B-2994

Distribution and Analysis of SPIs and Prophage Insertions in a Diverse Set of *Salmonella* Serogroups Sequenced by SOLiD™ System Sequencing Technology

Gregory Govoni
Life Technologies
850 Lincoln Centre Dr.
Foster City, CA 94404
Gregory.Govoni@lifetech.com

Gregory Govoni[1], Craig A. Cummings[1], Lovorka Degoricija[1], Matthew L. Ranieri[2], Andrea Moreno Switt[2], Elena Bolchakova[1], Martin Wiedmann[2], Manohar Furtado[1]

[1] ***Applied Biosystems, 850 Lincoln Centre Drive, Foster City, CA, USA;*** [2] ***Cornell University, Ithaca, USA.***

ABSTRACT

Next generation sequencing technology provides researchers access to unprecedented amounts of genetic information. This genetic information can be parsed and analyzed for any number of disease determinants, including microbial pathogenicity. In this study, we sequenced whole genomes from 16 different serovars of *Salmonella enterica* subsp. *enterica* representing 13 different serogroups (D2, E1, E4, F, G, I, L, N, O, O:54, P, Q, and R). Mate-pair libraries for each genome were created and analyzed using the SOLiD™ sequencing system. From the data generated, a *de novo* assembly for each genome was constructed and analyzed for the presence of all 21 reported *Salmonella* Pathogenicity Islands (SPIs) as well as several LT2 prophages (fels-1 & -2, gifsy-1 & -2). For many of the SPIs, their presence was universal (SPIs 1-6, 9, 11-13). SPIs that were not found in this subset of strains, and seem to be serotype-specific include SPIs 10, 15, 17, and 19-21. Frequency of the LT2 prophage insertions varied between 31% and 62% of the strains analyzed. Further analysis finds that only the fels-2 insertion has been maintained in a similar context as in the LT2 genome. The other remaining prophage insertions have weak identity scores and lack consistent genomic context, which suggests these insertions may have occurred independently with distantly-related prophage. We also took a more detailed view of the genes that make up the SPIs. First, we looked for the presence or absence of every SPI gene orthologue and compiled a list of the absent in a table. Interestingly, several orthologues for known virulence factors (*avrA*, *sseA*, *sopB* and *STM3118*) were missing in at least 1 isolate. Second, we looked at sequence variations in several of these SPI virulence factors orthologues. While several had elevated overall distance averages (Dn), *sipD* had the highest variation with a Dn of 0.0543 (as compared to 0.004-0.006 for highly conserved genes). Preliminary phylogenetic analysis grouped these sequences into several clusters. Interestingly for the *sseD* orthologues, the Uganda and Senftenberg isolates shared their own independent cluster, while the Uganda and Give isolates clustered separately despite being from the same E1 serogroup. Since the Uganda and Senftenberg isolates both came from a bovine source and the Give isolate did not (human), these findings may have possible implications for discerning the molecular basis behind *Salmonella* strain host specificity.

INTRODUCTION

Salmonella enterica is a foodborne pathogen which causes an estimated 1.4 million human cases annually in the US (Mead et al.,1999). It is a highly diverse pathogen with more than 2,500 different serotypes of *Salmonella* described. The host range and disease severity varies depending on the serovar in question. For example, *Salmonella enteritidis* causes a limited gastroenteritis in a broad range of hosts, while *Salmonella typhi* can cause a systemic, typhoid fever disease specifically in primates. Many of the genes responsible for these variations in host range and disease severity are due to virulence factors found in *Salmonella* Pathogenicity Islands (SPIs). SPIs are genomic regions in the chromosome that have been horizontally acquired and are important for disease pathogenesis in one form or another. To date, 21 SPIs have been described in the literature. While many islands have been studied in-depth (such as SPI-1 and SPI-2), many SPIs have only been recently reported and are not well-characterized. In general, not much is known regarding the conservation of these SPIs across a broad range of serogroups. The development of new molecular--based technologies is making a comprehensive analysis of SPIs easier and faster to do. SOLiD™ next generation sequencing is an incredibly powerful tool that allows researchers to analyze a vast amount of genomic data in a relatively quick amount of time. Now it is possible to look in parallel at every SPI in a diverse set of strains in a matter of months without sacrificing on the level of detail required to do real in-depth analysis. In the study reported here, we describe the prevalence of each of these SPIs in the strains tested and perform some of the analyses that can be done to help discern which of these genes may play a role in determining host specificity and/or disease severity.

