Peptidome-Based Serotyping of The Food-Borne Pathogens Salmonella Enterica by Label-Free Mass Spectrometry

Xixi Wang
Sichuan University State Key Laboratory of Biotherapy
Yang Yang
Chengdu Center for Disease Control and Prevention
Lian Wang
Chengdu Center for Disease Control and Prevention
Ming Li
Chengdu Center for Disease Control and Prevention
Peng Zhang
Sichuan University West China Hospital
Shufang Liang (zizi2006@scu.edu.cn)
Sichuan University https://orcid.org/0000-0003-1000-7508

Methodology

Keywords: Label-free peptidomic, peptide marker, Salmonella enterica Serovars, strain similarity, food-borne pathogens, C5.0 decision tree.

Posted Date: November 18th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-107261/v1

Peptidome-based serotyping of the food-borne pathogens *Salmonella enterica* by label-free mass spectrometry

Xixi Wang [1,2¶], Yang Yang[2¶], Lian Wang[2], Ming Li [2], Peng Zhang [3], Shufang Liang [1 *]

[1] State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, Sichuan University, and National Collaborative Innovation Center for Biotherapy, Chengdu, 610041, P. R. China

[2] Chengdu Center for Disease Control and Prevention, Chengdu, 610041, P. R. China

[3]Department of Urinary Surgery, West China Hospital, West China Medical School, Sichuan University, Chengdu, 610041, P. R. China.

* Correspondence to Dr. Shufang Liang

¶Contributed equally

zizi2006@scu.edu.cn

Xixi Wang: fozhu.1984@163.com; Yang Yang: 489499012@qq.com; Lian Wang: septwolvesnjwl@163.com; Ming Li: lm781227@163.com; Peng Zhang: zpeng2001@163.com.

ABSTRACT

Background: Food-borne diseases caused by *Salmonella enteric serovars* represent a serious public health problem worldwide. More than 2500 different serovars have been reported to relate with food-borne diseases according to the classification of White-Kauffmann-Le Minor scheme by now. A quick identification for the pathogens is critical for controlling food pollution and disease spreading.

Results: Here we applied a peptidomic analysis for quickly and precisely identifying serovar-specific peptide markers based on LC-MS/MS profiling of epidemiologically important *Salmonella enterica* subsp. *enterica serovars* in China. By label-free quantitative peptidomics MS identification, the 53 most variable serovar-related peptides were screened as potential peptide biomarkers, based on which a C5.0 predicted model with 4 predictor peptides was generated and a test set of 17 *Salmonella enterica* strains were classified with the accuracy of 94.12%. It is effective to determine the genotypic similarity among *Salmonella enteric* isolates according to each strain peptidome profiling, which is indicative of potential incidence even breakout of food contamination. This high-throughput strain peptidomic fingerprints are complementary to the genomic patterns by PFGE analysis for precise identification of 5 *Salmonella enterica serovars* including *Enteritidis*, *Typhimurium*, *London, Rissen* and *Derby*. The biological analysis showed that most of the changed peptides/proteins were enzymes related to nucleoside phosphate and energy metabolism.

Conclusions: the LC-MS/MS based quantitative peptidomic dissection on *Salmonella enteric serovars* provides a novel insight and real-time monitoring of food-borne pathogens.

Keywords:

Label-free peptidomic/peptide marker/*Salmonella enterica Serovar*s/ strain similarity/ food-borne pathogens/C5.0 decision tree.

Introduction

Salmonella, more than 2500 serotypes, is a major zoonotic food-borne pathogen, which causes outbreaks and sporadic cases of gastroenteritis in humans. Approximately 300 serovars are reported in China, of which *Enteritidis*, *Typhimurium*, *Rissen*, *Derby* and *London* are among the top five *Salmonella enterica serovars* isolated from food-borne *Salmonella* infections. The Salmonella strains have a drastic change in virulence or expression under the condition of a single gene mutation. So the identification and characterization of species and subspecies are generally necessary for pathogen confirmation and clinical diagnostics. Moreover, food safety efforts require serovar and strain level specificity for trace-back the source of bacterial contamination.

