ASM Sc. J., 14, 2021

https://doi.org/10.32802/asmscj.2020.602

Molecular Serotyping by Phylogenetic Analyses of a 1498bp Segment of the *invA* Gene of *Salmonella*

Rance Derrick Neri Pavon[1] and Windell L. Rivera[1*]

[1]*Pathogen-Host-Environment Interactions Research Laboratory, Institute of Biology, College of Science, University of the Philippines Diliman, Quezon City 1101, Philippines*

The current gold standard for *Salmonella* serotyping is costly, labor-intensive and time-consuming. However, proper identification is key to monitor *Salmonella* transmission and implementation of necessary control measures. The onset of advanced molecular techniques has lessened resource and labor requirements; however, it still remains complex, unestablished and plagued with insufficiencies. Hence, a simpler serotyping method with sufficient resolution is needed. In this study, the *invA* virulence gene, associated with *Salmonella* invasion into host cells and is considered as a marker for *Salmonella* detection, was amplified and sequenced among isolates from meat samples in Metro Manila, Philippines. This was followed by sequence alignments with reference sequences (Refseqs), oversaturation and model tests, phylogenetic tree analyses and signal detections. Unfortunately, alignment of a 229bp amplified and sequenced *invA* gene segment with Refseqs generated little to no base variations and consequently provided insufficient phylogenetic resolution for molecular serotyping (0 of the 17 serotypes tested). However, another segment of 1498bp, outside the amplified region, showed considerable base variation in alignment and consequently resolved a maximum of 13 out of 17 (76.47%) serotypes tested, all generated trees considered. These suggest the potential of the *invA* virulence gene as a single-gene marker for molecular serotyping of *Salmonella* through phylogenetic analyses.

Keywords: *Salmonella;* base variation; *invA* gene; phylogenetic analysis; serotyping; taxonomy

I. INTRODUCTION

Salmonella is one of the main causes of foodborne illnesses worldwide (Pal *et al.,* 2015), causing numerous clinical manifestations depending on host variables and serotypes (WHO, 2018). In the Philippines, *Salmonella* is one of the leading causes of food poisoning outbreaks (Azanza *et al.,* 2019) and its high incidences have been previously reported in food animal products from abattoirs and wet markets of Metro Manila (Calayag *et. al.,* 2017; Ng & Rivera, 2015; Paclibre *et al*., 2017). At present, *Salmonella* is divided into two species, namely *S. enterica* and *S. bongori*, with the former further divided into six subspecies, namely *enterica* (I), *salamae* (II), *arizonae* (IIIa), *diarizonae* (IIIb), *houtenae* (IV) and *indica* (VI) (Porwollik *et al.,* 2004). There are more than 2,500 serotypes (CDC, 2019; WHO, 2018) with 2,300 belonging to subsp. *enterica* (I) alone, constituting all forms of salmonellosis among humans and animals (Porwollik *et. al.,* 2004). However, only a handful of serotypes is associated with serious diseases (e.g. Typhi, Typhimurium, Enteritidis) (Bell *et. al.,* 2016; Porwollik *et al.,* 2004). Hence, the challenges of identification and classification hinder progress toward the epidemiological control and prevention of outbreaks from these pathogens.

Traditional serotyping remains as the gold standard for *Salmonella* identification. However, numerous and large volumes of antisera are required, making the method costly, labor-intensive and time-consuming (McQuiston *et. al.,* 2011; Seong *et al.,* 2012). In addition, there is also variability in

*Corresponding author's e-mail: wlrivera@science.upd.edu.ph

antigen expression (Barco *et al.,* 2011). The advent of molecular tools attempted to provide alternatives such as single-gene analyses using non-protein coding genes like 16S rRNA (Trkov & Avgus˘tin, 2003), yet still suffer from insufficient resolution and discordance (Fox *et. al.,* 1992; Větrovský & Baldrian, 2013) and housekeeping genes such as *rpoB* (Case *et al.,* 2007), that have incongruences, inconsistencies (Christensen, 2004; Glaeser & Kämpfer, 2015) and difficulty in primer design (Fukushima *et al.,* 2002). Similarly, molecular serotyping using antigen-coding genes which has been applied in the Philippines (Ng & Rivera, 2015) provides better resolution, requires multiple primer sets, and is limited to only a few serogroups or serotypes (APHL, 2014). More recently, the use of multilocus sequence typing (MLST) provides high discriminatory power and may serve as a possible tool in investigating *Salmonella* outbreaks (Kotetishvili *et al.,* 2002). However, it requires analyses of multiple genes involving numerous reactions (Seong *et al.,* 2012) and in some cases, still plagued with insufficient discriminatory power (Fakhr *et al.,* 2005). The development of whole genome sequencing (WGS) provided more in-depth insights for pathogen evolution, transmission and outbreak surveillance. Studies have shown that WGS (e.g. SeqSero) can be a reliable and rapid tool for *Salmonella* serotyping with better performance and resolution than traditional methods in serotype predictions and antigenic discrimination (Diep *et al.,* 2019; Ibrahim & Morin, 2018). On the other hand, Check-Points, a company based in the Netherlands, developed Check & Trace Salmonella (CTS) using DNA microarray technology. Similarly, CTS revealed comparable predictions with traditional methods for most of the prevalent serotypes with higher agreement than even WGS (SeqSero) (Diep *et al.,* 2019). However, these methods still showed discrepancies and identification failure, possibly due to database gaps.

The pathogenicity of *Salmonella* is defined by pathogenicity islands (SPIs). These SPIs have numerous virulence genes to allow invasion and proliferation inside host cells. *invA*, among other genes, has been shown as the basis of invasive phenotypes among pathogenic *Salmonella* (Clark *et. al.,* 1998; Fàbrega *et. al.,* 2009; Galán & Curtiss, 1989). *invA* has also been demonstrated as a specific marker for rapid *Salmonella* detection through polymerase chain reaction (PCR) (Heymans *et. al.,* 2018; Rahn *et. al.,* 1992; Shanmugasamy *et al.,* 2011) due to its wide distribution among serotypes (Galán & Curtiss, 1991; Nolan *et al.,* 1995) and specificity (De Clercq *et. al.,* 2007; Rahn *et al.,* 1992). Despite extensive studies on its detection and distribution among *Salmonella*, it has yet to be considered for molecular serotyping. For these reasons, this study analyzed gene sequences based on PCR marker for *invA* (Chiu, 1996) through phylogenetic analyses with available reference sequences (Refseqs) of *Salmonella* species, subspecies and *S. enterica* subsp. *enterica* (I) serotypes from databases to evaluate the delineation ability and resolution of the *invA* gene for *Salmonella* serotyping involving isolates from meat samples obtained from markets and abattoirs in Metro Manila, Philippines.

II. MATERIALS AND METHODS

A. Isolation of Salmonella

Three *Salmonella* isolates obtained from raw porkchop cuts from Alabang (sample APC1R1A) and Pasay (sample PPC1R1A) wet markets, and tonsil from slaughtered swine from Kayang abattoir, Pasay (sample 11) were randomly selected. Samples were placed in sterile zip-lock bags upon collection from site and into a cooler and brought to the laboratory to process. Based on standard protocols (Andrews *et al.,* 2019), 25 g of each sample was aseptically weighed, minced and placed inside Whirl-Pak® bags. Then, 225 ml of buffered peptone water was added. After incubation for 18-24 h at 37°C, 100 µl of the solution was then transferred to 10 ml of Rappaport-Vassiliadis (RV) broth. After incubation for 18-24 h at 42°C, 10 µl was plated and streaked onto xylose lysine deoxycholate (XLD) agar. After another 18-24 h of incubation at 37°C, black colonies that grew on red-coloured media were sub-cultured to nutrient agar (NA) for further confirmation and analysis.

