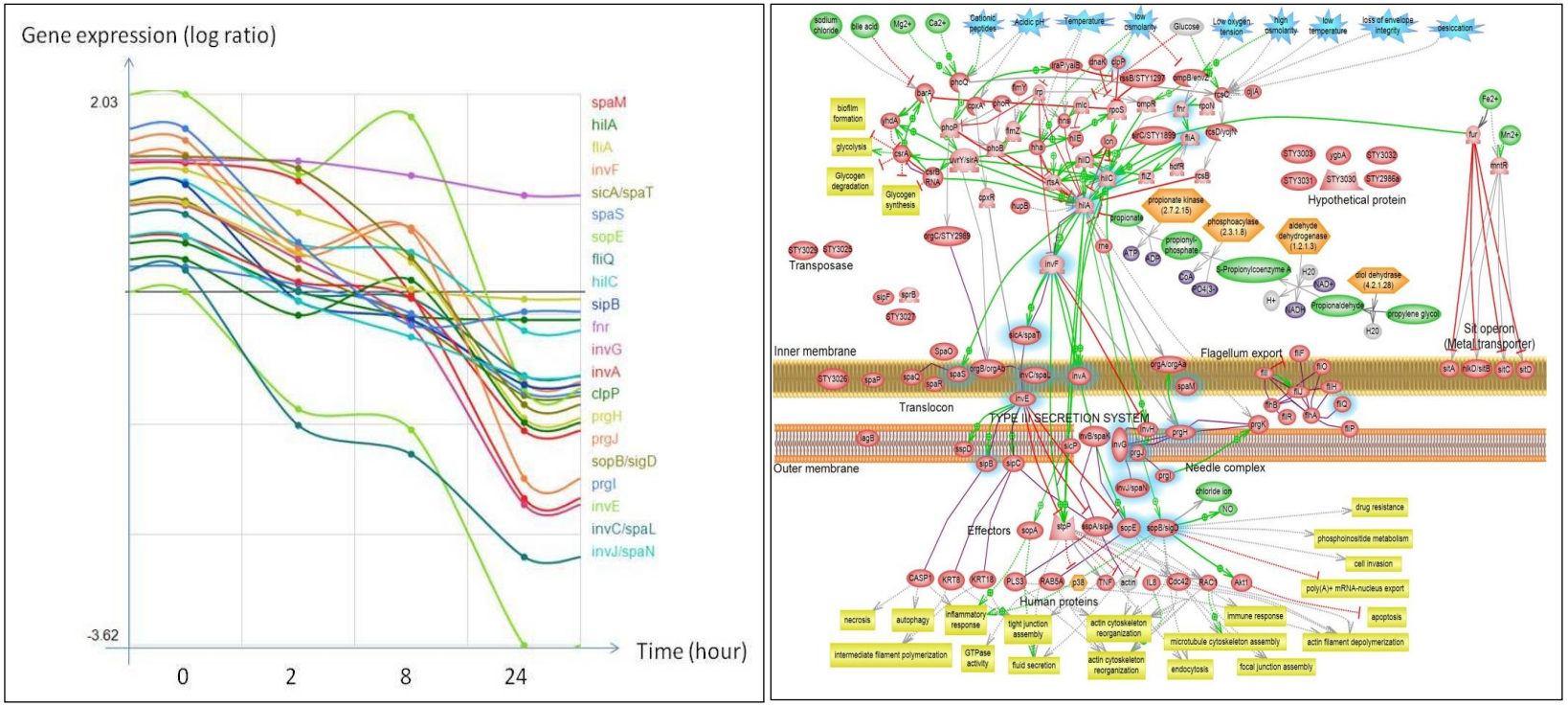


Figure 11a: Cluster 1A. This cluster consists of genes which have positive correlation. It appears that this group of genes plays a significant role especially during invasion in the macrophage (left). The expression profile corresponds to the proteins as highlighted in blue. It clearly depicts that the signals are being transmitted from the regulator to the Type III secretion system proteins and effector proteins which finally interact with the human proteins (right). Cluster 1A also revealed that many genes in SPI-1 pathway have expression profile similar to *hilA* and *hilC* profile, suggesting that the genes in this cluster are under stringent control of these two transcription factors. Their common expression profile also supports functional commonality of proteins in SPI-1 pathway. The most upstream transcription factor in this cluster is oxygen sensor *fnr* that controls the expression of *fliA* sigma factor to turn on *hilA* and *hilC* expression. This suggests that low oxygen concentration is the main trigger initiating genetic program for invasion of macrophages. Our findings are consistent with previously reported *fnr* role for *Salmonella* survival inside the host cells [42].