The *Salmonella* serotyping method based on White-Kauffmann-Le Minor scheme is accepted worldwide as a gold standard for the differentiation of *salmonella* below the subspecies level [1]. It is determined by a combination of biochemical reactions and serotyping of the somatic O, flagellar H, and capsular Vi antigens. However, the antigen-based serotyping, often performed by slide agglutination, is laborious, time-intensive, and expensive due to more than 200 different antisera and the wide-ranging quality of antibodies.

The pulsed field gel electrophoresis (PFGE) profiling has become a gold standard for molecular subtyping of *Salmonella*, and PCR-based amplification and genetic sequencing are also becoming increasingly popular for strain identification. But both of the two approaches have difficulty to precisely distinguish two highly similar serovars such as *S. Typhimurium* and *S. Heidelberg*. Other DNA-based techniques, including plasmid profile, ribosomal DNA intergenic spacer amplification, multi-locus sequence typing (MLST), multi-locus variable numbers tandem repeat analysis (MLVA) and clustered regularly interspaced short palindromic repeats (CRISPR), are usually available, which are still challenging due to the inefficiency to indicate phenotypes and multiple primers required for amplifying the sequences of untargeted genes.

Mass spectrometry (MS) provides a high-throughput and relatively unbiased view of the protein profiling in bacteria [2], which facilitates differentiation of genetically-related bacteria and decoding the new nonsynonymous single-nucleotide polymorphisms. In recent years, a technique known as direct bacterial profiling, has increasingly been applied to dissect proteome for bacterial species identification *via* the matrix-assisted laser desorption ionization–time of flight mass spectrometry (MALDI-TOF MS) [3-10]. A sufficient number of stable mass signals of major housekeeping proteins, such as ribosomal proteins, are reproducibly detected for bacterial species identification by using simple mass pattern-matching approaches or more sophisticated algorithms to compare and estimate the similarities between spectra [6, 11,12,13,14, 15, 16, 17, 18]. This identification approach is not dependent on actually specific ion peaks of MS spectrum but on the characteristic mass profiles (patterns) generated by a set of ion peaks called “fingerprint.”

This “fingerprint” platform successfully produces data for bacterial subtyping at the species level but hardly recognizating the serovar level (i.e., the H and O antigen levels) [19] due to its bias toward small ribosomal proteins with a limited range of 2−20 kDa. With the wider mass range, better sample to sample reproducibility, and greater number of proteins ionized, the electrospray ionization (ESI) based MS platform provides access to a more diverse range of proteins, potentially providing greater specificity for bacteria [20,21,22]. It has already been used to identify marker masses that differentiate thermophilic vs nonthermophilic groups of Cronobacter sakazakii [23] to identify proteins characteristic of specific outbreak strains of V. parahaemolyticus and guide the development of PCR probes19 and to differentiate closely related species within the enterobacteriaceae family [22,24]. In addition, the approach has been shown to quantify protein expression differences by using certain housekeeping proteins as internal standards.

Data mining science can extract useful knowledge which is hidden in the data. Predictive analytics is the process by which information is extracted from existing data sets for determining patterns and predicting the forthcoming trends or outcomes.

For clinical trials, Using the extracted prediction rules have the potential to predict the pathogenic factors[25], disease progression and prognosis[26]. Among the different methods of data mining, the decision tree is one of the powerful and common tools for creating predictions.

In this article, we demonstrate a peptidomic analysis for identifying serovar-specific peptide markers among epidemiologically important *Salmonella enterica subsp. enterica serovars* in China. And 53 most variable serovar-related peptides have been identified as potential biomarkers. Taking the 53 peptides as variables and serotypes as target, a C5.0 predicted model with 4 predictor peptides was generated and a test set of 17 *Salmonella enterica* strains were classified with the accuracy of 94.12%. It has also turned out to be effective applying the whole peptidomic profiling to determine the genotypic similarity among *Salmonella enteric* isolates as a sign of the breakout of a food contamination incidence by comparing the results with PFGE.