B. DNA Extraction and Polymerase Chain Reaction (PCR)

Colonies (2 to 3) of each isolate from NA were transferred into 50 µl Tris-EDTA buffer and subjected to boil-lysis DNA extraction method (100°C for 10 min). The mixture was then subjected to centrifugation at 2,656 x g for 5 min and the supernatant (containing the DNA) was then transferred to a

clean microfuge tube. *Salmonella* confirmation was done through *invA* gene detection (Chiu, 1996) with primers F-ACAGTGCTCGTTTACGACCTGAAT and R-AGACGACTGGTACTGATCTAT optimized for PCR using Soguilon-Del Rosario & Rivera's (2015) conditions: initial denaturation at 95°C for 2 min, followed by 30 cycles of denaturation at 95°C for 30 sec, annealing at 60°C for 30 sec, and extension at 72°C for 30 sec with final extension at 72°C for 5 min. Each PCR reaction consisted of 12.5 µl of 2× GoTaq® Green Master Mix (Promega), 1 µl each of 10 µM forward and reverse primers of *invA*, 8.5 µl of sterile nuclease-free water, and 2µl DNA template. *invA* gene positive result was based on the visualization of a 244bp product with UV transilluminator after electrophoresis with 1.5% agarose gel stained with SYBR® Safe.

C. DNA Sequencing and Contig Assemblies

invA PCR products and their corresponding primers were submitted to Macrogen, Inc. (South Korea) for sequencing. The resulting sequences were analyzed and processed with Geneious Prime® 2020.0.3 (https://www.geneious.com/). Forward and reverse sequences were subjected to *de novo* assembly and consensus sequences for all three isolates were generated based on base call-quality also in Geneious Prime® 2020.0.3. The consensus sequences obtained were then subsequently aligned with reference sequences and subjected to phylogenetic analyses.

D. Reference Sequences and Alignment

Reference sequences (Refseqs) of *invA* were obtained from the National Center for Biotechnology Information (NCBI) – GenBank database. Available *invA* sequences for *S. bongori* and different *S. enterica* subspecies and serotypes were downloaded in FASTA format. Sequences were aligned and trimmed using Geneious Prime® 2020.0.3. Two datasets were generated. For *invA* gene with the three isolates, all sequences were trimmed based on the smallest sequence size. For *invA* gene without samples, Refseqs were trimmed based on the variable sites among serotypes observed in the software. Alignments were then exported as FASTA format for further analyses. The full list of sequences and their accession numbers are shown in Table 1.

E. Test for Oversaturation

Sequences were tested for the presence of extreme substitution saturation to determine whether models can still correct for multiple hits (Morrison, 2006). PAUP* (Swofford, 2002) command prompt was used to generate uncorrected and corrected distances. Similarly, transitions and transversion distances were also generated. Uncorrected and corrected distances were used to generate a scatter plot, which if formed, a curve or showed a linear increase, then the dataset would be considered unsaturated and thus could proceed to other analyses, but if the plots showed a plateau, then models can no longer correct for multiple hits (Strimmer & von Haeseler, 2003). Similarly, a plot generated from uncorrected distances against transition and transversion distances would show that the dataset could still be corrected for multiple hits if both transitions and transversions were increasing linearly and that transitions were still above transversions. However, if a plateau was evident, then the datasets were deemed oversaturated (Brown *et al.,* 1982). Xia test (Xia *et al.,* 2003; Xia & Lemey, 2009) at 60 replicates was also conducted using DAMBE 7.2.1 (Xia, 2018) to test the amount of saturation among the datasets. This can be obtained by computing for the index of substitution saturation (I_{SS}). The I_{SS} is then compared to critical values ($I_{SS.c}$) for the dataset based on completely symmetrical and completely asymmetrical trees. The I_{SS} should ideally be significantly lower than the critical value for the dataset to be considered with little saturation. After oversaturation tests, the best-fit models for datasets that do not suffer from oversaturation were then determined.

F. Substitution Model Selection with jModeltest

jModeltest 0.1.1 by Posada (2008) was implemented. Exported FASTA alignments were converted to NEX format using DAMBE 7.2.1 (Xia, 2018). The resulting format was then subjected to likelihood scores computation under 88 candidate models with 11 substitution schemes, including both equal/unequal base frequencies, with/without a proportion of invariable sites (+I) and rate variation among sites (+G), using an ML tree as the base tree. Akaike information criterion (AIC) determined the best-fit model and the generated base frequencies, substitution rates, and if

applicable, gamma shape value, number of substitution categories and proportion of invariable sites were then used under various command prompts for phylogenetic analyses.

G. Phylogenetic Analyses with Command Prompts

Phylogenetic analyses with distance-based method using Neighbor-Joining (NJ) tree (Saitou & Nei, 1987), character-state methods using Maximum Parsimony (MP) (Eck & Dayhoff, 1966; Fitch, 1977), Maximum Likelihood (ML) (Cavalli-Sforza & Edwards, 1967; Felsenstein, 1981) and Bayesian Inference (BI) (Yang, 1997) were used to generate phylogenetic trees. PAUP* version 4b10 (Swofford, 2002) command prompt was used to generate NJ and MP trees with bootstrap support of 1,000 replicates. PhyML version 3.0 (Guindon *et al.,* 2010) was used to generate ML trees (using PHY formatted alignments also converted from DAMBE 7.2.1) with 1,000 replicates for boostrap support. Lastly, MrBayes version 3.1.2 (Ronquist & Huelsenbeck, 2003) was used to generate BI trees with 10,000,000 generations and posterior probabilities for support. NJ and ML trees were drawn using TreeExplorer version 2.12 (Tamura, 1997) while MP and BI trees were drawn using TreeView 1.6.6 (Page, 2002). The outgroup selected for the phylogenetic analyses is *S. bongori* since it is a different species of *Salmonella*. All subspecies and serotypes in this study belong to the species, *S. enterica*. However, for the serotypes of *S. enterica* subsp. *enterica*, the nearer outgroups/relatives are of other *S. enterica* subspecies.

H. Detection of Phylogenetic Signal

To know whether the variations (especially among *Salmonella* serotypes) evaluated through phylogenetic analyses are; due to underlying phylogenetic signal and not by chance, MP analysis was applied to all datasets (*invA* with samples, *invA* without samples). Again, using PAUP* version 4b10 (Swofford, 2002), a random set of 10,000 trees were generated through MP. If the tree lengths of all tree topologies generated follow a normal distribution, then the differences are simply by chance. However, if the distribution is skewed enough, then it suggests the existence of a phylogenetic signal. This is further supported by a g1 test based on Hillis & Huelsenbeck (1992), in which if g1 scores obtained are smaller than the critical values depending on the number of parsimony-informative sites and number of taxa, then a phylogenetic signal is exhibited.