MATERIALS AND METHODS

Isolates. Sixteen *Salmonella* isolates were selected for whole genome sequencing. The isolates used in the study are listed in Table I along with tissue source, host, serovar derivation and serogroup. Isolates from bovine source are highlighted. In addition, the reference strains used to do the bioinformatic analyses are also listed and outlined in green.

Nucleic Acid Isolation. Nucleic acid was isolated from stationary phase cultures by phenolic extraction, and cleaned up by Qiagen (Valencia, CA) genomic tips. Twenty micrograms of isolated genomic DNA from each strain were used to start the mate-pair library construction.

Genome sequencing and assembly. Genomes were sequenced using the SOLiD™ system (Applied Biosystems, Foster City). Mate-paired libraries with approximately 1.5 kb inserts were constructed and deposited on one-eighth of a flowcell. Then, 25 bp reads were obtained from each of the F3 and R3 tags. After correcting errors in colorspace reads using a modified version of the spectral alignment tools from the EULER-USR package (Chaisson et al. 2009), *de novo* assembly was performed with the SOLiD™ *de novo* pipeline, which employs the Velvet assembly engine (Zerbino & Birney, 2008).

Scaffolds were aligned to two reference genomes (*S. Typhimurium* LT2 and *S. Enteritidis*) and concatenated into pseudogenomes. Scaffolds that did not match the chromosomes of the reference genomes were considered to be putative plasmids or strain specific transposable elements. Read coverage was approximately 120 - 180 fold per nucleotide.

Whole genome and SPI Sequence Analysis. Automated annotation was performed with the RAST server (http://rast.nmpdr.org; Aziz *et al.* 2008). The reference sequences used to delineate the SPIs and LT2 prophages are listed in Table II and were obtained from multiple sources such as the PAI database (Yoon *et al.* 2007). Individual ORF analysis was performed by taking ORF sequences from reference strain and querying them against the *Salmonella* genome assemblies with BLASTN. To ensure blast hits were in the proper genomic context, the sequence based comparison tool on the RAST server was used to align the pseudogenomes against select reference genomes.

Table I. List of Strains Used in this Study

Source	Host	Serogroup	Serovar	Strain
food	human	O:35 (O)	35:f,g:-	Adelaide
stool	human	O:35 (O)	35:z4,Z23:-	Alachua
stool	human	O:9,46 (D2)	9,46:a:enx	Baildon
stool	human	O:16 (I)	16:d:1,7	Gaminara
stool	human	O:3,10 (E1)	3,{10}{15}{15,34}:l,v:1,7	Give
stool	human	O:16 (I)	16:b:e,n,x	Hvittingfoss
n/a	human	O:38 (P)	38:k:1,6	Inverness
urine	human	O:40 (R)	1,40:b:e,n,x	Johannesburg
stool	human	O:21 (L)	21:b:e,n,x	Minnesota
stool	human	O:13 (G)	1,13,23:b:1,5	Mississippi
stool	human	O:54	{6,7,14}{54}:g,m,s:-	Montevideo
stool	human	O:11 (F)	11:r:e,n,x	Rubislaw
	bovine	O:1,3,19 (E4)	1,3,19:g,[s],t: -	Senftenberg
	bovine	O:3,10 (E1)	3,{10}{15}:l,z13:1,5	Uganda
stool	bovine	O:30 (N)	30:b:e,n,x	Urbana
stool	human	O:39 (Q)	39:b:1,2	Wandsworth
Ref. Strains from public databases		B	1,4,[5],12:i:1,2	Typhimurium str. LT2
		D1	9,12[Vi]:d:-	Typhi str. CT18
		C1	6,7:c:1,5	Choleraesuis str. SC-B67
		D1	1,9,12:-:-	Gallinarum str. 287/91
		O:62 (IIIa)	62:z4,z23:-	arizonae serovar 62:z4,z23:-