METHODS

Bacterial strains

42 *Salmonella enterica* strains belonging to 5 different serovars (Table 1) were collected from Chengdu Center for Disease Control and Prevention. All strains were biochemically differentiated on the subspecies level and serotyped by slide agglutination with O antigen-specific and H antigen-specific sera (Sifin Diagnostics, Germany, Berlin) according to the White-Kauffmann-Le Minor scheme [1]. The selected strains were isolated from infected humans and contaminated food. All the *Salmonella enterica* strains were grown for 24 h at 37 °C on LB agar plates (Teknova, Hollister, CA).

Table 1. *Salmonella enterica subsp. enterica* strains (n =42) used in this study

Group	Serovar	No. of Strains	Source(s)
Training	*Enteritidis* (9,12:g,m:-)	5	Human, food
	Typhimurium (4,5,12:i:1,2)	7	Human, food
	Derby (4,5,12:f,g:-)	6	Human, food
	Rissen (6,7:f,g:-)	5	Human, food
	London (3,10:l,v:1,6)	2	Human, food
Testing	*Enteritidis* (9,12:g,m:-)	12	Human, food
	Typhimurium (4,5,12:i:1,2)	1	Human, food
	Derby (4,5,12:f,g:-)	1	Human, food
	Rissen (6,7:f,g:-)	1	Human, food
	London (3,10:l,v:1,6)	1	Human, food
	Sagona. (4,5,12:f,g,s:-)	1	Human, food

Cell lysis and protein extraction

Bacterial cells were centrifuged at 5000 × g for 10 minutes to collect pellet, in which 5mL of B-PER Complete Reagent (B-PER™ Complete Bacterial Protein Extraction Reagent, Thermo Scientific, USA) per gram of cell pellet was added to mix up and down. The suspension was incubated 15 minutes at room temperature with gentle rocking, following soluble proteins were separated from the insoluble parts by centrifuge at 16,000 × g for 20 minutes. Finally, cell supernatant was transferred to a new tube for protein concentration determination by BCA assay (Beyotime, China).

Trypsin digestion and peptide enrichment

We applied the filter-aided sample preparation (FASP) [27] and stop-and-go-extraction tips (StageTips) protocols [28] for protein digestion and desalting. Briefly samples were heated for 30 min at 50ºC for reduction. YM-30 membrane filters (Millipore, Cat. No. 42410) were activated with 200μL 100mM NaOH, then equilibrated with 200μL 8M urea buffer, and centrifuged for 15min. 200ug samples were added into the filters, centrifuged and washed two times with urea buffer. The alkylation was performed by 10μL of 500mM IAA in 90μL urea buffer for 30 minutes at 37ºC in the dark and centrifuged. 200μL 50mM Ammonium Bicarbonate Buffer (ABC) was added to the filter and centrifuged for three times. Subsequently, 4μL of 0.5 ug/μL MS-grade trypsin (V5280, Trypsin Gold, Promega,

USA) in 100μL ABC buffer was added to incubate at 37ºC overnight. The enzymatic digestion was stopped by centrifuging and the filter was washed by 200μL ABC buffer again. The filtrate was selected, which then subjected to SpeedVacuum to dry out. All centrifugation steps were performed at 16000g. The sample was resuspend with 100μL 0.2% acetic acid. By activating the self-made C_{18} embedded tips with 200μL methanol and water, the sample was added into the tip and centrifuged at 4600g. After washing the tip with 200μL 0.2% acetic acid, the sample were eluted by using 200μL acetonitrile (ACN)/water(v/v=40/60) and 400μL acetonitrile (ACN)/water (v/v=80/20). The eluent was dried out and resolved in 40μL 0.1% formic acid, 4μL of which for LC-MS/MS analysis.

Nanoflow-high performance liquid chromatography (HPLC)

Peptide samples were separated by HPLC on a PicoFrit analytical column (75 μm×10 cm, 5 μm BetaBasic C_{18}, 150 Å, New Objective, MA) at a flow rate of 300 nl/min. A 130 min LC gradient was applied. The gradient started with 98% solvent A (0.1% formic acid in water), and increased to 35% solvent B (0.1% formic acid in acetonitrile) over 110 min, followed by a steeper gradient to 80% solvent B over 15 min.