Table 1. *Salmonella* reference sequences of *invA* gene and accession numbers

No.	Species	Subspecies	Serovar	Strain	Accession No.
1	*S. enterica*	*enterica* (I)	Agona	SL483	NC 011149
2	*S. enterica*	*enterica* (I)	Cubana	CFSAN002050	NC 021818
3	*S. enterica*	*enterica* (I)	Choleraesuis	SC-B67	NC 006905
4	*S. enterica*	*enterica* (I)	Dublin	CT 02021853	NC 011205
5	*S. enterica*	*enterica* (I)	Enteritidis	P125109	NC 011294
6	*S. enterica*	*enterica* (I)	Gallinarum	287/91	NC 011274
7	*S. enterica*	*enterica* (I)	Gallinarum/Pullorum	CDC1983-67	NC 022221
8	*S. enterica*	*enterica* (I)	Gallinarum/Pullorum	RKS5078	NC 016831
9	*S. enterica*	*enterica* (I)	Heidelberg	B182	NC 017623
10	*S. enterica*	*enterica* (I)	Heidelberg	CFSAN002069	NC 021812

11	*S. enterica*	*enterica* (I)	Heidelberg	SL476	NC 011083
12	*S. enterica*	*enterica* (I)	Javiana	CFSAN001992	NC 020307
13	*S. enterica*	*enterica* (I)	Newport	SL254	NC 011080
14	*S. enterica*	*enterica* (I)	Paratyphi A	ATCC 9150	NC 006511
15	*S. enterica*	*enterica* (I)	Paratyphi C	RKS4594	NC 012125
16	*S. enterica*	*enterica* (I)	Pullorum	S06004	NC 021984
17	*S. enterica*	*enterica* (I)	Schwarzengrund	CVM19633	NC 011094
18	*S. enterica*	*enterica* (I)	Thompson	RM6836	NC 022525
19	*S. enterica*	*enterica* (I)	Typhi	CT18	NC 003198
20	*S. enterica*	*enterica* (I)	Typhi	Ty2	NC 004631
21	*S. enterica*	*enterica* (I)	Typhi	Ty21a	NC 021176
22	*S. enterica*	*enterica* (I)	Typhimurium	14028S	NC 016856
23	*S. enterica*	*enterica* (I)	Typhimurium	UK-1	NC 016863
24	*S. enterica*	*enterica* (I)	Typhimurium	ST4-74	NC 016857
25	*S. enterica*	*enterica* (I)	Weltevreden	2007-60-3289-1	NT 187115
26	*S. enterica*	*salamae* (II)	-	CNM-176	DQ644617
27	*S. enterica*	*arizonae* (IIIa)	-	ATCC 13314	MK017930
28	*S. enterica*	*arizonae* (IIIa)	-	CNM-247	DQ644621
29	*S. enterica*	*arizonae* (IIIa)	-	CNM-259	DQ644622
30	*S. enterica*	*arizonae* (IIIa)	-	CNM-771-03	DQ644620
31	*S. enterica*	*arizonae* (IIIa)	-	CNM-822-02	DQ644619
32	*S. enterica*	*diarizonae* (IIIb)	-	ATCC 43973	MK017931
33	*S. enterica*	*diarizonae* (IIIb)	-	CNM-750-02	DQ644624
34	*S. enterica*	*diarizonae* (IIIb)	-	CNM-834-02	DQ644625
35	*S. enterica*	*diarizonae* (IIIb)	-	CNM-2667-02	DQ644623
36	*S. enterica*	*houtenae* (IV)	-	ATCC 43974	MK017942
37	*S. enterica*	*houtenae* (IV)	-	CNM-2556-03	DQ644626
38	*S. enterica*	*indica* (VI)	-	CDC-811	DQ644630
39	*S. bongori*	-	-	N268-08	NC 021870
40	*S. bongori*	-	-	NCTC 12419	NC 015761

III. RESULTS AND DISCUSSION

A. The invA *Amplified and Sequenced Gene Region (229bp) Lacks Discriminatory Power for* Salmonella *Molecular Serotyping*

All three isolates obtained from three meat samples: sample 11 (tonsils - Kayang, Pasay abattoir), APC1R1A (porkchop – Alabang, Muntinlupa market), and PPC1R1A (porkchop – Pasay market), were confirmed as *Salmonella* by *invA* gene detection. Products were sent to Macrogen, Inc. (South Korea) for sequencing and contig assembly using Geneious Prime® 2020.0.3 resulted in 229bp as the shortest product (sample 11) followed by 231bp (sample APC1R1A) and 232bp (sample PPC1R1A). Alignment of sample *invA* gene sequences with Refseqs from GenBank databases (Figure 1) showed that the amplified and sequenced region (229bp) had little to no base variation (black bars) among *Salmonella* serotypes with only a total of 23 variable sites found mostly among subspecies and species levels but only six variable sites among serotypes with four of those variable sites coming from sequences of the samples as a result of ambiguous bases from sequencing. This suggests that the amplified region of *invA* may not be feasible as a serotyping marker. The 229bp *invA* dataset showed an increasing plot of corrected vs uncorrected distances (Figure 2A) and uncorrected vs transitions and transversions (Figure 2B) suggesting that the dataset is not oversaturated and can still be corrected for multiple hits using substitution models (Strimmer & von Haeseler, 2003). Using the Xia test (Xia *et al.,* 2003; Xia and Lemey, 2009) at 60 replicates, and 43 taxa (which the test dictates that at 32 OTUs, the completely symmetrical tree had I_{ss} of 0.023 which is significantly less

(p value < 0.01) than the critical value ($I_{SS.c}$) of 0.685. Similarly, the completely asymmetrical tree had 0.363 critical value ($I_{SS.c}$) which is still significantly less than (p value < 0.01) the I_{SS}, the *invA* (229bp) dataset was interpreted to have little saturation. Using jModeltest 0.1.1 (Posada, 2008), the AIC selected TPM2uf+G (Kimura, 1981) (base frequencies: A=0.2126, C=0.2109, G=0.1979, T=0.3786, relative substitution rates: AC=AT=302.0277, AG=CT=1813.5804, CG=GT=1.0000 gamma shape=0.7050) out of 88 candidate models as the best evolutionary model for the *invA* (229bp) dataset. This model was applied for NJ, and ML trees but since TPM2uf+G is not supported by MrBayes, GTR+G is instead used as substitute (Vea & Grimaldi, 2016).

All trees (NJ, MP, ML, & BI) for the 229bp segment were not able to delineate *Salmonella* serotypes (figures not shown). Based on MP, of the 299 characters (220bp), 209 were constant with only 20 parsimony-informative characters reiterating the minimal number of base variations in the marker. However, although with <50% bootstrap support and <0.7 posterior probability, all three isolates clustered among *S. enterica* subsp. *enterica* (I) serotypes. In contrast, other subsp. except for *salamae* (II), all clustered separately with high bootstrap values and posterior probabilities in all trees. *S. bongori* also clustered in all trees at 100 bootstrap support and 1.00 posterior probability values separate from *S. enterica* subspecies. In addition, in a study by Ng & Rivera (2015) on swine tonsils and jejunum samples from abattoirs, serogroup B (i.e. serotype Typhimurium, Heidelberg) predominated suggesting that the isolates in this study may be of the same serotypes. These results suggest that although the segment (229bp) of *invA* may not be suitable for molecular serotyping of *Salmonella,* the marker was still able to provide some discriminatory power at higher taxonomic levels. This was reflected in Christensen *et al.,* (1998) where the use of 16S rRNA gene on *Salmonella* also had poor separation at lower taxonomic levels but some variation observed at species level.

To test the reliability of phylogenetic analyses of *invA* (229bp) dataset, phylogenetic signal was detected. After analysis using PAUP*, a strong skewness was observed with a g1 value of -1.618702 which at 20 parsimony informative sites and a total of 43 taxa, is less than or more negative than the critical value which at 25 taxa and max of 50 parsimony informative sites, of -0.12 (the critical values change little beyond 15 taxa so the same critical values used at 25 taxa can be used for more than 25 taxa) (Hillis & Huelsenbeck 1992). This suggests strong support for the results and interpretations above that using primers from Chiu (1996) may be useful for *Salmonella* detection but not for molecular serotyping.

