International Journal of Infectious Diseases 49 (2016) 134–140

Contents lists available at ScienceDirect

International Journal of Infectious Diseases

journal homepage: www.elsevier.com/locate/ijid

Genotyping of *Salmonella* with lineage-specific genes: correlation with serotyping

Qing-Hua Zou [a], Ren-Qing Li [b], Gui-Rong Liu [c,d], Shu-Lin Liu [c,d,e,*]

[a] Department of Microbiology, School of Basic Medical Sciences, Peking University Health Science Center, Beijing, China
[b] Institute of Immunology, Beijing Center for Disease Control and Prevention, Beijing, China
[c] Genomics Research Center, Harbin Medical University, 157 Baojian Road, Harbin, 150081, China
[d] HMU-UCFM Centre for Infection and Genomics, Harbin Medical University, Harbin, China
[e] Department of Microbiology, Immunology and Infectious Diseases, University of Calgary, Calgary, Alberta, Canada

ARTICLE INFO

Article history:
Received 14 January 2016
Received in revised form 25 May 2016
Accepted 27 May 2016

Corresponding Editor: Eskild Petersen, Aarhus, Denmark.

Keywords:
Salmonella
Serotyping
rpoB
MLST
Serotype-specific gene

SUMMARY

Background: The bacterial genus *Salmonella* encompasses a large number of serotypes that are genetically very similar but biologically quite different, especially in pathogenic properties and host specificity. Serotyping has been used for the classification, identification, and epidemiological investigation due to its excellent discriminating power, but it cannot distinguish the different pathogenic lineages within a polyphyletic serotype. Additionally, very few institutions have the comprehensive set of antisera for typing. Therefore various studies have been performed to explore alternative assays to differentiate *Salmonella* isolates, such as the search for genes that can be used as potential molecular substitutes for serotyping. However, the genes tested so far have often given inconsistent results.

Methods: In this study, the discriminating power of seven genes to differentiate 309 *Salmonella* strains representing 26 serotypes was evaluated and the results were compared with those of other methods.

Results: The seven newly selected genes have a good power to differentiate different serovars. The tree based on the concatenated sequences of these genes revealed phylogenetic relationships of the bacteria consistent with that of the whole genome tree.

Conclusion: Individual *Salmonella* lineages each have specific genes that can be used to differentiate *Salmonella* isolates on a phylogenetic basis.

1. Introduction

Since the divergence from a common ancestor with *Escherichia coli* more than 100 million years ago,[1] *Salmonella* have developed into more than 2600 serotypes that cause a variety of illnesses in humans and other animals. All *Salmonella* lineages are closely related, as revealed initially by DNA–DNA re-association assays[2] and then by physical mapping[3] and genomic sequencing.[4–9] Notwithstanding the observed genetic relatedness among the *Salmonella* lineages, they differ profoundly in host range and pathogenic features,[10,11] causing clinical consequences ranging from no obvious disease to mild gastroenteritis to potentially fatal systemic infections such as typhoid fever in humans. Therefore, the timely and accurate identification of *Salmonella* isolates is of great clinical significance.

Various typing methods, such as multilocus sequence typing (MLST), pulsed-field gel electrophoresis (PFGE), and phage typing, have been developed to discriminate these closely related bacteria,[12–16] of which serotyping has been the most widely used assay owing to its excellent discriminating power. However, serotyping has multiple disadvantages, such as low throughput, high expense, the need for expertise, and the requirement of a comprehensive set of antisera, which is not available to many institutions. Most importantly, however, is the fact that many *Salmonella* serotypes are polyphyletic, containing more than one phylogenetic (and pathogenic) lineage. As a result, alternative methods have been attempted for discriminating *Salmonella*, such as PFGE, ribotyping,[17] sequencing of H antigen genes (*fliC* and *fljB*),[18] and 16S–23S rRNA spacer restriction fragment length polymorphism (RFLP).[19] These methods may yield important information for epidemic analysis, but their discriminating ability is usually insufficient for accurate identification. MLST is an excellent method to discriminate strains based on their sequence

* Corresponding author. Tel.: +86 45186669150; fax: +86 45187502720.
E-mail address: slliu@ucalgary.ca (S.-L. Liu).

http://dx.doi.org/10.1016/j.ijid.2016.05.029

differences at seven house-keeping loci and has been used for the molecular typing of many bacteria.[20,21] Recently, MLST has been proposed for *Salmonella* typing due to its high resolution performance in delineating the serotypes.[12] Other genes such as *rpoB* have also demonstrated utility in *Salmonella* identification.[22,23] However, no sets of genes hitherto reported have shown a discriminating power similar to that of serotyping.

In a previous study by the present study group, 27 *Salmonella* genomes (which were completed before March 2012) were compared and the polymorphisms of a selected set of genes were analyzed, including some conserved genes (i.e., genes common to all compared genomes) and some from genomic islands, among the different serotypes.[24] It was found that 10 of the analyzed genes were polymorphic among most of the serotypes compared and thus it was considered that they may be useful in delineating *Salmonella*. In the present study, a phylogenetic analysis of seven selected genes was conducted and comparisons were made with the *rpoB* gene and the seven MLST genes to examine their discriminating power among the different *Salmonella* serotypes. A series of additional genes present only in individual serotypes were also evaluated and it was found that the combined use of the genes could significantly improve gene-based *Salmonella* typing.

