P-1833

Comparative Analysis of *the rfb* Locus, Encoding O-antigen Biosynthesis Genes in *Salmonella enterica*

Matthew L Ranieri[1], Andrea Moreno Switt[1], Henk C. den Bakker[1], Lavorka Degoricija[2], Craig A. Cummings[2], Greg Govoni[2], Elena Bolchacova[2], Manohar R. Furtado[2], Martin Wiedmann[1]

[1]Department of Food Science, Cornell University, Ithaca, NY 14853; [2]Life Technologies, Foster City, CA 94404

ABSTACT

Salmonella is an important pathogen, and serotyping has proved useful in understanding host specificity and in supporting epidemiological investigations. More than 2,500 serotypes of *Salmonella* have been identified using the Kauffman-White immunologic classification scheme, which is based on somatic (O) and flagellar (H) antigens. The O antigen is determined by an outer membrane lipopolysaccharide component, and currently 46 O serogroups of *Salmonella* are recognized. O antigens, which exhibit significant structural diversity due to variations in sugar composition, arrangements, and linkages between sugars, are encoded in the *rfb* gene cluster, which varies substantially between serotypes. Currently, *rfb* gene clusters of more common *Salmonella* serotypes are available, including serogroups O:2 (A), O:4 (B), O:7 (C1), O:8 (C2-C3), O:9 (D1), O:3,10 (E1), O:13 (G), O:17 (J) and O:18 (K). To expand our knowledge of the *rfb* locus and to support DNA-based approaches for serotyping, we used whole genome sequencing technology with the SOLiD™ system to analyze the *rfb* region of 16 less common human disease associated *S. enterica* subsp. *enterica* serotypes: Adelaide (serogroup O:35), Alachua (O:35), Baildon (O:9,46), Gaminara (O:16), Give (O:3,10), Hvittingfoss (O:16), Inverness (O:38), Johannesburg (O:40), Minnesota (O:21), Mississippi (O:13), Montevideo {O:6,7,14}{54), Rubislaw (O:11), Senftenberg (O:1,3,19), Uganda (O:3,10), Urbana (O:30), and Wandsworth (O:39). The *rfb* clusters ranged in size from 6.6 to 26.5 Kb and harbored 7 to 26 putative genes, the majority of which were related to sugar biosynthesis, sugar transfer, and O-antigen processing. GC content of the *rfb* clusters ranged from 34.3% to 49.0%, which is below the genome average for *Salmonella,* suggesting that recent transfer from different bacterial species may contribute to O-antigen diversity. Within *Salmonella* serogroups or across serogroups sharing a common antigenic factor, there was a high degree of similarity, especially with genes related to sugar biosynthesis. Comparisons among serogroups revealed considerably less homology in gene content. Overall, *rfb* cluster analysis will expand our knowledge of serogroup diversity and provide data for the development of molecular based serotyping methods.

INTRODUCTION

Salmonella enterica is a foodborne pathogen which causes an estimated of 1.4 million of human cases annually in the US (Mead et al., 1999). Currently, more than 2,500 different serotypes of *Salmonella* have been reported. Traditional serotyping aids in epidemiological studies, however, it has many limitations including production and quality control of hundreds of antisera, time limitations (it takes a minimum of 3 days to identify all antigens of a single isolate), and some strains are untypable (i.e. rough, mucoid). Molecular based serotyping methods targeting the *rfb* gene cluster, *fliC* and *fljB* genes responsible for O, H1 and H2 antigens, respectively, have been investigated to provide an alternative method to traditional serotyping, but have mainly focused on serotypes commonly associated with foodborne pathogen illnesses. Here we describe the analysis of 16 *rfb* gene clusters from uncommon serotypes.