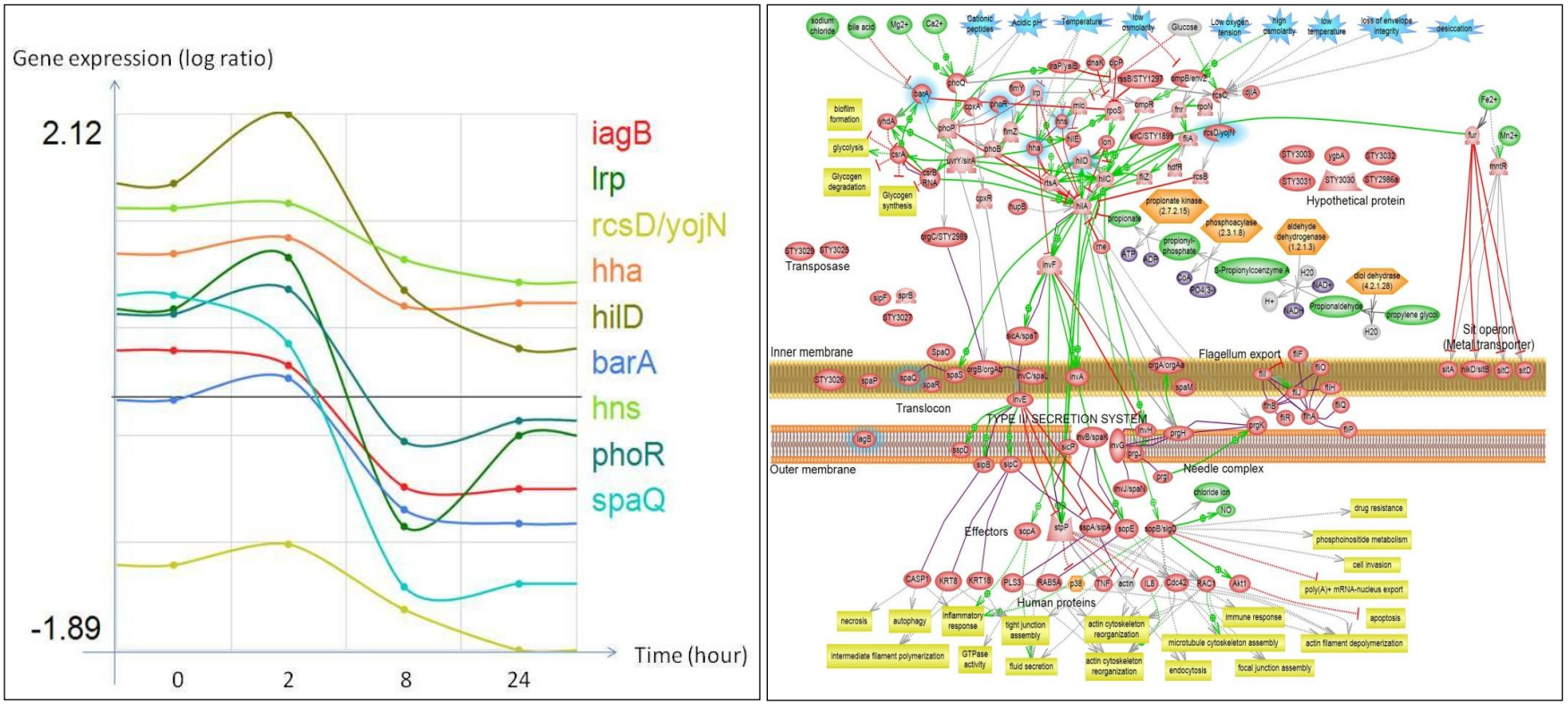


Figure 11b: Cluster 1B. SPI-1 genes form two distinct gene expression clusters in the time-course of *Salmonella* infection of macrophages. First cluster is shown on Figure 11a. Expression profile of second cluster is shown here. This cluster consists of genes which have positive correlation. It appears that this group of genes plays a significant role in the signaling pathway (left). The expression profile corresponds to the proteins as highlighted in blue. It shows that in this cluster, the sensor and transcriptional factors are positively correlated (right). Environmental sensors *barA, rcsD* and *phoR* have expression profile similar to *lrp* and *hilD* profile. This analysis also shows that *hilD* expression is controlled by *lrp* activity through *hns* global transcription regulator. Both *hns* and *hha* transcription factor are controlled by low osmolarity suggesting that this environmental signal is sensed by *Salmonella* during the macrophage invasion.

Table 5: Description for the genes in Figure 11a. The colour of gene corresponds to the colour of the line in the gene expression graph in SPI-1.

List of genes	Description	List of genes	Description
spaM	Needle complex assembly protein	*fnr*	DNA-binding transcriptional dual regulator, global regulator of anaerobic growth
hilA	Invasion protein transcriptional activator	*invG*	Type III secretion apparatus protein
fliA	RNA polymerase, sigma 28 (sigma F) factor	*invA*	Needle complex export protein
invF	Putative regulatory protein for type III secretion apparatus	*clpP*	ATP-dependent Clp protease proteolytic subunit
sicA/spaT	Type III secretion low calcium response chaperone LcrH/SycD	*prgH*	Needle complex inner membrane protein; pathogenicity 1 island effector protein
spaS	Surface presentation of antigens protein SpaS	*prgJ*	Putative Type III secretion apparatus protein
sopE	Invasion-associated secreted protein	*sopB/sigD*	Secreted effector protein
fliQ	Flagellar biosynthesis protein	*prgI*	Type III secretion protein
hilC	Invasion regulatory AraC family transcription regulator	*invE*	Putative secreted protein
sipB	Cell invasion protein	*invC/spaL*	ATP synthase SpaL
		invJ/spaN	Needle length control protein

Table 6: Description for the genes in Figure 11b. The colour of gene corresponds to the colour of the line in the gene expression graph in SPI-1.

List of genes	Description
iagB	Invasion protein IagB; Lytic transglycosylase, catalytic
lrp	DNA-binding transcriptional dual regulator, leucine-binding
rcsD/yojN	Phosphotransfer intermediate protein in two-component regulatory system with RcsBC
hha	Modulator of gene expression, with H-NS
hilD	Invasion AraC family transcription regulator
barA	Hybrid sensory histidine kinase, in two-component regulatory system with UvrY
hns	Global DNA-binding transcriptional dual regulator H-NS
phoR	Sensory histidine kinase in two-component regulatory system with PhoB
spaQ	Type III secretion apparatus protein