2. Materials and methods

2.1. Retrieval of gene sequences

Among the 10 highly conserved genes identified in the previous study that were polymorphic across most of the 15 serotypes analyzed,[24] three, i.e., *STM2379*, *cpsG*, and *STM4261*, were not included in this current study, because *STM2379* is a pseudogene in *Salmonella* Heidelberg B182, *cpsG* has duplicates in several genomes, and *STM4261* is 16 680 bp, which is too long to be of practical use. To find *Salmonella* strains with all of the seven selected genes (*nuoG*, *srfC*, *napA*, *yhgE*, *priA*, *cpdB*, and *entF*), *rpoB*, and the seven MLST genes (*aroC*, *dnaN*, *hemD*, *hisD*, *purE*, *sucA*, and *thrA*) being sequenced, a BLAST search of the nucleotide sequences of each of these genes in *Salmonella* Typhimurium LT2 was performed against the NCBI non-redundant and whole-genome shotgun contigs database, and only strains that had all of these genes sequenced were picked up for further analysis.

2.2. Construction of phylogenetic trees

Nucleotide sequences of the seven selected genes, *rpoB*, and the seven MLST genes were aligned using ClustalW in BioEdit software with default parameters. Phylogenetic trees were constructed by neighbor-joining method with MEGA software (version 5.0). The reliability of the neighbor-joining trees was estimated by bootstrap analysis with 1000 replicate datasets. The substitution model was 'maximum composite likelihood', substitutions included 'transitions + transversions', rates among sites were set as 'uniform', the pattern among lineages was set as 'same' (homogeneous), and the gaps/missing data treatment was 'complete deletion'.

The core genome tree of 32 *Salmonella* strains representing 18 serotypes (Supplementary Material Table S1) was constructed on coding sequences common to all compared genomes using the all-blast-all program in the NCBI Basic Local Alignment Search Tool (BLAST), with the criteria set at identity >75% and e-value <1e−10. For each query sequence, only the highest-scoring match above the defined identity and e-value cut-off in the 32 genomes was retained. Matched genes were then made into clusters using a Perl script. Genes present in all 32 genomes were aligned using ClustalW 2.1 and were concatenated to construct the core genome for each strain. A phylogenetic tree based on the core genome was constructed, as described above.

2.3. Lineage-specific genes

Genes present only in all analyzed strains (Supplementary Material Table S1) of one particular *Salmonella* lineage but absent in all the other lineages were considered lineage-specific. For further confirmation, all identified lineage-specific genes were searched against the NCBI non-redundant database using BLAST to exclude genes that had homologues in other *Salmonella* lineages.

2.4. Identification of clinical strains

Clinical strains of *Salmonella* were single-colony isolated and cultured in Luria-Bertani (LB) broth; DNA was extracted with a DNA extraction kit (Sangon Biotech, China). Primers for amplifying the segments of the selected genes were synthesized by Sangon Biotech, China. PCR fragments were sequenced using the Sanger AB3130 platform and the phylogenetic trees were constructed using MEGA 5.0.

3. Results

A total of 309 *Salmonella* strains representing 26 serotypes and having all the seven selected genes, *rpoB*, and the seven MLST genes being completely sequenced in the NCBI database were included in this study. In this analyzed collection, *Salmonella* Enteritidis was the predominant serotype (n = 86), followed by *Salmonella* Agona (n = 66) and *Salmonella* Montevideo (n = 37). Other serotypes were represented by 1 to 32 strains (Table 1).

3.1. The performance of the rpoB gene, the individual MLST genes, and the newly selected genes in distinguishing different Salmonella serotypes

The strains representing 20 serotypes formed serotype-specific branches on the *rpoB* gene tree, with the remaining six serotypes not well discriminated (Figure 1A). While it was not surprising to see *Salmonella* Enteritidis and *Salmonella* Gallinarum clustered together and *Salmonella* 4,[5],12:i:- being mixed with *Salmonella* Typhimurium, as judged by their phylogenetic relationships, it was unexpectedly found that strains of *Salmonella* Newport formed two separate clusters and the two strains of *Salmonella* Saintpaul did not cluster together.

Among the seven MLST genes, *sucA* distinguished strains representing 19 serotypes, with the remaining six genes, *aroC, thrA, hemD*, *dnaN*, *hisD* and *purE*, distinguishing strains representing 16, 18, 17, 16, 16, and 15 serotypes, respectively (Supplementary Material Figure S1, Table 1).

With the exception of *nuoG*, which only distinguished strains representing 13 serotypes, the other six genes selected in this study discriminated the 309 strains very clearly. The clustering correlated well with serotyping (Supplementary Material Figure S1, Table 1). Eighteen to 20 serotypes could be discriminated successfully by these genes.

Salmonella Saintpaul, *Salmonella* 4,[5],12:i:-, *Salmonella* Newport, and *Salmonella* Typhimurium were the least discriminated serotypes for *rpoB*, the MLST genes, and the seven newly selected genes, although the seven newly selected genes could clearly discriminate strains of *Salmonella* Dublin, *Salmonella* Heidelberg, and *Salmonella* Gallinarum/Pullorum (Supplementary Material Figure S1, Table 1).

3.2. Performance of the concatenated sequences of the selected genes to reflect phylogenetic relationships of the Salmonella serotypes

While examining the combined performance of the concatenated sequences, trees were constructed and it was found that the