MS identification

The peptides were identified by LC-MS/MS analysis on an Ultimate 3000-nano LC apparatus and a Q Exactive mass spectrometer system coupled via a FLEX nano-electrospray ion source (all components from Thermo Scientific, West Palm Beach, FL). Eluting peptides were sprayed at a voltage of 2.3 kV and acquired in a MS data-dependent mode using XCalibur software (version 2.2, Thermo Scientific). Survey scans were acquired at a resolution of 70,000 over a mass range of m/z 250 to m/z 1,800 with an automatic gain control (AGC) target of 10^6. For each cycle, the top 20 most intense ions were subjected to fragmentation by high energy collisional dissociation with normalized collision energy of 27. The induced fragment ions from the MS/MS scans were acquired at a resolution of 17,500 with an AGC target of 5

$\times 10^4$. Dynamic exclusion was set to 20 s. Unassigned ions were rejected and only those with a charge ≥ 2 were subjected to HCD fragmentation.

Pulsed-field gel electrophoresis (PFGE)

All the *Salmonella* strain isolates were subjected to PFGE according to the standardized protocol of the CDC PulseNet (PNL05, April 2013). Briefly, cell suspension buffer (100 mM Tris, 100 mM EDTA, and pH 8.0) with a turbidity reading of 1 to 1.3 was mixed in equal volumn with molten 2% low-melting point agarose, pipetted into disposable molds and then stored at 4 ºC for 20 to 30 min, which were incubated overnight at 56 ºC in 1 ml of lysis buffer (0.5 M EDTA, 0.5 M Tris, 1% N-laurylsarcosine) with 250 μg/mL proteinase K (Promega, U.S.A). The sterile ultrapure water and 0.01M Tris-EDTA buffer, pH 8.0 were used to remove excess reagents and cell debris from the lysed plugs. Chromosomal DNA was digested with 30 U of XbaI (Fermentas, Lithuania) for 3 h at 37 ℃. Electrophoresis was carried out with 0.5xTBE buffer at 6 V/cm and 14℃ by CHEF DRIII system (Bio-Rad, USA). The running time was 20 h and the pulse ramp time was 5 to 30 s. *Salmonella enterica* serotype *Braenderup,* strain H9812 was used as a size marker. The gels were visualized on a UV transilluminator, and photographed by a digital imaging system (Gel Doc XR, Bio-Rad) which subsequently converted the gel images to the TIFF file format. DNA fragments patterns were analyzed with BioNumerics software (Applied Maths USA). All the isolates were clustered into different pulsotypes by genetic similarity cut-off ≥ 85%. Reproducibility power was confirmed by comparing the fingerprint patterns that were obtained from duplicate runs of the same isolates.

Quantitative peptidomic analysis and bioinformatics methods

Raw MS raw files were imported into the MaxQuant software suite (v1.6.0) [29] with the default settings for quantification via MS1 peak integration and normalization of proteomic data comparing multiple samples. We used the label-free quantification (LFQ) function to estimate protein abundances in all of the analyzed samples. The parameters for database searching were set as following, including UniProt KB databases (*Salmonella*, Aug 30, 2018), trypsin digestion with two missed

cleavages, carbamidomethyl (C) as a fixed modification, oxidation (M) and acetyl (protein N-term) as the variable modification. Initial peptide mass tolerance was set to 7 ppm and fragment mass tolerance was 0.5 Da, with + 2 as default charge state of each peptide. The false discovery rates (FDRs) of peptide were both set to 0.01. An automated R-based QC pipeline called Proteomics Quality Control (PTXQC) [30] for LC−MS/MS data generated by the MaxQuant software pipeline was applied to detect measurement bias, verifying consistency, and avoiding propagation of error. PTXQC created a QC report containing a comprehensive and powerful set of QC metrics, augmented with automated scoring functions. The replicates for each sample, which failed in Alignment performance for Retention time (RT) correction (RT difference (ΔRT) to Ref>0.7min, would be removed for downstream analysis. The acceptable peptide results were imported in Perseus to perform some data transference and Hierarchical clustering analysis.