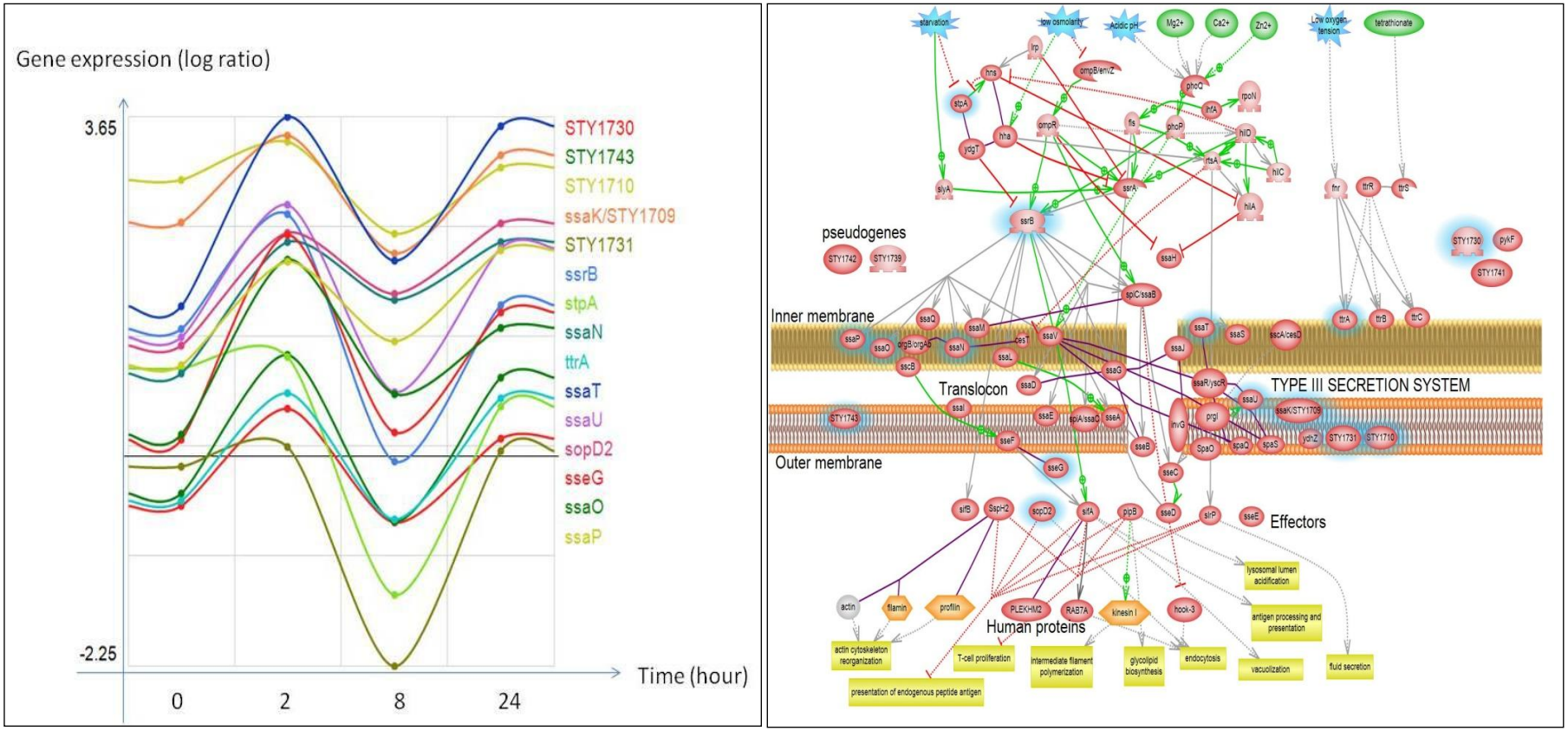


Figure 12a: Cluster 2A. SPI-2 genes form two distinct gene expression clusters during the time-course of Salmonella infection of macrophages. This cluster of genes showed positive correlation during the systemic infection (left). The expression profile corresponds to the proteins highlighted in blue. It can be seen that the main regulator is ssrB and most of the translocon, type III secretion system and effector genes have the similar profile (right). Expression profile graph of cluster 2A shows that the main environmental stimulus is starvation which is sensed by *stpA* and *slyA*. The signal is then transmitted to *ssrB*, the main regulator in cluster 2A. *SlyA* was also found a significant regulator by sub-network enrichment analysis after 2 hours of infection. Note the temporary down-regulation of entire cluster at 8 hours of infection. This can be explained by the switch in *ssrAB* control. Initially it may be activated by *slyA* in response to starvation and later in the infection s*srAB* expression can be controlled by either *stpA* and or *lrp* global regulators that are also respond to starvation.

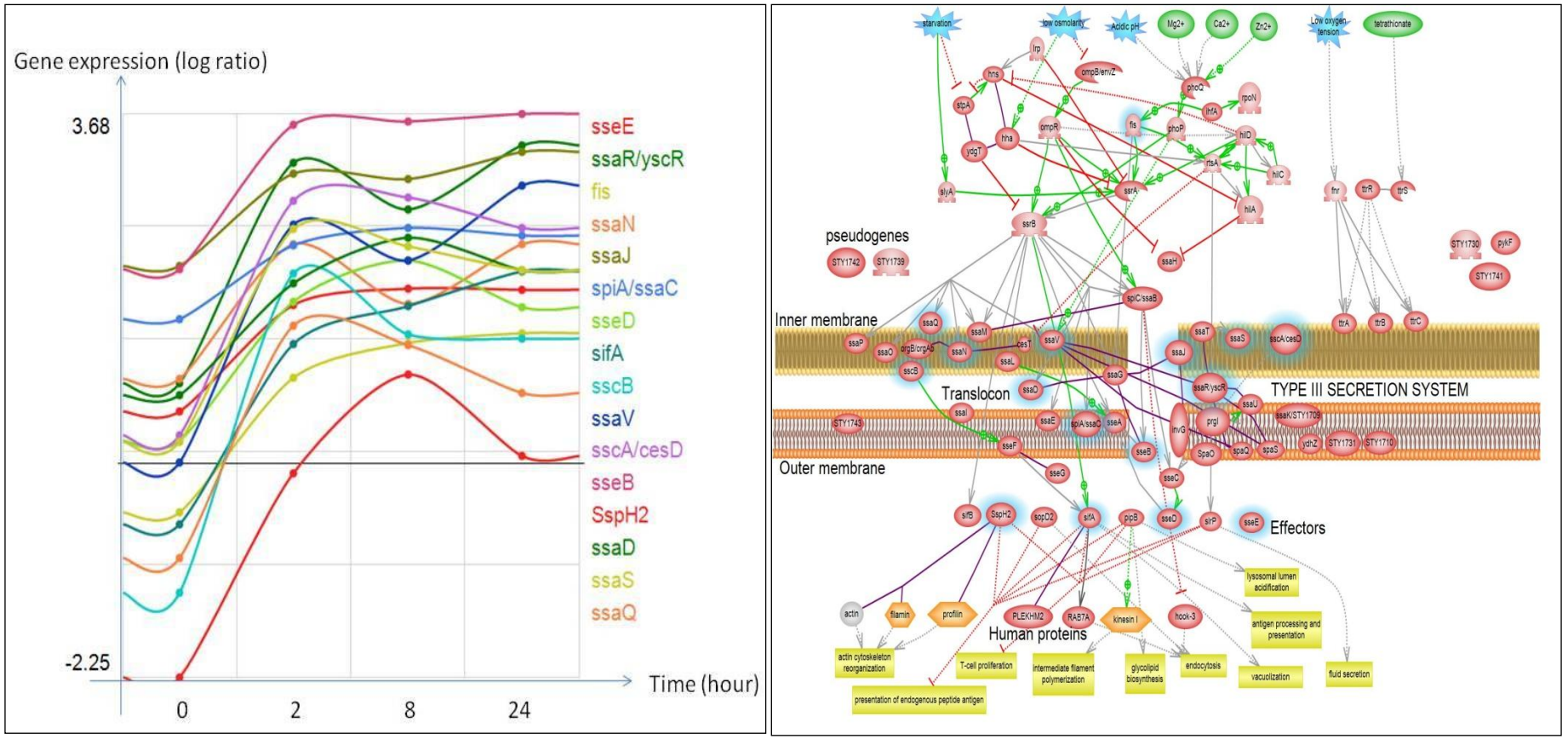


Figure 12b: Cluster 2B. This cluster of genes showed positive correlation during the systemic infection (left). The expression profile corresponds to the proteins highlighted in blue. It can be seen that the main regulator is ssrB and most of the translocon, type III secretion system and effector genes have the similar profile (right). The only transcription factor in cluster 2B is Fis protein. However, *fis* does directly regulate genes in this cluster but does it through expression of *ssrAB* operon according to our SPI-2 pathway. The only difference between profiles of cluster 2A and 2B containing *ssrAB* is expression at 8hrs of infection. *Fis* is required for activation of *ssrA* expression in murine macrophages through DNA relaxation [56]. It appears that genes in cluster 2A are more under *fis* controlled than *ssrAB* control perhaps because their expression is more sensitive to DNA relaxation than the expression of genes in cluster 2B which appear under stringent *ssrAB* control.