Figure 1. Alignment of *invA* samples with Refseqs showing amplified region (Unhighlighted: 33-261) with low base variation (black bars) and segment (1498bp) outside the amplified region (Highlighted: 253-1,750) showing considerable base variation among *Salmonella* species, subspecies and serotypes

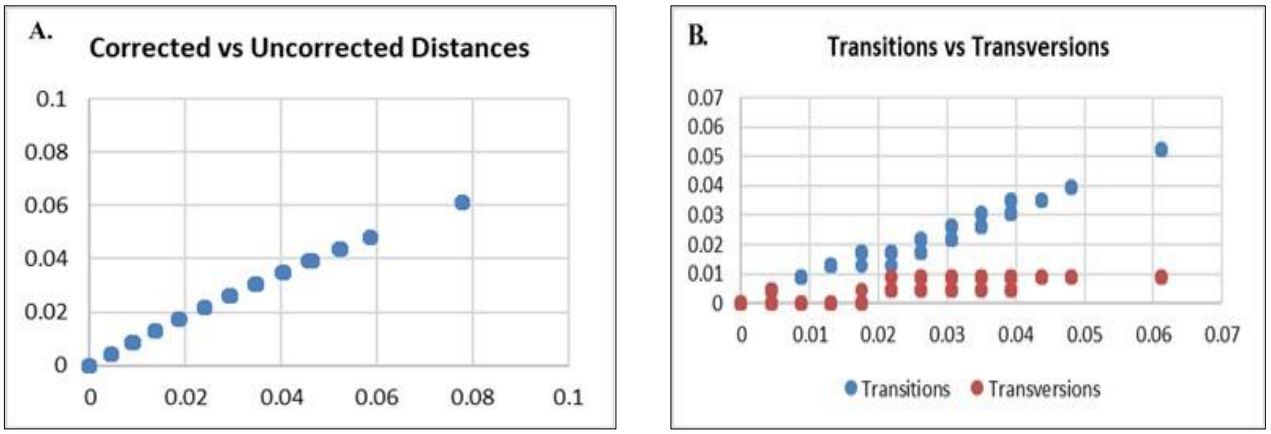

Figure 2. Test for oversaturation of *invA* with samples dataset (229bp) showing an increasing plot for corrected to uncorrected distances (A) and uncorrected distances to transitions and transversions distances (B)

B. Different invA *Segment (1498bp) May Provide Enough Variation for* Salmonella *Molecular Serotyping*

Base variations were observed outside the amplified segment of *invA* among Refseqs. After alignment, the Refseqs were trimmed based on the presence of variable sites among *S. enterica* subsp. *enterica* serotypes which resulted in a 1498bp segment (as shown in Figure 1). As observed, there is considerable base variations (black bars) among serotypes and more at subspecies and species levels, suggesting its potential as a marker for molecular serotyping of *Salmonella* by phylogenetic analysis. The 1498bp segment had a total of 279 variable sites among all taxonomic levels with 48 variable sites among *Salmonella* serotypes. Similarly, the *invA* (1498bp) dataset also showed an increasing plot for corrected vs uncorrected distances (Figure 3A) and uncorrected distances vs transitions and transversions (Figure 3B) suggesting that the dataset is not oversaturated and can still be corrected for multiple hits using substitution models (Strimmer & von Haeseler, 2003). Using the Xia test (Xia *et*.

al., 2003; Xia & Lemey, 2009) at 60 replicates, and 40 taxa, which the test dictates that at 32 OTUs, the completely symmetrical tree had I_{SS} of 0.165 which is significantly less (p value < 0.01) than the critical value ($I_{SS.c}$) of 0.775. Similarly, the completely asymmetrical tree had 0.491 critical value ($I_{SS.c}$) which is still significantly less than (p value < 0.01) the I_{SS}, the *invA* (1498bp) dataset was interpreted to have little saturation. Using jModeltest 0.1.1 (Posada, 2008), the AIC selected TIM3+I (Posada *et. al.,* 2003) (base frequences: A=0.2443, C=0.1975, G=0.2647, T=0.2936, relative substitution rates: AC=CG=0.8891, AG=5.4947, AT=GT=1.0000, CT=10.5125, proportion of invariant sites=0.7050) out of 88 candidate models as the best evolutionary model for the *invA* (1498bp) dataset. This model was applied for NJ and ML trees. However, since BI does not support the model, GTR+G+I was instead applied as a substitute (Lecocq *et al.,* 2013).

As predicted, most trees (NJ, ML and BI) (Figure 4 and 5 shown for NJ and ML tree, respectively) showed that the 1498bp segment of the *invA* gene was able to resolve 12 of the 17 serotypes (70.59%), analyzed with high bootstrap support (coloured dots in figures represent serotypes that were delineated with >50% bootstrap value). While the MP tree showed delineation of 11 out of 17 serotypes (64.71%) with 240 parsimony informative characters and 1,230 constant characters. All trees showed consistency in clustering and separation of serotypes, subspecies and species. Most serotypes (11 of 17) are well-separated in all trees showing consistency (bootstrap, posterior probability) namely, Typhimurium (99-100%, 1.00), Schwarzengrund and Javiana (99-100%, 1.00), Choleraesuis and Paratyphi C (98-100%, 1.00), Heidelberg (94-100%, 1.00), Thompson (88-92%, 1.00), Agona (80-92%, 0.98), Newport (80-82%, 1.00), Weltevreden (74-80%, 0.97), and Paratyphi A (66-69%, 0.88). However, clustering of some serotypes was only supported in some trees; namely, Typhi only in NJ (66%, Fig. 4) and ML (69%, Figure 5) and Cubana only in BI (0.79), although still with supported separations from other serotypes. Hence, adding all serotypes that were resolved by all trees would amount to a total of 13 out of 17 serotypes (76.47%) that were delineated using the 1498bp variable segment of *invA*. Unfortunately, the remaining four serotypes (Gallinarum, Pullorum, Enteritidis and Dublin) failed to provide clear separations. This is consistent with a study that showed that concatenating seven housekeeping genes also failed to differentiate serotypes Gallinarum and Enteritidis (Seong *et al.,* 2012). However, in that same study, the phylogenetic analysis of the complete sequence of housekeeping gene *rpoB* alone was able to separate these serotypes. In another study using MLST analysis of seven housekeeping genes, results were consistent in terms of separations of serotypes and closer clustering of serotypes Paratyphi C and Choleraesuis, Javiana and Schwarzengrund. However, in contrast, Gallinarum and Enteritidis were resolved with high bootstrap values and some serotypes such as Paratyphi A and Typhi clustered differently from this study, most probably due to analysis of more serotypes (Leekitcharoenphon *et al.,* 2012). A study by Kim *et al.*, (2006) using 38 primer sets, also showed some similar topologies particularly for the closer clustering of Paratyphi C and Choleraesuis along with Paratyphi A and Typhi and still, insufficient differentiations (closely clustered) among serotypes Enteritidis, Gallinarum, and Pullorum. Feng *et al.,* (2013) similarly stated that there were only minor differences in the genome of serotypes Enteritidis, Gallinarum and Pullorum. They can be considered as variants of the same bacterium. The lack of resolution among these 4 serotypes is supported in numerous studies mentioning close relation and difficulty in differentiation despite their varying host ranges and pathogenicity (Alzwghaibi *et. al.,* 2019; Barrow & Neto, 2011). These suggest that different genes and techniques offer different advantages and limitations. Hence, the *invA* gene is at par with other analyses with the advantage of being a specific single-gene marker for *Salmonella*. At higher taxonomical levels (subspecies and species), all trees have well-supported delineations of all *S. enterica* subspecies namely, *arizonae* (IIIa), *diarizonae* (IIIb), *indica* (VI), and *houtenae* (IV) and *salamae* (II) and accordingly of species (*S. enterica* and *S. bongori*). Results at higher taxonomical levels are consistent with previous studies using various virulence, i.e. *invE*, *spaM*, *spaN* (Boyd *et al.,* 1997) or housekeeping genes (McQuiston *et al.,* 2008) to delineate subspecies and species levels of *Salmonella*. Hence, these results suggest the potential of this region (1498bp) of the *invA* gene in molecular serotyping of *Salmonella*.