Table II. List of *Salmonella* Pathogenicity Islands (SPIs) and LT2 prophages Investigated in this Study

Pathogenicity Island or Prophage	Ref Strain	size (kb)	No. of CDS	Location 5'		Location 3'		insertion site
				coordinates	gene name/tag	gene name/tag	coordinates	
SPI-1	S. typhimurium (LT2)	34.9	47	**3005842**	ygbA/STM2860	TM2908	**3050120**	fhlA/mutS
SPI-2	S. typhimurium (LT2)	25.3	32	**1476167**	ssrB/STM1391	ssaU/STM1422	**1501489**	tRNA-valV
SPI-3	S. typhimurium (LT2)	30.4	12	**3954069**	STM3752	mgtC/STM3764	**3984421**	tRNA selC
SPI-4	S. typhimurium (LT2)	30.8	11	**4473213**	urvrA/STM4254	yjcC/STM4264	**4503982**	ssb/soxSR
SPI-5	S. typhimurium (LT2)	9.1	10	**1175321**	serT/STM1086	copR/STM1096	**1184389**	tRNA-serT
SPI-6	S. typhimurium (LT2)	36.8	33	**304665**	STM0266	STM0298	**341459**	tRNA-asp (CT18)
SPI-7	S. typhi (CT18)	133.3	143	**4409652**	STY4521	STY4680	**4542913**	tRNA-phe
SPI-8	S. typhi (CT18)	6.9	12	**3132530**	STY3273	STY3292	**3139414**	tRNA-phe
SPI-9	S. typhi (CT18)	15.7	4	**2743495**	STY2875	STY2878	**2759190**	?
SPI-10	S. typhi (CT18)	32.9	23	**4683605**	tRNA-leu/STY4821	STY4852	**4716538**	tRNA-leu
SPI-11	S. Choleraesuis (SC-B67)	15.7	18	**1350481**	SC1240	ysdB/SC1258	**1366166**	Gifsy-1 prophage
SPI-12	S. Choleraesuis (SC-B67)	11.1	11	**2354604**	tRNA-Pro	ccmD	**2365678**	tRNA-pro
SPI-13	S. typhimurium (LT2)	19.8	18	**3276146**	pheV/STM3116	STM3134	**3295923**	tRNA-pheV
SPI-14	S. typhimurium (LT2)	6.8	6	**926081**	STM0854	STM0859	**932913**	?
SPI-15	S. typhi (CT18)	6.4	5	**3053654**	STY3188	STYt047	**3060017**	tRNA Gly
SPI-16	S. typhi (CT18)	4.5	3	**605515**	STYt005	STY0607	**609992**	tRNA Arg
SPI-17	S. typhi (CT18)	5.1	2	**2460793**	STYt031	STY2631a	**2465914**	tRNA Arg
SPI-18	S. typhi (CT18)	2.6	2	**1454939**	hlyE/STY1498	STY1499	**1457500**	?
SPI-19	S. Gallinarum (287/91)	40.6	33	**1110982**	SG1023	SG1055	**1151559**	tRNA Ser
SPI-20	S. arizonae serovar 62:z4,z23:--	34.0	30	**2617493**	SARI_02707	SARI_02736	**2651483**	?
SPI-21	S. arizonae serovar 62:z4,z23:--	55.8	58	**2504768**	SARI_02578	SARI_02635	**2560524**	?
fels-1 prophage	S. typhimurium (LT2)	42.7	37	**962612**	STM0893	STM0929	**1004271**	attP-attP
fels-2 prophage	S. typhimurium (LT2)	33.7	46	**2844427**	STM2694	STM2739	**2877840**	attP-attP
gifsy-1 prophage	S. typhimurium (LT2)	48.5	53	**2728976**	gogB/STM2584	STM2636	**2776816**	attB-attP
gifsy-2 prophage	S. typhimurium (LT2)	45.8	48	**1098228**	STM1005	STM1052	**1141455**	attP-pseudo gene

References

Aziz RK, Bartels D, *et al*. 2008. The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics **9**:75.
Chaisson MJ, Brinza D, Pevzner PA. 2009. De novo fragment assembly with short mate-paired reads: Does the read length matter? Genome Res **19**(2): 336-46.
Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res vol. **14**(7):1394-40.
Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res vol. **18**(5):821-9.
Mead, P. S., L. Slutsker, V. Dietz, L. F. McCaig, J. S. Bresee, C. Shapiro, P. M. Griffin, and R. V. Tauxe. 1999. Food-related illness and death in the United States.Emerg. Infec. Dis. 5:607-625.
Yoon SH, Park YK, Lee S, Choi D, Oh TK, Hur CG, and Kim JF. 2007. Towards Pathogenomics: A Web-Based Resource for Pathogenicity Islands. Nucleic Acids Research 35:D395-D400.