The predicted model development

The SPSS Modeler of IMB has been implemented with various tools and algorithms to model and assess the impact of peptide biomarkers on Salmonella serotyping. To prepare and test the models, 42 isolates and their biological replicates were randomly categorized into two groups for model training (25 isolates) and testing (17 isolates). All the candidate peptides from the quantitative peptidomic analysis except serotype were employed as variables, the serotype group being the target.

RESULTS

Comprehensive peptidomic profiling for different Salmonella serotypes

42 *Salmonella* isolates with different serotypes were divided into the training and testing group (Table 1). The training one was analyzed by shotgun proteomics based LC-MS/MS. The total 7339 peptides were identified (FDR>0.01) and quantified by LFQ intensities in at least two technical repeats. To avoid the peptide markers'

variance, we limited the acceptable ones as being identified in all replicates within at least one certain serotype. And 1050 peptides were used for marker development, of which 115 peptides were found in *Enteritidis*, 77 in *Typhimurium*, 250 in *Rissen*, 472 in *Derby* and 456 in *London*. Meanwhile all the identified peptides were applied to differentiate similarity among Salmonella enteric isolates.

Peptide markers for *Salmonella enterica* serotyping

We first tested whether the peptide markers were feasible to identify Enteritidis, Typhimurium, London, Rissen and Derby in the training set. We used the Perseus computational platform [31] to filter the LC-MS/MS raw peptides results which were identified in at least two of three technical repeats by reversed identification and potential contamination, then the LFQ values of peptides were logarithmized and imputation was done by normal distribution with 0.3 width and 1.8 down shift. A multiple sample ANOVA t-tests were performed with a permutation-based testing correction that was controlled by using a FDR threshold of 0.05. The coming results showed 97 peptides were collected with ANOVA significant ($p<0.05$). Considering the data dependent acquisition (DDA) methods we applied in shotgun proteomics, the identified peptides may be acquired by MS randomly and couldn't be characterized by typical chromatographic peaks to quantify accurately. So we checked all the significantly changed peptides in Skyline [32] (Figure 1), and only 53 peptides (Table 2) were certificated with good peak shape ($\geqslant$6 data acquired points, and S/N$\geqslant$10). The power of 53 peptides to separate the five serotypes from each other was shown by a Hierarchical cluster analysis in the training group (Figure 2).

Figure 1. The part of LC-MS/MS spectra of *Salmonella enterica* serovar-identifying peptide markers in training group.

Figure 2. Hierarchical cluster analysis with 53 peptide markers in training group. 25 isolates from 5 serotypes were divided into 5 clusters without overlap.

Table 2. Representative 53 peptide markers detected from 5 serotypes in training groups