Table 7: Description for the genes in Figure 12a. The colour of gene corresponds to the colour of the line in the gene expression graph in SPI-2.

List of genes	Description	List of genes	Description
sty1730	Predicted DNA-binding transcriptional regulator	*ssaN*	Flagellum-specific ATP synthase
sty1743	Putative amino acid permease	*ttrA*	Tetrathionate reductase subunit A
sty1710	Secretion system apparatus	*ssaT*	Putative type III secretion protein
ssaK/STY1709	Type III secretion system apparatus protein	*ssaU*	Secretion system apparatus protein SsaU
STY1731	Conserved protein	*sopD2*	Secreted protein
ssrB	DNA-binding response regulator in two-component regulatory system with EvgS	*sseG*	Secreted effector protein
stpA	DNA binding protein, nucleoid-associated	*ssaO*	Archaeal flagella-related protein D, type III secretion protein
		ssaP	Type III secretion system apparatus protein

Table 8: Description for the genes in Figure 12b. The colour of gene corresponds to the colour of the line in the gene expression graph in SPI-2.

List of genes	Description	List of genes	Description
sseE	Secreted effector protein	*sscB*	Secretion system chaparone
ssaR/yscR	Type III secretion system protein	*ssaV*	Secretion system apparatus protein SsaV
fis	Global DNA-binding transcriptional dual regulator	*sscA/cesD*	Putative Type III secretion system chaperone protein
ssaN	Flagellum-specific ATP synthase	*sseB*	Secreted protein EspA
ssaJ	Needle complex inner membrane lipoprotein	*SspH2*	Leucine-rich repeat protein
spiA/ssaC	Putative outer membrane secretory protein	*ssaD*	Putative pathogenicity island protein
sseD	Translocation machinery component	*ssaS*	Flagellar biosynthesis protein Q
sifA	Secreted effector protein	*ssaQ*	Flagellar motor switch/type III secretory pathway protein

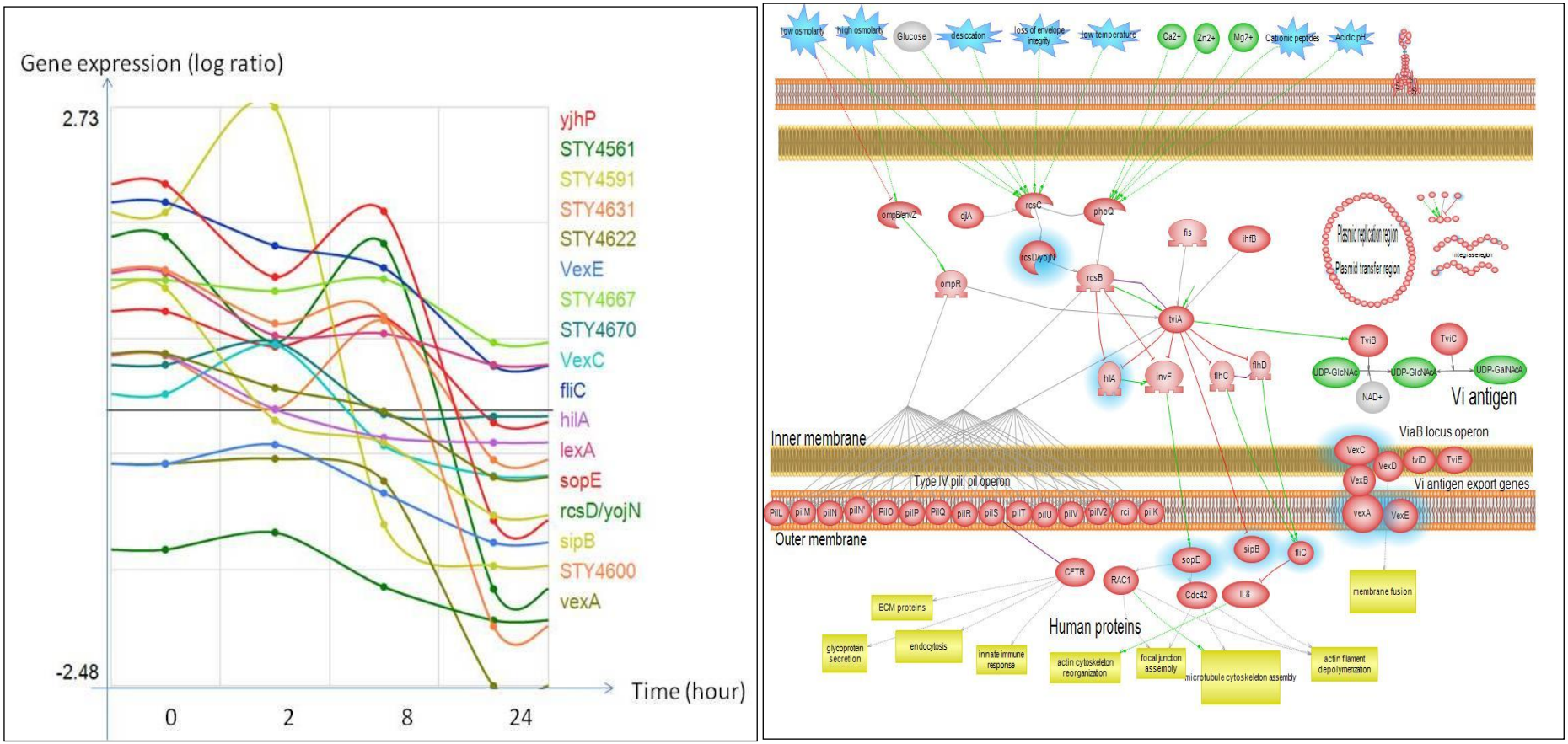


Figure 13: Cluster 7A. SPI-7 genes form one distinct gene expression cluster during the time-course of *Salmonella* infection of macrophages. Expression profile graph of cluster 7A shows that the expression of Vex genes/exopolysaccharide export genes is positively correlated. In this case, *rcsD, hilA, sopE, sipB, fliC* and some of the phage-related proteins have similar profile. According to [32], *rcsB* acts together with *tviA* which is encoded by the first gene of *viaB* locus in order to activate *viaB* transcription from the *tviA* promoter. Unfortunately, *tviA* is not measured on the chip and thus, its profile could not be determined.

Table 9: Description for the genes in Figure 13. The colour of gene corresponds to the colour of the line in the gene expression graph in SPI-7.