Figure 1. CDS Analysis for all 21 SPIs in the newly Sequenced Strains

Island	Ref Genome	# of CDS	*Ad*	Ala	B	Gm	Gv	Hv	*In*	*Jo*	Mn	Ms	Mv	R	S	Ug	Urb	W	%>90%	%>50%
SPI-1	LT2	35	97%	100%	100%	97%	100%	100%	100%	100%	100%	97%	97%	97%	100%	100%	100%	100%	100%	100%
SPI-2	LT2	32	100%	100%	97%	100%	100%	100%	100%	100%	100%	97%	100%	100%	100%	100%	100%	100%	100%	100%
SPI-3	LT2	12	92%	100%	75%	92%	67%	100%	67%	67%	100%	92%	67%	67%	67%	67%	67%	67%	38%	100%
SPI-4	LT2	11	91%	91%	100%	91%	100%	91%	100%	91%	91%	91%	100%	100%	91%	91%	100%	91%	100%	100%
SPI-5	LT2	10	100%	100%	100%	100%	90%	100%	100%	90%	90%	90%	100%	100%	90%	100%	100%	90%	100%	100%
SPI-6	LT2	33	90%	90%	90%	59%	77%	92%	85%	79%	85%	77%	77%	82%	74%	82%	77%	90%	31%	100%
SPI-7	*CT18*	*143*	*18%*	*25%*	*4%*	*19%*	*10%*	*0%*	*52%*	*22%*	*0%*	*13%*	*25%*	*56%*	*26%*	*21%*	*77%*	*17%*	*0%*	*19%*
SPI-8	CT18	12	8%	50%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	75%	0%	0%	0%	0%	6%
SPI-9	CT18	4	100%	100%	100%	100%	100%	100%	75%	100%	100%	100%	100%	75%	100%	100%	100%	75%	81%	100%
SPI-10	CT18	23	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	17%	0%	0%	39%	0%	0%
SPI-11	*S Cho*	*18*	*50%*	*50%*	*56%*	*56%*	*50%*	*56%*	*50%*	*56%*	*50%*	*56%*	*44%*	*50%*	*61%*	*50%*	*44%*	*56%*	*0%*	*44%*
SPI-12	S Cho	11	73%	73%	100%	73%	73%	82%	82%	73%	73%	91%	73%	73%	73%	100%	73%	82%	19%	100%
SPI-13	LT2	18	61%	61%	94%	94%	100%	100%	94%	100%	94%	94%	94%	100%	61%	94%	94%	94%	81%	100%
SPI-14	LT2	6	0%	0%	83%	83%	83%	100%	83%	83%	83%	83%	83%	83%	0%	83%	83%	83%	6%	81%
SPI-15	CT18	5	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
SPI-16	*CT18*	*3*	*67%*	*67%*	*0%*	*0%*	*33%*	*33%*	*0%*	*0%*	*0%*	*100%*	*67%*	67%	*67%*	*0%*	*0%*	*0%*	6%	38%
SPI-17	CT18	2	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
SPI-18	CT18	2	100%	100%	100%	100%	100%	100%	100%	0%	100%	100%	100%	100%	100%	0%	0%	100%	81%	81%
SPI-19	*S Gal*	*37*	*22%*	*24%*	*27%*	*27%*	*27%*	*32%*	*30%*	*30%*	*27%*	*32%*	*30%*	*30%*	*32%*	*27%*	*27%*	49%	0%	0%
SPI-20	S Az	28	*39%*	*32%*	*50%*	*39%*	*39%*	*36%*	*43%*	*43%*	*39%*	*36%*	*46%*	*39%*	*36%*	*43%*	*39%*	*57%*	0%	6%
SPI-21	*S Az*	*42*	*26%*	*26%*	*33%*	*23%*	*23%*	*21%*	*28%*	*28%*	*23%*	*19%*	*30%*	*26%*	*19%*	*23%*	*26%*	*23%*	0%	0%
		No >90%	7	7	8	8	6	10	6	5	7	9	7	6	5	7	6	5		