Serotypes	Peptide	Mass(Da)	Leading protein ID
L*	VETISYVK	937.512	G5LCI8
R	NLNLTDAQR	1043.54	A0A4Q8PFP4
L, E, T	IFYNDFQADDADLSDYTNK	2253.97	G5RAQ2
L, T, R, D	LQYVDESLSDDQWICGQR	2210.99	V7ITE4
R, D	DDASQTLTTDWVSWNR	1893.85	M7S778
D	VDINSGAVVTDAVAPNK	1668.87	Q06970
L,T,R,D	TLLGQQGYATLADIPEK	1816.96	V1WQL9
R, D	TAEVLAPLGINVTGIHR	1759.99	V7IQV5
R	ADITPVNVDTVTR	1399.73	G5RQP9
R	HLVDLYQQQGIDK	1555.8	V7IUA8;
D	AEASQYDALANAR	1378.65	G4C2Q2
R	LAIALCEQESHLDLR	1766.9	A0A2C9NZ70
L, T	AGFATSQQAYDEAVDK	1699.77	Q8ZLW8
R	LLADGMESFNQHCASGIEPNR	2345.05	G5R0G3
R	NQLTAAALFPLYVNAAAK	1875.03	V1HFI7
R	GFDSLLSEVK	1093.57	A0A0H2WRG6
D	ITFNAPTVPVVNNVDVK	1825.99	V7IT80
L, T	YGVVEFDQK	1083.52	P26393
R	LDEWENAFAEWR	1564.69	A0A0W5T2G5
R, D	TNSAQYDDSNMGQNK	1671.68	V1WYV3
D	VNLIESLASLSVTK	1472.84	G5R943
L, T	TNNLTADPTNPLAQVPAGEIR	2191.12	Q8ZRQ2;
R	DAYIDHLLGYISVNNLTPLK	2258.19	Q8Z5H4
R	WDNTPVMEEILALR	1685.84	V1H6Q6
D	EGSSLLGSDAGELAGVGK	1645.82	Q57TJ7
D	NVYTSVVNGQFTFDDK	1832.86	Q6V2X1
D	SLHSPGLAFR	1083.58	G5L558
R, D	ETNVIDKDGNPQTLK	1670.85	V7IPT6
D	TFFANSVLTNAVDQAK	1724.87	G5S2P3
L, T	LPTDFNEGASNNTYSR	1784.8	A0A0H3BSP7
R	TDGLSMSFADWR	1384.61	Q8Z5H4
D	SIQQGMLR	931.491	Q8Z3X0
R	DIGLPGIADAHIVLTNLASQIGR	2343.29	A9MGZ9
D	GDDIAGLLAVVQPVPPADAR	1973.06	V1WTT2
R	DLVASGFTR	964.498	B5F944
R	TEEVVAENPGK	1171.57	A0A117HYN3
D	LVDEAESHNLNLTYR	1772.87	A0A4P5EY10
R	GQALPLSVSYVSTTAEGAQR	2034.04	G5Q4H5

Serotypes	Peptide	Mass(Da)	Leading protein ID
R,D	FQQPVNSVLAPTDVVTR	1869.99	A0A1S0ZQL6
R	DYYLAENRDESFDMAENDK	2323.95	G5S6V3
R	LAEPAAAIAR	981.561	Q8Z5H4
R	DQYNLHPVYK	1275.62	B5F747
R	VEAHFAEEAQAVDR	1570.74	Q8Z5H4
R	DLGVTLSPAEHAER	1493.75	G4C3W3
D	VTEVGITGLNADFLR	1603.86	Q8Z9L7
D	YVEDNYTTK	1131.51	A0A3Z6VVS9
R	HELAQLLGFESYAFK	1751.89	A0A100V908
R	ITQWLATYVEK	1350.72	V1VX57
D	QLLPDDTVWR	1241.64	V7IUF0
R	ISLVVPVFNEEATIPIFYK	2178.2	G4BXV1
D	EGVLADGIQTFPDR	1516.75	B5F862
L, E, T	EVPALMAGGHLDPEK	1562.78	C0PVT9
D	VNAADLLTILQALK	1481.88	A0A2A6D6Z1

* Abbreviated form of different serotypes: *Enteritidis* (E), *Typhimurium* (T), *Derby* (D), *Rissen* (R), *London* (L).

Accuracy of models and important predictor variables

It still seemed too much for the daily testing work with 53 peptides as a profile pattern for serotyping markers, so we applied the SPSS Modeler to model and obtain the most effective peptides for *Salmonella enterica* serotyping. Taking the 53 peptides as variables and serotypes as target, 25 of 42 *Salmonella* isolates in the training group were processed and two models generated by quick, unbiased, efficient statistical tree (QUEST)[33] and C5.0[34] algorithms with 100% accuracy were collected for further assessment. The most important predictor peptides based on the C5.0 method were the 4 peptides including “IFYNDFQADDADLSDYTNK, YGVVEFDQK, VETISYVK, VANNDLLTILQALK”(Figure 3A). For QUEST way, there were also 4 predictors as “AEASQYDALANAR,VETISYVK,YGVVEFDQK,LQYVDESLSDDQVVICGQR” (Figure 3B). The two peptides “VETISYVK”,“YGVVEFDQK” were common for both methods.