To test the reliability of the phylogenetic analyses of *invA* (1498bp) dataset, phylogenetic signal was detected. After analysis using PAUP*, a strong skewness was observed with a g1 value of -0.769258, which at 240 parsimony informative sites and a total of 40 taxa, is less than or more negative than the critical value which at 25 taxa and max of 250 parsimony informative sites, of -0.08 (the critical values change little beyond 15 taxa so the same critical values used at 25 taxa can be used for more than 25 taxa) (Hillis & Huelsenbeck, 1992). This suggests strong support for the results and interpretations above.

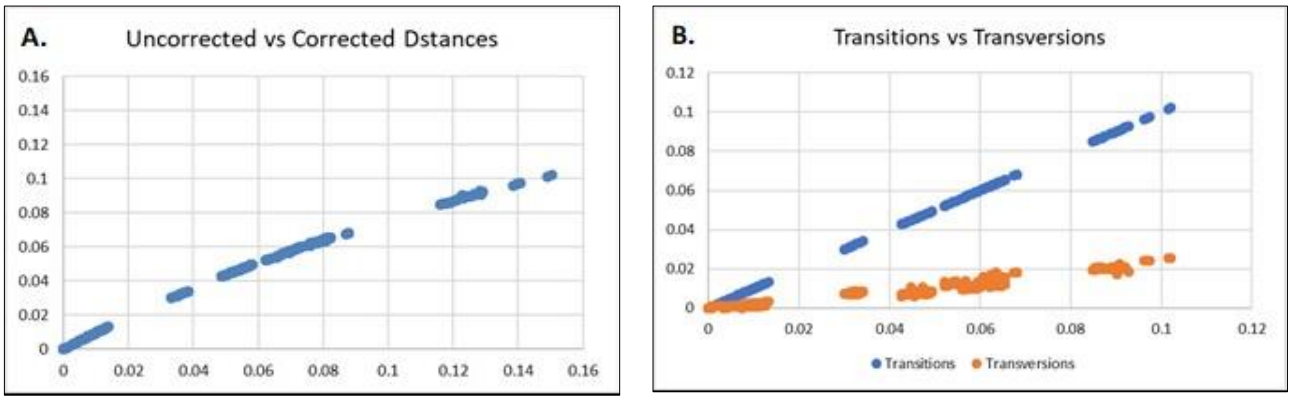

Figure 3. Test for oversaturation of *invA* Refseqs dataset (1498bp) showing an increasing plot for corrected to uncorrected distances (A) and uncorrected distances to transitions and transversions distances (B)

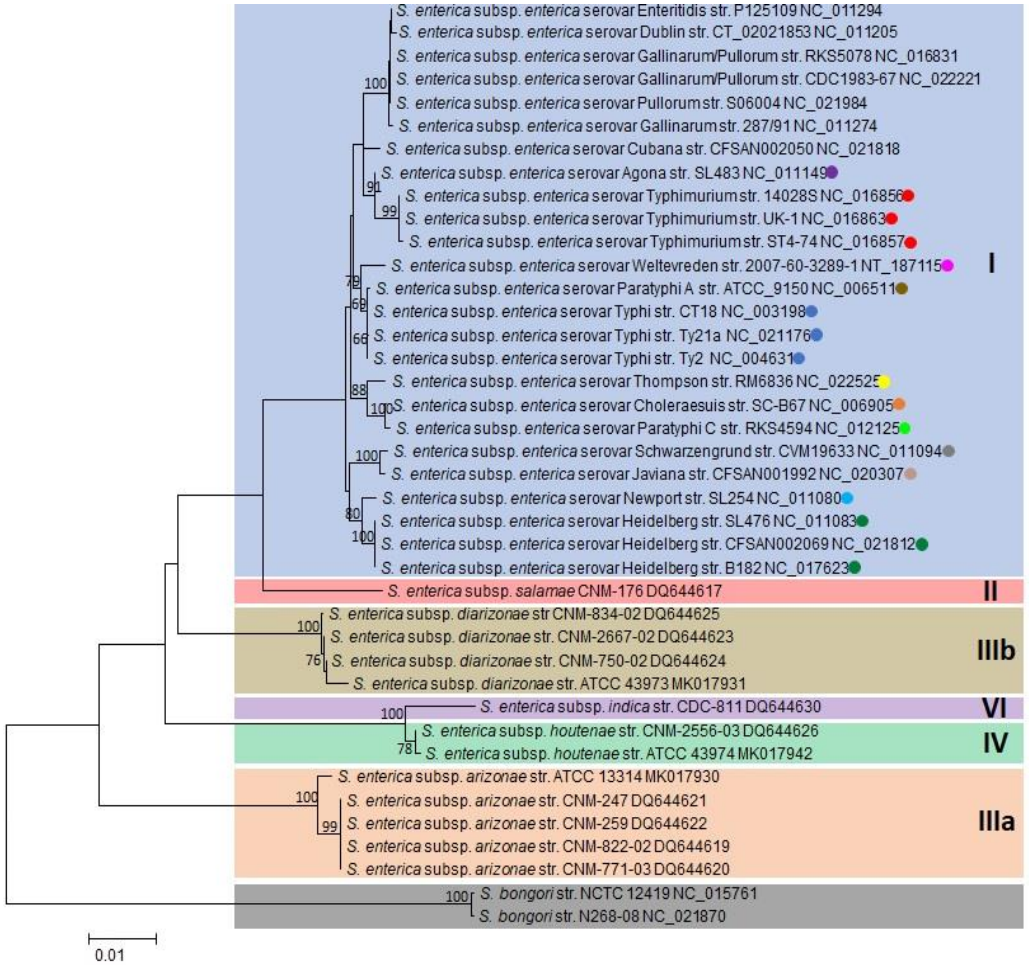

Figure 4. Neighbor-Joining tree of *S. enterica* serotypes based on 1498 nucleotides of the *invA* gene and using the TIM3+I model of DNA substitution. The tree is rooted on *S. bongori*. Values on nodes represent bootstrap percentage out of 1,000 bootstrap samples; values <50% are not shown. Scale bar represents one nucleotide substitutions per 100 nucleotides