> 90% of ORFs | 90% < ORFs > 50% | < 50% of ORFs | ORFs = 0%

Fig1. *Legend*: A combination of blast queries and the sequence based comparison tool on the RAST server was used to determine the presence or absence of every coding sequence (CDS) in every SPI. The % of CDSs present in each isolate relative to the number in the reference strain are reported for each SPI. Isolates with CDSs present in the SPI but do NOT appear to share the same genomic context as the reference strain are denoted with a box around the %. Strains with similar CDS order but with low CDS homology have percentages in *purple.*

Results: The preponderance of each SPI in this subset of isolates varies greatly. Ten islands are found in all the strains tested and include SPIs 1-6, 9, 11-13. Of those 11 SPIs, SPIs 1, 2, 4 and 5 are retained with over 90% of the reference ORFs present, suggesting there is evolutionary pressure to keep most of those islands intact. Islands that are not present in any of the isolates and can be classified as highly serover-specific include SPIs 10, 15, and 17. SPIs 19-21 are also most likely serovar specific, but may be present in these strains due to an independent horizontal acquisition of a related piece of DNA. SPIs 7 and 11 seem to be present in many of the strains tested but located in different regions of the genome. Strains with significant portions of SPI-7 include Inverness, Rubislaw, and Urbana.

Figure 2. Comparative representation of newly sequenced *Salmonella* genomes vs reference genomes

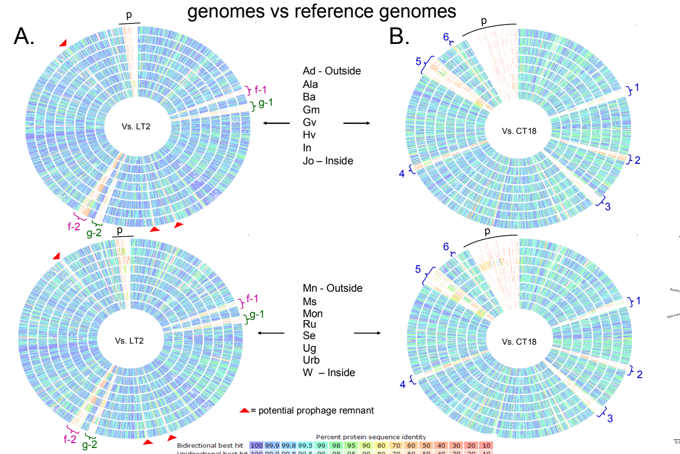


Fig 2. *Legend*: The sequence based comparison tool on the RAST servers was used to generate color-coded comparative representations of the *S. Typhimurium* LT2 (A) or *S. typhi* CT18 (B) genome versus the pseudogenomes from the 16 newly sequenced isolates. Each putative coding sequence (CDS) is color-coded according to its percent protein identity to the reference genome. In fig 2A, the 4 LT2 prophage insertions and several potential prophage remnants are denoted with symbols: f-1 (fels-1), f-2 (fels-2), g-1 (gifsy-1) g-2 (gifsy-2) and red triangles (prophage remnant). CDS associated with the extra-chromosomal plasmids are denoted with a “p”. Prophage elements associated with CT18 are numbered 1-6.

Results: From the analysis, none of the prophage insertions found in LT2 are fully conserved in the strains tested. Only the fels-2 prophage insertion came close to being conserved, as a high homology region covering ~50 – 60% of the fels-2 ORFs were found in the Ala, Mon, S and Urb isolates at the same location as in the LT2 reference strain. A similar size region was found for the gifsy-1 prophage insertion in the Hv, In, Ms, R and W isolates, but were not in the same genomic location as in the LT2 strains, which suggests that these insertions most likely occurred independently of the insertion in the reference strain. Similar observations were found for the 6 prophages identified in the CT18 strain with significant portions of prophage 5 region being found in Ala, In, Mon, S, and Urbana isolates. Since most insertions had at most 50-60% of the ORFs in the reference strain and an average ORF homology to the LT2 reference strain of between 90-95% (as compared to 99.1% for the average isolate genome), it is clear that these prophage insertions are decaying at a faster rate than most of the SPIs.