Typically, the *rfb* region contains genes necessary for the biosynthesis of O-antigens, an important membrane component of Gram-negative bacteria. The O-antigen is a repeat unit polysaccharide comprised of O-unit repeats containing two to six sugar residues. As indicated by the high number of Salmonella serogroups (46), the O-antigens are quite variable. This variation is mainly because of order and linkage variation of different sugars within the polysaccharide. Genetic variation within the *rfb* region parallels the polysaccharide variabilty. The main genes coded for in the *rfb* region are involved in sugar biosynthesis, glycosyl transferases, and O-antigen processing genes.

MATERIALS AND METHODS

Isolates. Sixteen *Salmonella* isolates were selected for whole genome sequencing. These isolates represent the serotypes Adelaide, Alachua, Baildon, Gaminara, Give, Hvittingfoss, Inverness, Johannesburg, Minnesota, Mississippi, Montevideo, Rubislaw, Senftenberg, Uganda, Urbana, and Wandsworth (Table 1).

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 quarter 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 using 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 considered to be putative plasmids or strain specific transposable elements

Whole genome alignments. Automated annotation was performed with the RAST server (http://rast.nmpdr.org; Aziz et al. 2008) and whole genome alignments were performed using the Mauve Genome Alignment Software (Darling et al., 2004).

Table 1. Isolates chosen to represent uncommon serotypes for full genome sequencing, and results from genome analysis.

Isolate	Serotype	Serogroup	Genome Size (Mb)	Genome % GC	*rfb* gene cluster size (Kb)	*rfb* gene cluster %GC
A4-669	Adelaide	O:35 (O)	4.67	51.9	14.2	35.8
R6-377	Alachua	O:35 (O)	4.72	52.0	14.3	35.6
R6-199	Baildon	O:9,46 (D2)	4.75	52.0	23.9	39.8
A4-567	Gaminara	O:16 (I)	4.68	52.0	13.1	37.3
S5-487	Give	O:3,10 (E1)	4.61	52.1	16.9	38.6
A4-620	Hvittingfoss	O:16 (I)	4.74	51.9	23.0	43.6
R8-3668	Inverness	O:38 (P)	5.02	51.9	6.6	35.2
S5-703	Johannesburg	O:40 (R)	4.66	52.0	11.0	34.7
A4-603	Minnesota	O:21 (L)	4.61	51.9	13.5	34.3
A4-633	Mississippi	O:13 (G)	4.82	52.0	10.6	42.4
S5-403	Montevideo	O:6,7,14 (C1)	5.04	52.0	26.5	49.0
A4-653	Rubislaw	O:11 (F)	5.06	51.8	12.7	37.7
A4-543	Senftenberg	O:1,3,19 (E4)	5.01	51.8	14.4	39.5
R8-3404	Uganda	O:3,10 (E1)	4.8	51.9	16.9	38.6
R8-2977	Urbana	O:30 (N)	4.88	52.0	10.1	40.0
A4-580	Wandsworth	O:39 (Q)	4.86	52.0	18.0	40.2

Figure 1. Organization of 16 *Salmonella rfb* gene clusters, representing less common disease associated serotypes. Putative ORFS are represented by arrows, with corresponding assignment of the gene name or putative protein function. The serotype of each isolate is followed by serogroup, determined by the Kaufmann-White immunological typing scheme.

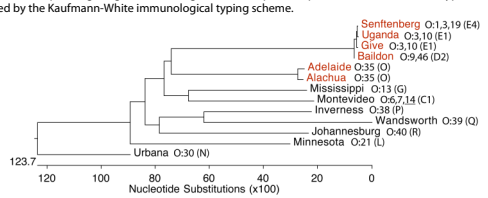

Figure 2. Neighbor joining phylogram of 13 *wzy* genes extracted from Salmonella *rfb* clusters. Sequences were aligned using the Clustal W algorithm in DNAStar (Madison, WI). For closely related *rfb* clusters (two groups highlighted in red) the *wzy* gene is highly conserved. Serotypes containing partial *wzy* ORFs were omitted from the analysis (Rubislaw, Gaminara, and Hvittingfos).