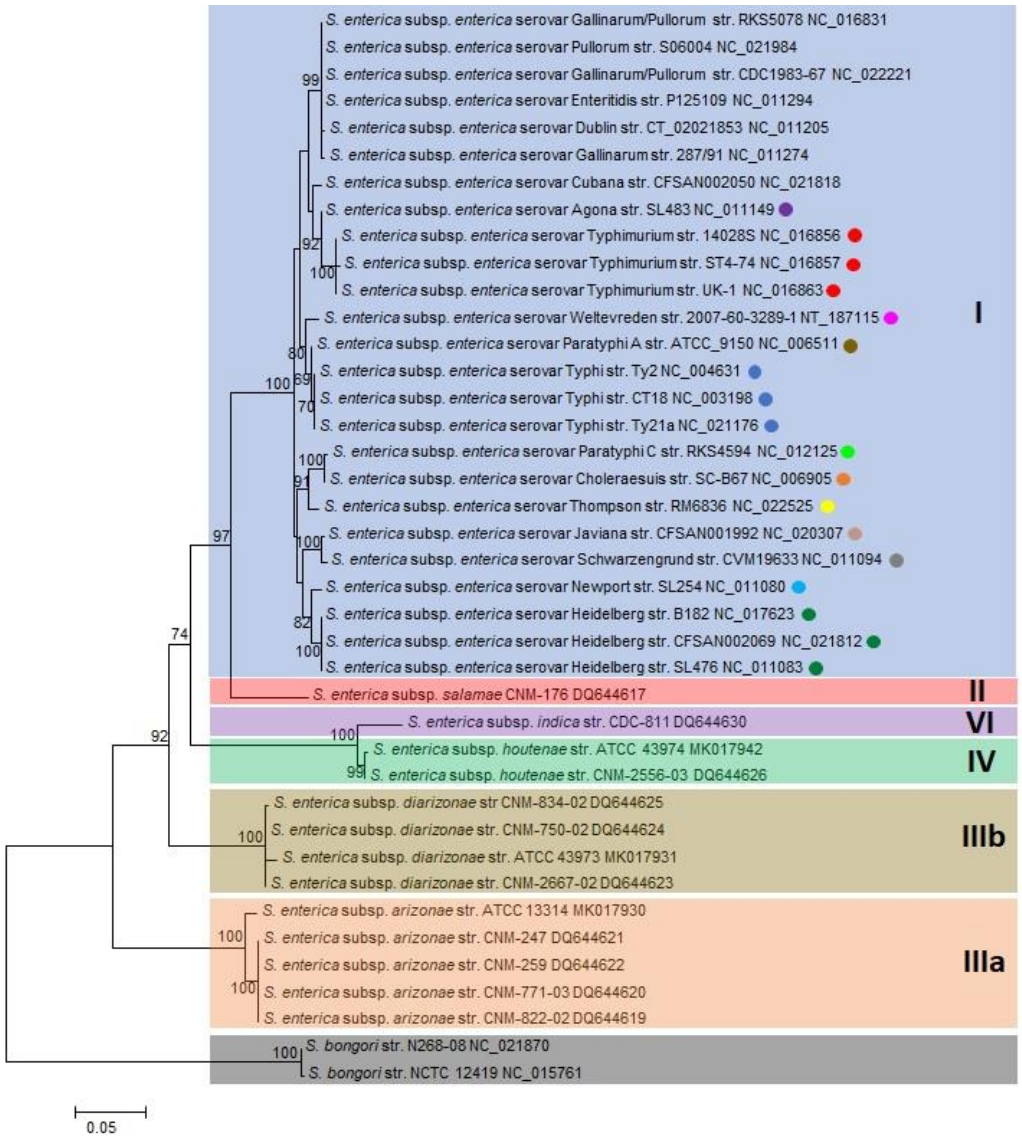

Figure 5. Maximum Likelihood tree of *S. enterica* serotypes based on 1498 nucleotides of the *invA* gene and using the TIM3+I model of DNA substitution. The tree is rooted on *S. bongori*. Values on nodes represent bootstrap percentage out of 1,000 bootstrap samples; values <50% are not shown. Scale bar represents five nucleotide substitutions per 1,000 nucleotides.

IV. CONCLUSION

Although the amplified and sequenced region of the *invA* gene was not able to resolve *Salmonella* serotypes due to the lack of base variations within the 229bp segment, it was able to provide some discriminatory power at higher taxonomic levels. Unfortunately, the *Salmonella* isolates from meat samples obtained from markets and abattoirs in Metro Manila, Philippines failed to be serotyped under this marker. However, a 1498bp segment outside the amplified region showed base variations that were sufficient to resolve most of the tested serotypes with high support values and completely at higher taxonomic levels. These suggest that the *invA* gene possesses enough variability to delineate lower *Salmonella* taxonomic levels and is thus a potential marker for molecular serotyping of *Salmonella*. Thus, the *invA* gene may have comparable resolution to MLST and other housekeeping genes analyses as seen in comparable clustering or separations of serotypes, yet has the advantage of being a single, virulence gene that offers simpler, faster method and without the disadvantages of housekeeping genes. However, more analyses should be done to evaluate the extent of variations of the *invA* gene, such as the ratio of synonymous and non-synonymous mutations, more *Salmonella* serotypes and markers for the developed *invA* gene to target the more

variable regions. More *invA* gene sequencing should also be done since there are still many serotypes lacking representatives in databases. In addition, the complete *rpoB* gene may also be used in complement to delineate some serotypes (e.g. Gallinarum, Pullorum and Enteritidis) that remained unresolved using *invA*. Lastly, considering that *Salmonella* is an efficient pathogen, other virulence genes such as *mgtC*, which is also widely-distributed among *Salmonella*, should also be considered especially in concatenation to establish more refined phylogenetic analyses and to resolve some serotype clustering that remains unsupported in the current study.

V. ACKNOWLEDGEMENTS

This study was supported by the Philippine Department of Agriculture-Biotechnology Program Office (DABIOTECH-R1706). The authors are grateful to Dr. Ian Kendrich C. Fontanilla for his technical support.

VI. REFERENCES

Alzwghaibi, AB, Yahyaraeyat, R, Fasaei, BN, Langeroudi, AG & Salehi, TZ 2019, 'Identification and discrimination of *Salmonella* Enteritidis, *S.* Pullorum, *S.* Gallinarum and *S.* Dublin using *Salmonella* specific genomic regions amplification assay', Iranian Journal of Veterinary Medicine, vol. 13, no. 2, pp. 131-142.

Andrews, WH, Wang, H, Jacobson, A & Hammack, T 2019, Chapter 5: *Salmonella* in: Bacteriological analytical manual, 8th edn, <https://www.fda.gov/food/laboratory-methods-food/bacteriological-analytical-manual-bam-chapter-5-salmonella>.

Association of Public Health Laboratories 2014, *Salmonella* serotyping in US public health laboratories, <https://www.aphl.org/aboutAPHL/publications/Documents/FS_SalmonellaSustainabilityWhitePapr_Nov2014.pdf>.

Azanza, MAV, Membrebe, BNQ, Sanchez, RGR, Estilo, EEC, Dollete, UGM, Feliciano, RJ & Garcia, NKA 2019, 'Foodborne disease outbreaks in the Philippines (2005-2018)', Philippine Journal of Science, vol. 148, no. 2, pp. 323-342.

Barco, L, Lettini, AA, Ramon, E, Longo, A, Saccardin, C, Pozza, MCD & Ricci, A 2011, 'A rapid and sensitive method to identify and differentiate *Salmonella enterica* serotype Typhimurium and *Salmonella enterica* serotype 4,[5],12: i:- by combining traditional serotyping and multiplex polymerase chain reaction', Foodborne Pathogens and Disease, vol. 8, no. 6, pp. 741–743.

Barrow, PA & Neto, OCF 2011, 'Pullorum disease and fowl typhoid—new thoughts on old diseases: a review', Avian Pathology, vol. 40, no. 1, pp. 1–13.