Figure 3. The decision tree for the prediction of Salmonella enteric serotypes. The model based on the C5.0(A) and QUEST(B) method.

To evaluate the capability of these three models, a testing group including 17 isolates were predicted. The predictive accuracy of the C5.0 and QUEST methods were 94.12% and 88.24% respectively. For C5.0 model, 16 of 17 strains were identified correctly, while only one isolate couldn't be categorized among the five serotypes. The exceptional undetectable one was *Sagona* which was recognized through a serological test according to White-Kauffmann-Le Minor scheme. QUEST model unrecognized *Sagona* and identified a *Rissen* as *Eneritidis.* So the C5.0 based predicted model was more reliable for *Salmonella enterica* serotyping in this research.

Discovering the genetic and biological explanations for the distinct peptidomic profiles in the Salmonella serotypes

Instead of the proteins, which were indexed in UniprotKB database with redundancy for Salmonella, we applied shotgun proteomic strategy for peptide identification to achieve the more veracity in case the marked protein IDs here were obsoleted. Associating biological process with distinct peptidomic profiles in these 5 Salmonella serotypes, we had to align the peptide markers to certain proteins (Table 3). The Maxquant software help us to relate these two parts, but there are redundant proteins which need to be verified and checked manually by following rules. 1) the protein is active by query. 2) All the redundant protein IDs for one peptide attribute to one single gene. 50 proteins of 53 peptides were finally listed [see Additional file 3]. Of which 25 proteins tended to be located at the cellular anatomical entity, and more than 80% proteins significantly enriched for the GO (Gene Ontology) molecular function terms of catalytic and binding activity. These proteins were involved in cellular processes, metabolic processes [see Additional file 1].

Table 3. The EC numbers and enzyme names of the proteins derived from changed peptides

Protein ID	EC* number	Accept name
B5F747	6.3.5.5	carbamoyl-phosphate synthase (glutamine-hydrolysing)
Q8Z9L7	6.3.5.5	carbamoyl-phosphate synthase (glutamine-hydrolysing)
Q8Z5H4	5.4.2.8	phosphomannomutase
G5S2P3	5.4.2.12	phosphoglycerate mutase
B5F862	5.1.3.15	glucose-6-phosphate 1-epimerase
G5L558	4.1.3.30	methylisocitrate lyase
G5R943	4.1.1.65	phosphatidylserine decarboxylase
G4C3W3	4.1.1.23	orotidine-5′-phosphate decarboxylase
V7IUF0	3.5.2.7	imidazolonepropionase
B5F944	3.5.2.3	dihydroorotase
V7IQV5	3.4.13.21	dipeptidase E
G4C2Q2	3.2.1.28	phosphomannomutase
V1HFI7	3.2.1.28	α,α-trehalase
P26393	2.7.7.24	glucose-1-phosphate thymidylyltransferase
A0A3Z6VVS9	2.7.2.1	acetate kinase
V7IT80	2.3.1.39	[acyl-carrier-protein] S-acetyltransferase
G5RQP9	2.3.1.117	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase
Q8Z3X0	1.4.4.2	glycine dehydrogenase (aminomethyl-transferring)
Q57TJ7	1.17.1.8	4-hydroxy-tetrahydrodipicolinate reductase
C0PVT9	1.15.1.1	superoxide dismutase
A0A2A6D6Z1	1.11.1.6	catalase
A0A117HYN3	1.1.1.44	phosphogluconate dehydrogenase

*EC: Enzyme Commission

With catalytic and binding activity, we supposed the changed proteins possibly belong to some kind of enzymes. So all candidate protein were searched against the ExplorEnz (https://www.enzyme-database.org). 25 proteins were aligned to the certain EC (Enzyme Commission) number with exact name and Reaction equation. [see Additional file 3].