List of genes	Description	List of genes	Description
yjhP	KpLE2 phage-like element; predicted methyltransferase	*VexC*	VI polysaccharide export ATP-binding protein
sty4561	Restriction endonuclease	*fliC*	Flagellar filament structural protein (flagellin)
sty4591	Type I site-specific restriction-modification system, R subunit	*hilA*	Invasion protein transcriptional activator
sty4631	ATP/GTP binding protein	*lexA*	LexA repressor
sty4622	phage tail protein X	*sopE*	Invasion-associated secreted protein
VexE	VI polysaccharide export protein	rcsD/yojN	Phosphotransfer intermediate protein in two-component regulatory system with RcsBC
sty4667	CopG-like DNA-binding	*sipB*	Cell invasion protein
sty4670	Glucosamine-6-phosphate deaminase-like protein	*STY4600*	DNA-binding transcriptional regulator prophage P2 remnant
		vexA	Predicted exopolysaccharide export protein

4 Discussion

4.1 Construction and applications of SPI regulatory pathways for Salmonella

To date, 17 SPIs have been discovered in *S.* enterica [12, 43, 44, 45]. Nine of these SPIs are present in the genome of S.Typhi CT18 and were chosen for pathway reconstruction because experimental data is available to validate them. Our pathways are readily available for the analysis of future experimental data and for comparison of different *Salmonella* species. In total, our SPI pathways have 463 interactions with 157 of them classified as direct physical interactions. Pathways are consistent with previously published literature on *Salmonella* infection since only interactions reported in the literature were used for construction. We showed how to use new SPI pathways for analysis of gene expression microarray data inside Pathway Studio software. Because proteins in our SPI pathways are annotated with identifiers from multiple *Salmonella* species, the pathways can also be used for comparison of invasion mechanisms between different *Salmonella* strains. Our SPI pathways also provide a list of candidate biomarkers for *Salmonella* infection. The most suitable biomarkers for clinical diagnostics are proteins secreted and exposed outside the *Salmonella* cell and induced during the infection. The list of such proteins is readily available from our SPI pathways and can be used for development of diagnostics using ELISA assay. Further challenges associated with the development of diagnostics kit which must be specific to *Salmonella* species and at the same time provide comprehensive coverage of all enteric species can be addressed by comparison of SPI pathways between invasive *S.*enterica and other *Salmonella* species.

While literature suggests that SPI pathways can be activated by different environmental factors such as osmolarity, oxygen level, temperature acidic pH and cation concentration we found that the major factors activating Salmonella infection in macrophages are changes in starvation and osmolarity.

Our pathways also revealed that there is a lack of literature knowledge about SPI-6, SPI-8, SPI-9, and SPI-10 regulation. This knowledge gap does not allow complete reconstruction of the regulatory pathways for these regions and point to the areas for further experimental research, thus helping to develop most efficient research strategy for full understanding of *Salmonella* invasion mechanism which leads to typhoid fever outbreaks.

4.2 Experimental validation of SPI regulatory pathways

We have validated our SPI pathways by comparing them with the publically available microarray data. For comparison, we used statistical methods that have never been used for analysis of GSE3096 dataset. Therefore our network analysis provides novel findings never previously reported. The GSE3096 dataset measures the expression of the entire *Salmonella* genome and therefore represents an unbiased and independent sample that can be used for cross-validation of any pathways and networks constructed for *Salmonella* based on the information from other sources. Our only source of data for construction of SPI pathways was *Salmonella* protein interactions reported in peer-reviewed scientific literature. Most of these interactions were measured either prior to publication of the GSE3096 dataset or were determined by different methods and in different experiments unrelated to GSE3096. Comparison with GSE3096 showed that the behavior of genes in our SPI pathways is consistent with the current view on *Salmonella* infection. The SPI-1 pathway is turned on during the first hours of host cell invasion, while the SPI-2 and SPI-3 pathways are necessary for survival inside host cell phagosomes and are activated at later stages of the infection.

Sub-network enrichment analyses of the expression time-course of *Salmonella* genes during human macrophage invasion identified several transcription factors (PhoP, IHF, SlyA, and Lrp) that were previously shown in the literature to be significant for infection and survival in phagosomes and therefore were components of our SPI pathways. SNEA also found novel significant transcription factors (RpoN and BaeR) that have never been reported playing a role during infection. RpoN is significantly down-regulated during the infection. This is evident from the levels of its mRNA expression as well as from the expression of its targets. One possible biological function of RpoN down-regulation is activation of PhoP transcription factor. PhoP acts upstream of SsrB and is essential for intra-macrophage control of T3SS [46]. PhoP was reported to bind to the *ssrB* promoter when *Salmonella* are inside macrophages [46]. It has been shown that RpoN opposes PhoP activation *in vivo*: the deletion of *rpoN* attenuates *S.* Typhi virulence and increases resistance to the cationic antimicrobial peptide polymyxin B [47]. Polymyxin B resistance is mediated by the PhoQ-PhoP system and *rpoN* deletion appears to act independently from PhoP by providing an alternative mechanism to develop Polymyxin B resistance [47]. Thus, down-regulation of *rpoN* during macrophage invasion may provide additional boost to PhoP activation.

Identification of major regulators by SNEA combined with analysis of SPI regulatory pathways allows identification of major environmental stimuli used by *Salmonella* to initiate program of macrophage host invasion. For example, PhoQ-PhoP system can be activated either by acidic pH or by lower concentration of divalent cations (Ca^{2+} or Mg^{2+}) according to our SPI-1 pathway [48, 49]. *Salmonella* forms a capsule in the macrophage lysosome to escape host intracellular defense mechanism. The intra-lysosomal environment is very acidic. The link between Mg^{2+} concentration, PhoQ-PhoP, and transcriptional regulation of *Salmonella* invasion genes was reported previously [50]. It has been further suggested that PhoP-activated genes are highly expressed within the host cells due to the low intraphagosomal Mg^{2+} concentration and these genes are necessary for intramacrophage survival [51]. The inactivation of Leucine-responsive regulatory protein (Lrp) appears to be noteworthy at the first 2 hours after invasion. It was reported that Lrp is a master regulatory protein that coordinates expression of most bacterial operons in response to nutrient availability [52, 53]. It has been reported recently that lrp deletion promotes *Salmonella* virulence [54]. This is consistent with our findings that lrp is down-regulated after the first 2 hours of infection.

IHF (IhfA) and SlyA are also known SPI-2 regulators [16] and are included in our SPI-2 pathway. Expression of IhfA appears to be significant during 8 hours and 24 hours after invasion. According to [17], IHF was found to be essential for SPI-1 expression at early to late exponential growth phase and IHF levels possibly coordinate the expression of SPI-1 and SPI-2 genes. This is further supported by the previous work by [57] that shows IHF integrates stationary-phase and virulence gene expression and plays a critical role in the co-regulatory process. Expression of SlyA is significant during the first 2 hours after invasion. This is in accordance with the findings by [58] which reported that SlyA regulon is activated during infection of the host and at least 2 proteins expressed in macrophages were found to be SlyA-dependent.