Bell, RL, Jarvis, KG, Ottesen, AR, McFarland, MA & Brown, EW 2016, 'Recent and emerging innovations in *Salmonella* detection: a food and environmental perspective', Microbial Biotechnology, vol. 9, no. 3, pp. 279–292.

Boyd, EF, Li, J, Ochman, H & Selander, RK 1997, 'Comparative genetics of the *inv-spa* invasion gene complex of *Salmonella enterica*', Journal of Bacteriology, vol. 179, no. 6, pp. 1985–1991.

Brown, WM, Prager, EM, Wang, A, & Wilson, AC 1982, 'Mitochondrial DNA sequences of primates: tempo and mode of evolution', Journal of Molecular Evolution, vol. 18, pp. 225-239.

Calayag, AMB, Paclibare, PAP, Santos, PDM, Bautista, CAC & Rivera, WL 2017, 'Molecular characterization and antimicrobial resistance of *Salmonella enterica* from swine slaughtered in two different types of Philippine abattoir', Food Microbiology, vol. 65, pp. 51–56.

Case, RJ, Boucher, Y, Dahllof, I, Holmstrom, C, Doolittle, WF & Kjelleberg, S 2007, 'Use of 16S rRNA and *rpoB* genes as molecular markers for microbial ecology studies', Applied and Environmental Microbiology, vol. 73, no. 1, pp. 278–288.

Cavalli-Sforza, LL & Edwards, AWF 1967, 'Phylogenetic analysis: models of estimation procedures', American Journal of Human Genetics, vol. 19, pp. 233-257.

Centers for Disease Control and Prevention 2019, *Salmonella*, <https://www.cdc.gov/salmonella/index.html>.

Chiu, COJ 1996, 'Rapid identification of *Salmonella* serovars in feces by specific detection of virulence genes, *invA* and *spvC*, by an enrichment broth culture-multiplex PCR

combination assay', Journal of Clinical Microbiology, vol. 34, no. 10, pp. 2619-2622.

Christensen, H 2004, 'Comparative phylogenies of the housekeeping genes *atpD*, i*nfB* and *rpoB* and the 16S rRNA gene within the Pasteurellaceae', International Journal of Systematic and Evolutionary Microbiology, vol. 54, no. 5, pp. 1601–1609.

Christensen, H, Nordentoft, S & Olsen, JE 1998, 'Phylogenetic relationships of *Salmonella* based on rRNA sequences', International Journal of Systematic Bacteriology, vol. 48, no. 2, pp. 605–610.

Clark, MA, Hirst, BH & Jepson, MA 1998, 'Inoculum composition and *Salmonella* pathogenicity island 1 regulate M-cell invasion and epithelial destruction by *Salmonella* Typhimurium', Infection and Immunity, vol. 66, no. 2, pp. 724-731.

De Clercq, D, Ceustermans, A, Heyndrickx, M, Coosemans, J & Ryckeboer, J 2007, 'A rapid monitoring assay for the detection of *Salmonella* spp. and *Salmonella* Senftenberg strain W775 in composts', Journal of Applied Microbiology, vol. 103, no. 6, pp. 2102–2112.

Diep, B, Barretto, C, Portmann, A-C, Fournier, C, Karczmarek, A, Voets, G, Li, S, Deng, X & Klijn, A 2019, '*Salmonella* serotyping; comparison of the traditional method to a microarray-based method and an in silico platform using whole genome sequencing data', Frontiers in Microbiology, vol 10, no. 2554.

Eck, RV & Dayhoff, MO 1966, In: M. O. Dayhoff, eds Atlas of protein sequence and structure, National Biomedical Research Foundation, Silver Spring, Md, pp. 161- 169.

Fàbrega, A, du Merle, L, Le Bouguénec, C, Jiménez de Anta, MT & Vila, J 2009, 'Repression of invasion genes and decreased invasion in a high-level fluoroquinolone-resistant *Salmonella* Typhimurium mutant', PLoS ONE, vol. 4, no. 11, pp. 1-11.

Fakhr, MK, Nolan, LK & Logue, CM 2005, 'Multilocus sequence typing lacks the discriminatory ability of pulsed-field gel electrophoresis for typing *Salmonella enterica* serovar Typhimurium', Journal of Clinical Microbiology, vol. 43, no. 5, pp. 2215–2219.

Felsenstein, J 1981, 'Evolutionary trees from DNA sequences: a maximum likelihood approach', Journal of Molecular Evolution, vol. 17, pp. 368-376.

Feng, Y, Johnston, RN, Liu, G-R & Liu, S-L 2013, 'Genomic comparison between *Salmonella* Gallinarum and Pullorum: differential pseudogene formation under common host restriction', PLoS ONE, vol. 8, no. 3, pp. 1-6.

Fitch, WM 1977, 'On the problem of generating the most parsimonious tree', The American Naturalist, vol. 111, pp. 223-257.

Fox, GE, Wisotzkey, JD & Jurtshuk, P 1992, 'How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity', International Journal of Systematic Bacteriology, vol. 42, no. 1, pp. 166–170.

Fukushima, M, Kakinuma, K & Kawaguchi, R 2002, 'Phylogenetic analysis of *Salmonella*, *Shigella*, and *Escherichia coli* strains on the basis of the *gyrB* gene sequence', Journal of Clinical Microbiology, vol. 40, no. 8, pp. 2779–2785.

Galán, JE & Curtiss, R 1989, 'Cloning and molecular characterization of genes whose products allow *Salmonella* Typhimurium to penetrate tissue culture cells', Proceedings of the National Academy of Sciences, vol. 86, no. 16, pp. 6383–6387.

Galán, JE & Curtiss, R 1991, 'Distribution of the *invA*, *-B*, *-C*, and *-D* genes of *Salmonella* Typhimurium among other *Salmonella* serovars: *invA* mutants of *Salmonella* Typhi are deficient for entry into mammalian cells', Infection and Immunity, vol. 59, no. 9, pp. 2901-2908.

Glaeser, SP & Kämpfer, P 2015, 'Multilocus sequence analysis (MLSA) in prokaryotic taxonomy', Systematic and Applied Microbiology, vol. 38, no. 4, pp. 237–245.

Guindon, S, Dufayard, J-F, Lefort, V, Anisimova, M, Hordijk, W & Gascuel, O 2010, 'New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0', Systematic Biology, vol. 59, no. 3, pp. 307–321.

Heymans, R, Vila, A, van Heerwaarden, CAM, Jansen, CCC, Castelijn, GAA, van der Voort, M & Biesta-Peters, EG 2018, 'Rapid detection and differentiation of *Salmonella* species, *Salmonella* Typhimurium and *Salmonella* Enteritidis by multiplex quantitative PCR', PLOS ONE, vol. 13, no. 10, pp. 1-15.

Hillis, DM & Huelsenbeck, JP 1992, 'Signal, noise, and reliability in molecular phylogenetic analyses', Journal of Heredity, vol. 85, pp. 189-195.

Ibrahim, GM & Morin, PM 2018, '*Salmonella* serotyping using whole genome sequencing', Frontiers in Microbiology, vol. 9, no. 2993.

Kimura, M 1981, 'Estimation of evolutionary distances between homologous nucleotide sequences', Proceedings of the National Academy of Sciences USA, vol. 78, pp. 454-458.

Kim, H-J, Park, S-H & Kim H-Y 2006, 'Comparison of *Salmonella enterica* serovar Typhimurium LT2 and non-LT2 *Salmonella* genomic sequences, and genotyping of *Salmonellae* by using PCR', Applied and Environmental Microbiology, vol. 72, no. 9, pp. 6142–6151.