DISCUSSION

SOLiD™ next generation sequencing is an incredibly powerful tool that allows researchers to analyze a vast amount of genomic data in a relatively quick amount of time. Using this technology, we were able to sequence in one experimental run the whole genomes from 16 different *Salmonella* isolates representing 13 separate serogroups. With this data, it was possible to look at the evolutionary distance of multiple genomic markers, such as SPIs, across a broad spectrum of strains all at once. Previously, this type of in-depth analysis across a plethora of serogroups would not be possible in a timley manner. As part of the analyses, we were able to identify both large (such as loss of an ORF) and small (nucleotide sequence substitutions) genetic variations that may effect SPI function and possibly cause changes in disease severity and/or host specificity.

Figure 3A. Protein Percent Identity for each putative CDS in SPI-1

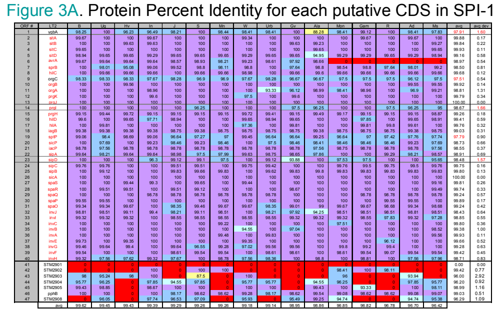


Fig3A. *Legend*: Analysis of the protein identity for each CDS in SPI-1 was performed and color coded. Genes reported confirmed to be important for pathogenicity/virulence are highlighted in red in LT2 column. The average identity and average deviation from mean were calculated for each ORF across all isolates (minus isolates where CDS was reported not present). CDSs with the lowest % identity to reference sequence (red) and highest avg deviation ([red-boxed]) are highlighted in avg and avedev columns. The average identity for all CDSs across the whole island was also calculated for each isolate (minus ORFs 41-47) and listed at the bottom of each column. Isolates were positioned from left to right in order of highest to lowest % identity across the entire island.

Results: The analysis shows that the SPI-1 region in the Baildon isolate has diverged the least from the LT2 reference SPI-1 (99.62% identity), while the SPI-1 region in the Mississippi isolate has diverged the most (96.42% or 98.89% not including *avrA*). Orthologues with the lowest homology to the LT2 reference sequence include *ygbB*, *orfC* and *sptB*. Interestingly, both *ygbA* and *orfC* have been shown NOT to be important for virulence in LT2, while *avrA* has. CDSs with the highest protein sequence variation include *ygbA*, *prgI* and *sipD*. CDSs corresponding to genes *STM2901* through *STM2908* are all missing in at least 1 isolate, and may suggest that they were inherited in the same horizontal acquisition event as the rest of the island, but are not important for pathogenicity and so are lost over time.

Figure 3B. List of SPI Orthologues Missing in at least 1 Isolate

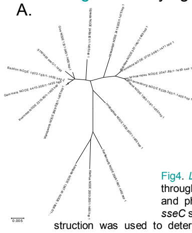
Island	SPI-1	SPI-2	SPI-3	SPI-4	SPI-5	SPI-6		SPI-7	SPI-8	SPI-9	SPI-10	SPI-11		SPI-12	SPI-13	SPI-14 thru SPI-21
Genes Missing in at least 1 Isolate	avrA	sseA	STM3752	STM4255	STM1089	STM0274A	STM0292	all	all	STY2878	all	SC1240	SC1249	tfa	STM3117	all
	STM2901		sugR		STM1093	STM0275	STM0293					sopB	SC1250	vtf	STM3118	
	STM2902		STM3754			STM0277	STM0294					SC1242	cspH	SC2248	STM3119	
	STM2903		rhuM			STM0283	STM2094.1N					mppA	pagD		STM3120	
	STM2904					STM0287	STM0295					SC1244			STM3121	
	STM2905					STM0288	STM0297					SC1245			STM3122	
	pphB					STM0290						SC1246			STM3123	
	STM2908					STM0291						icdA				

Fig3B. *Legend*: The analysis shown in figure 3A was performed on the multiple SPIs. Orthologues not present in at least 1 isolate in this study are listed here. Genes known to be required for pathogenesis in the reference strain are highlighted in red.