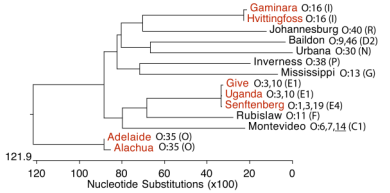

Figure 3. Neighbor joining phylogram of 14 *wzx* genes extracted from *Salmonella rfb* clusters. Sequences were aligned using the Clustal W algorithm in DNAStar (Madison, WI). For closely related *rfb* regions (three groups highlighted in red) the *wzx* gene is highly conserved. Serotypes containing partial *wzx* were omitted from the analysis (Minnesota and Wandsworth).

RESULTS AND DISCUSSION

Comparison of *Salmonella rfb* regions indicates variable gene content across serogroups, mirroring the diversity of *Salmonella* O-antigens. The number of genes within each *rfb* region was highly variable, ranging from 7 genes in Inverness to 26 in Montevideo (Figure 1). While the majority of genes in the *rfb* region were related to sugar biosynthesis, sugar transfer, O-unit export and O-unit polymerization, gene order and content within the *rfb* region differed considerably between serogroups. For example, *Salmonella* Baildon (O:9,46) contains sugar biosynthesis genes related to rhamnose, paratose and mannose synthesis, while *Salmonella* Johannesburg (O:40) only contains sugar biosynthesis genes related to mannose synthesis (Figure 1). Also, the *rfb* region of Johannesburg is 12.9 Kb smaller than the Baildon *rfb* region. In addition to differences in presence of sugar biosynthesis genes, the location of *wzx* and *wzy* is variable across different serogroups. In *Salmonella* Hvittingfoss the *wzx* and *wzy* genes are located at the 5' of the *rfb* gene cluster, while in Salmonella Adelaide (O:35) and Mississippi (O:13) the *wzx* and *wzy* genes are located in the central region of the *rfb* region. Further differences were found in *Salmonella* Montevideo and Inverness, as *wzx* and *wzy* genes are not adjacent to one another, but located at the start and end of the rfb region (Figure 1).

Serotypes exhibiting identical or similar O-antigens contain many homologous genes within the *rfb* cluster. Serotypes Give (O:3,10) and Uganda (O:3,10) were found to exhibit identical gene content within the *rfb* cluster (Figure 1). A closely related serotype, Senftenberg (O:1,3,10) is nearly identical to Give and Uganda, except that the Senftenberg *rfb* cluster contains one rearrangement of a hypothetical protein and lacks *wbaL*, which encodes an O-acetyl transferase related protein. This subtle difference in gene content highlights the impact that minor gene changes can have on phenotypic expression. *Salmonella* Adelaide and Alachua, both representing serogroup O:35, were found to have highly similar *rfb* regions, including conservation of order, content and size of genes. *Salmonella* Hvittingfoss (O:16) contained identical genes to Gaminara (O:16), but also contained an additional ten sugar biosynthesis genes and one hypothetical protein (Figure 1). A GC content analysis indicated that genes spanning from *wcaD* to *wcaI* (Figure 1) may not be native to the *rfb* region, as the GC content is as high as 70% in some genes and well above the *rfb* average of 43.6% GC. Additionally, a comparison of Montevideo *rfb* genes with a previously sequenced *Salmonella* Montevideo *rfb* cluster (Lee et al., 1992) indicated a number of gene differences. The Montevideo sequenced in this study (isolate S5-403) contains an additional 20 genes (uridine kinase C1 to *wcaE;* Fig 1). This putative insertion was included in one single contig during the genome assembly and showed a higher GC content (as high as 66% GC) as compared to other genes within the *rfb* region (typically 25% GC), suggesting that this gene cluster was correctly assembled and may have been introduced into *Salmonella* Montevideo S5-403 via horizontal gene transfer.