Kotetishvili, M, Stine, OC, Kreger, A, Morris, JG & Sulakvelidze, A 2002, 'Multilocus sequence typing for characterization of clinical and environmental *Salmonella* strains', Journal of Clinical Microbiology, vol. 40, no. 5, pp. 1626–1635.

Lecocq, T, Vereecken, NJ, Michez, D, Dellicour, S, Lhomme, P, Valterová, I, Raplus, J & Rasmont, P 2013, 'Patterns of genetic and reproductive traits differentiation in mainland vs. corsican populations of bumblebees', PLoS ONE, vol. 8, no. 6, pp. 1-14.

Leekitcharoenphon, P, Lukjancenko, O, Friis, C, Aarestrup, FM & Ussery, DW 2012, 'Genomic variation in *Salmonella enterica* core genes for epidemiological typing', BMC Genomics, vol. 13, no. 88, pp. 1-11.

McQuiston, JR, Herrera-Leon, S, Wertheim, BC, Doyle, J, Fields, PI, Tauxe, RV & Logsdon, JM 2008, 'Molecular phylogeny of the *Salmonellae*: relationships among *Salmonella* species and subspecies determined from four housekeeping genes and evidence of lateral gene transfer events', Journal of Bacteriology, vol. 190, no. 21, pp. 7060–7067.

McQuiston, JR, Waters, RJ, Dinsmore, BA, Mikoleit, ML & Fields, PI 2011, 'Molecular determination of H antigens of *Salmonella* by use of a microsphere-based liquid array', Journal of Clinical Microbiology, vol. 49, no. 2, pp. 565–573.

Morrison, DA 2006, 'Phylogenetic analyses of parasites in the new millennium', Advances in Parasitology, vol. 63, pp. 1-124.

Ng, KCS & Rivera, WL 2015, 'Multiplex PCR–based serogrouping and serotyping of *Salmonella enterica* from tonsil and jejunum with jejunal lymph nodes of slaughtered swine in Metro Manila, Philippines', Journal of Food Protection, vol. 78, no. 5, pp. 873–880.

Nolan, L, Giddings, CW & Brown, J 1995, 'The distribution of *invA*, *pagC* and *spvC* genes among *Salmonella* isolates from animals', Veterinary Research Communications, vol. 19, no. 3, pp. 167–177.

Pal, M, Merera, O, Derra, FA, Rahman, MT & Hazarika, RA 2015, 'Salmonellosis: a major foodborne disease of global significance', Beverage and Food World, vol. 42, no. 12, pp. 21-24.

Paclibare, PAP, Calayag, AMB, Santos, PDM & Rivera, WL 2017, 'Molecular characterization of *Salmonella enterica* isolated from raw and processed meats from selected wet markets in Metro Manila, Philippines', The Philippine Agricultural Scientist, vol. 100, no. 2, pp. 375-382.

Page, RDM 2002, 'Visualizing phylogenetic trees using Treeview', in: Current protocols in bioinformatics, John Wiley & Sons Inc., New York, US.

Porwollik, S, Boyd, EF, Choy, C, Cheng, P, Florea, L, Proctor, E & McClelland, M 2004, 'Characterization of *Salmonella enterica* subspecies I genovars by use of microarrays', Journal of Bacteriology, vol. 186, no. 17, pp. 5883–5898.

Posada, D 2008, 'jModelTest: phylogenetic model averaging', Molecular Biology and Evolution, vol. 25, no. 7, pp. 1253–1256.

Posada, D, Baxevanis, AD, Davison, DB, Page, RDM, Petsko, GA, Stein, LD & Stormo, GD 2003, 'Using Modeltest and PAUP* to select a model of nucleotide substitution', in: Current protocols in bioinformatics, John Wiley & Sons Inc., New York, US.

Rahn, K, De Grandis, SA, Clarke, RC, McEwen, SA, Galán, JE, Ginocchio, C, Curtis, R & Gyles, CL 1992, 'Amplification of an *invA* gene sequence of *Salmonella* Typhimurium by polymerase chain reaction as a specific method of detection of *Salmonella*', Molecular and Cellular Probes, vol. 6, no. 4, pp. 271–279.

Ronquist, F & Huelsenbeck, JP 2003, 'MrBayes 3: Bayesian phylogenetic inference under mixed models', Bioinformatics, vol. 19, pp. 1572-1574.

Saitou, N & Nei, M 1987, 'The neighbor-joining method: a new method for reconstructing phylogenetic trees', Molecular Biology and Evolution, vol. 4, pp. 406-425.

Seong, W-J, Kwon, H-J, Kim, T-E, Lee, D-Y, Park, M-S & Kim, J-H 2012, 'Molecular serotyping of *Salmonella enterica* by complete *rpoB* gene sequencing', Journal of Microbiology, vol. 50, no. 6, pp. 962–969.

Shanmugasamy, M, Velayutham, T & Rajeswar, J 2011, '*invA* gene specific PCR for detection of *Salmonella* from broilers', Veterinary World, vol. 4, no. 2, pp. 562-564.

Soguilon-Del Rosario, S & Rivera, WL 2015, 'Incidence and molecular detection of *Salmonella enterica* serogroups and *spvC* virulence gene in raw and processed meats from selected wet markets in Metro Manila, Philippines', International Journal of Philippine Science and Technology, vol. 8, no. 2, pp. 52-55.

Strimmer, K & von Haeseler, A 2003, 'Nucleotide substitution models', in: Salemi, M. & A-M, eds Vandamme,

The Phylogenetic Handbook: A Practical Approach To DNA and Protein Phylogeny, Cambridge University Press, UK.

Swofford, DL 2002, 'PAUP* 4.0b10', in: Phylogenetic analysis using Parsimony (*and other methods), Sinauer Associates, Sunderland, MA.

Tamura, K 1997, TreeExplorer Version 2.12, <http://www.ctu.edu.vn/~dvxe/Bioinformatic/Software/BIT%20Software/TE_man.html>.

Trkov, M & Avgusˇtin, G 2003, 'An improved 16S rRNA based PCR method for the specific detection of *Salmonella* enterica', International Journal of Food Microbiology, vol. 80, no. 1, pp. 67-75.

Vea, IM & Grimaldi, DA 2016, 'Putting scales into evolutionary time: the divergence of major scale insect lineages (Hemiptera) predates the radiation of modern angiosperm hosts (electronic supplementary materials)', Nature, vol. 6 pp. 1-11.

Větrovský, T & Baldrian, P 2013, 'The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses', PLoS ONE, vol. 8, no. 2, pp. 1-10.

World Health Organization 2018, *Salmonella* (non-typhoidal), <https://www.who.int/news-room/fact-742 sheets/detail/salmonella-(non-typhoidal)>.

Xia, X 2018, 'DAMBE7: New and improved tools for data analysis in molecular biology and evolution', Molecular Biology and Evolution, vol. 35, pp. 1550–1552.

Xia, X & Lemey, P 2009, 'Assessing substitution saturation with DAMBE', in: Philippe Lemey, Marco Salemi and Anne-Mieke Vandamme, (ed.), The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny, 2nd edn, Cambridge University Press. pp. 615-630.

Xia, X, Xie, Z, Salemi, M, Chen, L & Wang, Y 2003, 'An index of substitution saturation and its application', Molecular Phylogenetics and Evolution, vol. 26, pp. 1-7.

Yang, Z 1997, 'PAML: A program package for phylogenetic analysis by maximum likelihood', Bioinformatics, vol. 13, no. 5, pp. 555–556.