Results: Several genes associated with pathogenesis (avrA, *sseA*, *sopB*, and *STM3118*) are not present in at least 1 isolate. How that relates to disease severity or host specificity for that isolate needs to be characterize further.

Figure 4. Phylogenetic Analysis of Specific SPI Orthologues

A.



B.

		SPI-1					SPI-2				SPI-5
		prgJ	prgI	sptP	sipB	sipD	ssaS	sseC	sseD	sseF	sopB
(B)	LT2	1	1	1	1	1	1	1	1	1	1
O:3,10 (E1)	Ug	1	1	3	1	1	1	2	2	1	1
O:9,46 (D2)	B	1	1	1	3	1	1	1	1	1	1
O:16 (I)	Hv	1	1	1	2	1	1	3	1	1	1
O:38 (P)	In	1	1	1	2	1	1	1	1	1	1
O:39 (Q)	W	1	1	1	2	1	1	1	1	1	1
O:35 (O)	Ala	1	2	1	2	1	1	4	3	3	3
O:1,3,19 (E4)	S	2	2	1	2	1	1	4	2	3	4
O:35 (O)	Ad	2	2	1	1	1	1	4	3	3	3
O:16 (I)	Gam	2	1	2	1	2	2	1	1	1	2
O:21 (L)	Mn	2	1	2	1	2	2	2	1	1	2
O:13 (G)	Ms	2	3	3	1	2	3	2	4	4	1
O:54	Mon	2	1	2	1	2	1	2	5	2	2
O:30 (N)	Urb	1	1	2	1	2	2	2	5	2	2
O:40 (R)	Jo	1	1	2	1	2	2	2	5	2	2
O:11 (F)	R	3	3	2	1	2	2	2	1	1	2
O:3,10 (E1)	Gv	3	3	2	1	2	2	1	1	2	2
	overall avg distance	0.0043	0.0109	0.0229	0.0061	0.0546	0.0065	0.0283	0.0258	0.0190	0.0163

Fig4. *Legend*: Phylogenetic analysis was performed on 10 SPI orthologues scattered throughout SPIs 1, 2 and 5. Pairwise distance calculations (averages listed in fig4B) and phylogeny reconstruction using a neighbor-joining method (phylogeny tree for *sseC* shown in fig4A) were performed. Cluster Analysis the of the phylogeny reconstruction was used to determine which isolates cluster together. Clusters were numbered according to distance from reference strain LT2 (with 1 being in same cluster) and color coded as shown in Fig 4B.

Results: From the analysis, we were able to make 3 observations. 1. Similar gene cluster patterns can be found for several of the CDSs and suggests that many of these CDSs have co-evolved (such as *sseCDF* and *sopB* in isolates Alachua, Senftenberg and Adelaide). The *sseD* orthologues found in the Uganda and Senftenberg isolates have diverged in their own cluster from the other *sseD* genes. Since both of these strains were isolated from a bovine source, it may suggest that alterations in SseD help determine host specificity. 3. Isolates Uganda and Give have evolved independently, as demonstrated by an almost total lack of common clusters (only share *sipB* cluster), despite being from the same serogroup (E1).

CONCLUSIONS

- SPIs 1-5, 9, and 13 are universal among a wide range of serogroups.
- SPIs 10, 15, 17, and 19-21 seem to have a highly serovar-restricted profile.
- SPIs 7 and 11 seemed conserved in multiple serogroups but located in different genomic regions.
- Most Prophage insertions are random and are not conserved between strains.
- ORF analysis for each SPI may be helpful in re-defining pathogenicity island boundaries and in determining if a virulence factor has a serovar-restricted or a more universal phenotype
- Phylogenetic analysis of SPI CDSs may help determine which genes may be shaping disease outcome and host specificity by looking for individual or groups of virulence factors that have evolved together.