The involvement of transcriptional factor BaeR in the invasion process has not been reported previously. BaeR is identified in this work as the major regulator of gene expression in *Salmonella* after 8 hours of infection. BaeR was shown to regulate multidrug and metal efflux resistance systems [36] and is a component of our SPI-9 pathway. In *E.coli*, the BaeRS system was shown to influence indirectly the expression of PhoR-PhoB system which is part of our SPI-1 pathway [55]. PhoB is downstream of PhoP and necessary for PhoP regulation of HilA expression according to our SPI-1 pathway. Thus, our results suggest that BaeR can synergize with PhoP in response to the acidification and low cation concentration inside host

phagosomes during the infection. It also suggests that *Salmonella* needs the increased production of multidrug efflux resistance pump in order to survive inside lysosomes, which is a convenient target for anti-typhoid drug development.

SNEA also found other transcription factors from our SPI pathways such as HilA, FlgM, InvF, MarA, and RfaH with p-values higher than 0.05. The p-value range calculated by SNEA depends greatly on the size of the microarray chip, which defines the size of reference distribution of expression values. Smaller chips tend to produce larger SNEA p-values due to the smaller statistical power provided by reference distribution. Therefore, SNEA p-values for smaller chips such as *Salmonella* genome chip can be used only as a relative rather than absolute measure of transcription factors activity. We reported and discussed only transcription factors with SNEA p-values below the conventional 0.05 cut off emphasizing those that previously were not reported to play a role during macrophage infection. Other transcription factors in our SPI pathways should be active during the infection, suggesting that the 0.05 cutoff was too stringent for the *Salmonella* chip.

4.3 Overview of pathogenicity islands' interaction

The construction of pathogenicity island pathways enables us to identify the higher level interdependencies between SPIs which are regulated by the common global regulators. Understanding of these interdependencies is necessary to predict pathogenicity of different *Salmonella* strains carrying various combinations of SPI regions in the genome. We found that SPI-1 is interconnected with SPI-4, SPI-5, and SPI-7. Activation of SPI-4 proteins is dependent on the regulators in SPI-1, secretion of SigD/SopB encoded by SPI-5 is via T3SS in SPI-1, and SopE encoded by SPI-7 is also secreted through SPI-1 T3SS. Similarly, SPI-2 is interconnected with SPI-5, whereby PipA and PipB from SPI-5 are secreted through T3SS encoded by SPI-2. STY3274 and STY3277 which are encoded in SPI-8 are related to SPI-6; STY3274 is secreted via T6SS and STY3277 is a T6SS Vgr family protein. SPI-6 and SPI-10 both have chaperon-usher fimbrial operon; *saf* and *sef* operon respectively. It was also shown that both SPI-4 and SPI-9 encodes for T1SS [22, 25]. Interestingly, genes in SPI-3 are not connected to other pathways but this SPI is controlled by PhoQ-PhoP system which is found in SPI-1, 2, 4, 5, and 7. SPI-3 is very important for the ability of *S.*Typhi to survive in the macrophage with Mg^{2+} limiting conditions. A summary of the interactions between the different SPIs is shown in figure 14.

5 Conclusion

We have built the collection of nine pathways regulating different stages of *S.* Typhi infection including host invasion, intracellular host survival, and drug resistance. Our collection shows that nine of the SPIs are interconnected and play an important role for Typhoid Fever. In general, *S.*Typhi is capable of responding to various environmental challenges such as acidic pH, low temperature, high osmolarity, and in response to divalent cations (Ca^{2+}, Mg^{2+}, Zn^{2+}). The pathways were validated by analysis of gene expression data. Sub-network enrichment analysis of gene expression has confirmed several major regulators crucial for SPI regulation and identified one novel transcription factor activated during macrophage infection. We have identified several clusters of genes co-expressed during macrophage infection in our SPI pathways.

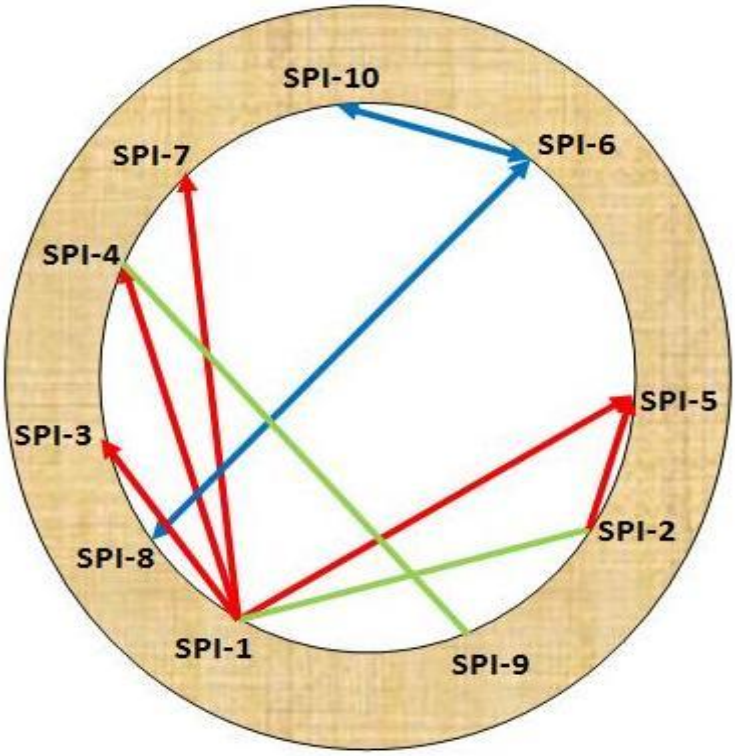


Figure 14: Schematic diagram showing the interdependencies between the 10 SPIs. Single-headed red arrow indicates that function of one SPI region (target) depends on the function of another SPI region (regulator). Green line indicates that both have the similar secretion system while double-headed blue arrow indicates that the gene/operon is interrelated between the SPIs.

Acknowledgements

We thank Hock Siew Tan for initial participation in the study and Dr. Jennifer Saito for critical reading of the manuscript. The project was funded by an intramural grant from Universiti Sains Malaysia.

References

[1] Typhoid vaccines: WHO position paper. Wkly Epidemiol Rec, 83(6):49-59, 2008.

[2] R. W. Crawford, D. L. Gibson, W. W. Kay and J. S. Gunn. Identification of a bile-induced exopolysaccharide required for *Salmonella* biofilm formation on gallstone surfaces. Infect Immun, 76(11):5341-5349, 2008.

[3] J. Hacker and J. B. Kaper. Pathogenicity islands and the evolution of microbes. Annu Rev Microbiol, 54:641-679, 2000.

[4] E. F. Vanin. Processed pseudogenes: characteristics and evolution. Annu Rev Genet, 19:253-272, 1985.

[5] S. Novichkova, S. Egorov, and N. Daraselia. MedScan, a natural language processing engine for MEDLINE abstracts. Bioinformatics, 19(13): 1699-706, 2003.