Further comparison of *rfb* clusters for uncommon serotypes with previously sequenced *rfb* clusters from identical serogroups revealed some other examples of *rfb* cluster diversification. *Salmonella* Baildon (O:9,46) and previously sequenced *Salmonella* Strasbourg (O:9,46) showed virtually identical organization and gene content of *rfb* cluster, except that Baildon strain R6-199 lacks *tyv*, which encodes a CDP–paratose epimerase involved in tyvelose synthesis (Xiang et al., 1994). A comparison of serotypes Give and Uganda (O:3,10) with previously sequenced *Salmonella* Anatum (O:3,10; Wang et al., 1992) indicated complete gene conservation and order within the *rfb* cluster.

Widely distributed *rfb* genes *wzx* and *wzy* represent possible targets for molecular seroypting. *wzx* and *wzy*, encoding for an O-antigen export unit and O-unit polymerase, respectively, are present in the *rfb* clusters associated with most *Salmonella* serogroups (except serogroups O:2(A), O:4(B), or O:9(D1). *wzx* and *wzy* are thus commonly used as targets for molecular serotyping (Herrera-Leon et al., 2006). Analysis of *wzy* sequences (Figure 2) indicates *wzy* genes from different serogroups from distinct clusters (with the exception of serogroups O:1,3,19, O:3,10, and O:9,46), which are highly conserved. Analysis of *wzx* (Figure 3) also indicates sequence conservation within serogroups, with wzx genes for different serogroups typically representing distinct clusters. Overall, these data support that specifically *wzx* is an appropriate initial target for serogroup classification, with use of additional genes (in particular wzy, but also *prt, abe* and glycosyltransferase encoding genes) providing for improved serogroup classification.

CONCLUSIONS

The *rfb* region gene content is largely conserved within serogroups, but varies considerably between serogroups with evidence for lateral gene transfer events in this region.

The O-antigen export and O-antigen polymerase genes, *wzx* and *wzy*, respectively, are conserved within most serogroups, although *wzx* appears to be a more suitable target for molecular serotyping approaches.

Full genome sequencing with de novo assembly represents a rapid method for analysis of gene content, even within highly variable genomic regions. The sequencing and annotation of diverse serotypes and serogroups will allow for the development of robust, reliable typing methods. Such analysis can contribute to understanding foodborne pathogen diversity and aid in developing improved tools to track and characterize pathogens.

REFERENCES

Chaisson MJ, Brinza D, Pevzner PA. 2009. De novo fragment assembly with short mate-paired reads: Does the read length matter? Genome Res. 19:336-46.

Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14:1394-40.

Herrera-Leon, S., R. Ramiro, M. Arroyo, R. Diez, M. Usera, N. Echeita. 2007. Blind comparison of traditional serotyping with three multiplex PCRs for the identification of Salmonella serotypes. Research in Microbiology. 158: 122-127.

Lee, S.J., Romano, L.K. and Reeves, P.R.. 1992. Cloning and structure of group C1 O antigen (*rfb* gene cluster) from *Salmonella enterica* serovar montevideo. Journal of General Microbiology 138: 305-312.

Mead, P. S., L. Slutsker, V. Dietz, F. L. McCaig, J. S. Bresee, C. Shapiro,P. M. Griffin, and R. V. Tauxe. 1999. Food-related illness and death in the United States. Emerg. Infect. Dis. 5:607–625.

Samuel, G. and Reeves, P.R. 2003. Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly. Carbohydrate Research. 338: 2503-2519.

Wang, L., Romana, L.K. and Reeves, P.R. 1992. Molecular analysis of a *Salmonella enterica* group E1 *rfb* gene cluster: O antigen and the genetic basis of the major polymorphism. Genetics. 130: 429-443.

Xiang, S.H, Hobbs, M. and Reeves, P.R. 1994. Molecular analysis of the *rfb* gene cluster of a group D2 *Salmonella enterica* strain: evidence for its origin from an insertion sequence-mediated recombination event between group E and D1 strains. Journal of Bacteriology 176: 4357-4365,

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res.. 18:821-9.