[6] A. Yuryev, Z. Mulyukov, E. Kotelnikova, S. Maslov, S. Egorov, A. Nikitin, N. Daraselia and I. Mazo: Automatic pathway building in biological association networks. BMC Bioinformatics, 7:171, 2006.

[7] R. An, S. Sreevatsan and P. S. Grewal: Comparative in vivo gene expression of the closely related bacteria *Photorhabdus temperata* and *Xenorhabdus koppenhoeferi* upon infection of the same insect host, *Rhizotrogus majalis*. BMC Genomics, 10: 433, 2009.

[8] Haiyuan Yu, Nicholas M. Luscombe, Hao Xin Lu, Xiaowei Zhu, Yu Xia, Jing-Dong J. Han, Nicolas Bertin, Sambath Chung, Marc Vidal and M. Gerstein. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. Genome Res., 14(6):1107-1118, 2004.

[9] I. Ispolatov, A. Yuryev, I. Mazo and S. Maslov. Binding properties and evolution of homodimers in protein-protein interaction networks. Nucleic Acids Res, 33(11):3629-3635, 2005.

[10] Sébastien P. Faucher, Steffen Porwollik, Charles M. Dozois, Michael McClelland and F. Daigle. Transcriptome of *Salmonella enterica* serovar Typhi within macrophages revealed through the selective capture of transcribed sequences. Proc Natl Acad Sci U S A., 103(6):1906-1911, 2005.

[11] A. Y. Sivachenko, A. Yuryev, N. Daraselia and I. Mazo. Molecular networks in microarray analysis. J Bioinform Comput Biol, 5(2B):429-56, 2007.

[12] Morgan E. Salmonella Pathogenicity Islands. In: Salmonella: Molecular Biology and Pathogenesis (Eds Rhen M, Maskell D, Matroeni P and Threlfall J). Horizon Bioscience: pp. 67-88, 2007

[13] A. Janakiraman and J. M. Slauch. The putative iron transport system SitABCD encoded on SPI1 is required for full virulence of *Salmonella typhimurium*. Mol Microbiol, 35(5):1146-1155, 2000.

[14] D. G. Kehres, A. Janakiraman, J. M. Slauch and M. E. Maguire. SitABCD is the alkaline Mn(2+) transporter of *Salmonella enterica* serovar Typhimurium. J Bacteriol, 184(12):3159-3166, 2002.

[15] M. L. Zaharik, V. L. Cullen, A. M. Fung, S. J. Libby, S. L. Kujat Choy, B. Coburn, D. G. Kehres, M. E. Maguire, F. C. Fang and B. B. Finlay. The *Salmonella enterica* serovar typhimurium divalent cation transport systems MntH and SitABCD are essential for virulence in an Nramp1G169 murine typhoid model. Infect Immun, 72(9):5522-5525, 2004.

[16] E. Fass and E. A. Groisman. Control of *Salmonella* pathogenicity island-2 gene expression. Curr Opin Microbiol, 12(2):199-204, 2009.

[17] M. Hensel, A. P. Hinsley, T. Nikolaus, G. Sawers and B. C. Berks. The genetic basis of tetrathionate respiration in *Salmonella typhimurium*. Mol Microbiol, 32(2):275-287, 1999.

[18] M. Hensel, T. Nikolaus and C. Egelseer. Molecular and functional analysis indicates a mosaic structure of *Salmonella* Pathogenicity Island 2. Mol Microbiol, 31(2):489-498, 1999.

[19] E. L. Barrett and M. A. Clark. Tetrathionate reduction and production of hydrogen sulfide from thiosulfate. Microbiol Rev, 51(2):192-205, 1987.

[20] E. A. Groisman. The ins and outs of virulence gene expression: Mg2+ as a regulatory signal. Bioessays, 20(1):96-101, 1998.

[21] C. W. Dorsey, M. C. Laarakker, A. D. Humphries, E. H. Weening and A. J. Baumler. *Salmonella enterica* serotype Typhimurium MisL is an intestinal colonization factor that binds fibronectin. Mol Microbiol, 57(1):196-211, 2005.

[22] J. Parkhill, G. Dougan, K. D. James, N. R. Thomson, D. Pickard, J. Wain, C. Churcher, K. L. Mungall, S. D. Bentley, M. T. Holden, M. Sebaihia, S. Baker, D. Basham, K. Brooks, T. Chillingworth, P. Connerton, A. Cronin, P. Davis, R. M. Davies, L. Dowd, N. White, J. Farrar, T. Feltwell, N. Hamlin, A. Haque, T. T. Hien, S. Holroyd, K. Jagels, A. Krogh, T. S. Larsen, S. Leather, S. Moule, P. O'Gaora, C. Parry, M. Quail, K. Rutherford, M. Simmonds, J. Skelton, K. Stevens, S. Whitehead and B. G. Barrell. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. Nature, 413(6858):848-852, 2001.

[23] K. L. Main-Hester, K. M. Colpitts, G. A. Thomas, F. C. Fang and S. J. Libby. Coordinate regulation of Salmonella pathogenicity island 1 (SPI1) and SPI4 in *Salmonella enterica* serovar Typhimurium. Infect Immun, 76(3):1024-1035, 2008.

[24] T. Kiss, E. Morgan and G. Nagy. Contribution of SPI-4 genes to the virulence of *Salmonella enterica*. FEMS Microbiol Lett, 275(1):153-159, 2007.

[25] R. G. Gerlach, D. Jackel, B. Stecher, C. Wagner, A. Lupas, W. D. Hardt and M. Hensel. *Salmonella* Pathogenicity Island 4 encodes a giant non-fimbrial adhesin and the cognate type 1 secretion system. Cell Microbiol, 9(7):1834-1850, 2007.

[26] A. K. Bhunia. Salmonella enterica. In: Foodborne Microbial Pathogens: Mechanisms and Pathogenesis, Springer: pp. 201-215, 2008

[27] K. H. Hong and V. L. Miller. Identification of a novel *Salmonella* invasion locus homologous to Shigella ipgDE. J Bacteriol, 180(7):1793-1802, 1998.

[28] Leigh A. Knodler, B. Brett Finlay and O. Steele-Mortimer. The *Salmonella* Effector Protein SopB Protects Epithelial Cells from Apoptosis by Sustained Activation of Akt. J. Biol. Chem., 280(10):9058-9064, 2004.

[29] T. D. Lawley, K. Chan, L. J. Thompson, C. C. Kim, G. R. Govoni and D. M. Monack. Genome-wide screen for *Salmonella* genes required for long-term systemic infection of the mouse. PLoS Pathog, 2(2):e11, 2006.

[30] Alain Filloux, Abderrahman Hachani and S. Bleves. The bacterial type VI secretion machine: yet another player for protein transport across membranes. Microbiology, 154(1570-1583, 2008.

[31] H. M. B. Seth-Smith. SPI-7: *Salmonella*'s Vi-Encoding Pathogenicity Island. J Infect Developing Countries, 2(4):267-271, 2008.

[32] I. Virlogeux, H. Waxin, C. Ecobichon, J. O. Lee and M. Y. Popoff. Characterization of the *rcsA* and *rcsB* genes from *Salmonella typhi*: *rcsB* through *tviA* is involved in regulation off Vi antigen synthesis. J Bacteriol, 178(6):1691-1698, 1996.

[33] L. P. Randall and M. J. Woodward. The multiple antibiotic resistance (mar) locus and its significance. Res Vet Sci., 72(2):87-93 2002.

[34] I. Gentschev, G. Dietrich and W. Goebel. The *E. coli* alpha-hemolysin secretion system and its use in vaccine development. Trends Microbiol, 10(1):39-45, 2002.

[35] C. Latasa, C. Solano, J. R. Penades and I. Lasa. Biofilm-associated proteins. C R Biol, 329(11):849-857, 2006.

[36] K. Nishino, E. Nikaido and A. Yamaguchi. Regulation of multidrug efflux systems involved in multidrug and metal resistance of *Salmonella enterica* serovar Typhimurium. J Bacteriol, 189(24):9066-9075, 2007.

[37] S. Nagasawa, K. Ishige and T. Mizuno. Novel members of the two-component signal transduction genes in *Escherichia coli*. J Biochem, 114(3):350-357, 1993.

[38] A. L. Bishop, S. Baker, S. Jenks, M. Fookes, P. O. Gaora, D. Pickard, M. Anjum, J. Farrar, T. T. Hien, A. Ivens and G. Dougan. Analysis of the hypervariable region of the *Salmonella enterica* genome associated with tRNA (leuX). J Bacteriol, 187(7):2469-2482, 2005.

[39] Robert A. Edwards, Brian C. Matlock, Brian J. Heffernan and Stanley R. Maloy. Genomic analysis and growth-phase-dependent regulation of the SEF14 fimbriae of *Salmonella enterica* serovar Enteritidis. Microbiology, 147(2705–2715, 2001.

[40] M. Gerstein and R. Jansen. The current excitement in bioinformatics-analysis of whole-genome expression data: how does it relate to protein structure and function? Curr Opin Struct Biol, 10(5): 574-84, 2000.

[41] A. Lagreid, T. R. Hvidsten, H. Midelfart, J. Komorowski and A. K. Sandvik. Predicting gene ontology biological process from temporal gene expression patterns. Genome Res, 13(5): 965-79, 2003.

[42] R. C. Fink, M. R. Evans, S. Porwollik, A. Vazquez-Torres, J. Jones-Carson, B. Troxell, S. J. Libby, M. McClelland and H. M. Hassan. FNR is a global regulator of virulence and anaerobic metabolism in *Salmonella enterica* serovar Typhimurium (ATCC 14028s). J Bacteriol, 189(6): 2262-73, 2007.

[43] C. H. Chiu, P. Tang, C. Chu, S. Hu, Q. Bao, J. Yu, Y. Y. Chou, H. S. Wang and Y. S. Lee. The genome sequence of *Salmonella enterica* serovar Choleraesuis, a highly invasive and resistant zoonotic pathogen. Nucleic Acids Res, 33(5): 1690-8, 2005.

[44] D. H. Shah, M. J. Lee, J. H. Park, J. H. Lee, S. K. Eo, J. T. Kwon and J. S. Chae. Identification of *Salmonella gallinarum* virulence genes in a chicken infection model using PCR-based signature-tagged mutagenesis. Microbiology, 151(Pt 12):3957-68, 2005.

[45] G. S. Vernikos and J. Parkhill. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. Bioinformatics, 22(18): 2196-203, 2006.

[46] J. J. Bijlsma and E. A. Groisman. The PhoP/PhoQ system controls the intramacrophage type three secretion system of *Salmonella enterica*. Mol Microbiol, 57(1):85-96, 2005.

[47] J. Barchiesi, M. Espariz, S. K. Checa and F. C. Soncini. Downregulation of RpoN-controlled genes protects *Salmonella* cells from killing by the cationic antimicrobial peptide polymyxin B. FEMS Microbiol Lett, 291(1):73-79, 2009.

[48] E. Garcia Vescovi, F. C. Soncini and E. A. Groisman. Mg2+ as an extracellular signal: environmental regulation of *Salmonella* virulence. Cell, 84(1):165-74, 1996.

[49] E. A. Groisman. The pleiotropic two-component regulatory system PhoP-PhoQ. J Bacteriol, 183(6):1835-42, 2001.

[50] D. A. Pegues, M. J. Hantman, I. Behlau and S. I. Miller. PhoP/PhoQ transcriptional repression of *Salmonella typhimurium* invasion genes: evidence for a role in protein secretion. Mol Microbiol, 17(1): 169-81, 1995.

[51] S. I. Miller. PhoP/PhoQ: macrophage-specific modulators of *Salmonella* virulence? Mol Microbiol, 5(9): 2073-8, 1991.

[52] R. D'Ari, R. T. Lin and E. B. Newman. The leucine-responsive regulatory protein: more than a regulator? Trends Biochem Sci, 18(7): 260-3, 1993.

[53] J. M. Calvo and R. G. Matthews. The leucine-responsive regulatory protein, a global regulator of metabolism in *Escherichia coli*. Microbiol Rev, 58(3): 466-90, 1994.

[54] C. H. Baek, S. Wang, K. L. Roland and R. Curtiss, 3rd. Leucine-responsive regulatory protein (Lrp) acts as a virulence repressor in *Salmonella enterica* serovar Typhimurium. J Bacteriol, 191(4): 1278-92, 2009.

[55] K. Nishino, T. Honda and A. Yamaguchi. Genome-wide analyses of *Escherichia coli* gene expression responsive to the BaeSR two-component regulatory system. J Bacteriol, 187(5):1763-72, 2005.

[56] O. C. T, R. K. Carroll, A. Kelly and C. J. Dorman. Roles for DNA supercoiling and the Fis protein in modulating expression of virulence genes during intracellular growth of *Salmonella enterica* serovar Typhimurium. Mol Microbiol, 62(3): 869-82, 2006.

[57] M. W Mangan, S. Lucchini, V. Danino, T. O. Croinin, J.C. Hinton, C.J. Dorman. The integration host factor (IHF) integrates stationary-phase and virulence gene expression in *Salmonella enterica serovar* Typhimurium. Mol Microbiol. 59(6):1831-47, 2006.

[58] N. Buchmeier, S. Bossie, C. Y. Chen, F. C. Fang, D. G. Guiney, S. J. Libby. SlyA, a transcriptional regulator of Salmonella typhimurium, is required for resistance to oxidative stress and is expressed in the intracellular environment of macrophages. Infect Immun. 65(9):3725-30, 1997.