To furtherly verified the enrichment analysis results, we applied the genetic strategy using NCBI's Conserved Domain Database (CDD) [35] for the annotation of protein or peptide sequences with the location of conserved domain footprints, and functional sites inferred from these footprints which indicates local or partial similarity to other

proteins, some of which may have been characterized experimentally. 16 out of 53 peptides were queried in different gene super families including ribokinase_pfkB_like super family, PRK super family, GAT_1 super family, DUF1439 super family dnaK super family, fumC_II super family and cupin_like super family [see Additional file 4]. For protein and gene scaled biological analysis, it was showed that the most obvious changes within the five serotypes were concentrate upon enzymes related to nucleoside phosphate and energy metabolism.

Hierarchical clustering to differentiate similarity among *Salmonella* enteric isolates

Due to the importance of *Salmonella* as one of the most important causative pathogens of food-borne diseases, a variety of phenotypic and genotypic methods have been developed to trace the contaminated sources of disease outbreak and to elucidate the epidemiology of infection. The shotgun proteomics approach allowed us to profile and identify nearly all the peptides for the *Salmonella* isolates, which are tightly related to bacterial pheno- and genotypes. So, the peptidomic profiling is applicable to determine the genotypic similarity among different *Salmonella* serotypes.

A hierarchical clustering (HCL) of all peptides was performed in Perseus using Euclidean distances. The distance threshold was defined by the variances of the repetitions of Quality Control (QC) sample and used for the gene-closed cluster identification to distinguish from variance and difference. QC sample was produced by randomly pooling the protein extraction from all the isolates and acquired by LC-MS/MS every 5-10 unknown samples (Figure 4). 25 isolates from training group were clustered and the results showed that none of the 25 isolates have descended from a common ancestor due to their Euclidean distances compare to the QC sample (Figure 5). To furtherly test the HCL efficiency for genotypic similarity analysis, 11 isolates from a food poisoning incident by Enteritidis were analyzed by LC-MS/MS. And 10 Enteritidis isolates had the same trend confirmed by PFGE results except for the strain No.1210 which had a different gene-type with strain 1211 to 1219 (Figure

6).

Figure 4. The injection mode for the peptidomic analysis of *Salmonella enterica* serotypes. Each sample was performed in triplicate randomly. And the QC samples were analyzed by every five unknowns.

Figure 5. Hierarchical clustering to differentiate similarity among Salmonella enteric isolates. The cluster analysis by LC-MS/MS (A) and PFGE (B) In the training group. There was no evident similarity between Salmonella enteric strains.

Figure 6. Hierarchical clustering to identify similarity among Salmonella enteric isolates. The cluster analysis by LC-MS/MS (A) and PFGE (B) in testing group. The strains No.1211 to 1219 were certificated as the gene-closed strains in both ways.

The serotype specific peptidome and biological analysis

As the extra-serotypes peptidomic analysis, we also took a view of intra-serotype difference to evaluated the possible pathway for evolution. Four serotypes including *Enteritidis, Typhimurium, Derby, Rissen* in our experiment were all applied to peptidomic analysis mentioned above. All four serotypes' peptidome was significantly enriched for the GO molecular function terms of catalytic and binding activity [see Additional file 4]. The EC number was searched and the enzymes enriched in Energy metabolism-related phosphorylation and dehydrogenase activity for *Enteritidis,* nucleoside phosphate transferring for *Typhimurium,* Glucose-related energy metabolism for *Derby,* nucleoside phosphate metabolism for *Rissen*[see Additional files 5-8].

DISCUSSION

Comparing to traditional molecular typing methods include antigen-based

serotyping, the LC-MS/MS based peptidomic serotyping could fix the issues like distinguish the new serotypes without any antigen specificity or the coding gene of antigen being mutated. By profiling *Salmonella enterica* strain, both PFGE and LC-MS/MS show the certain "fingerprint" of DNA or proteins. But the LC-MS/MS provides more sensitivity and higher resolution. In the clinical testing, the serotyping and breakout indicating are always carried out parallel, the peptidomic way seems more convenient, time-saving due to its capability of settling it down in one injection which molecular typing method or PFGE can't finish separately.

The SPSS data mining methods in this work totally extracted 6 peptide markers for the serotyping of *Salmonella enterica*. We hope a multiple reaction monitoring (MRM) method can be developed to quantify these target peptides for serotyping, even breakout indicating. The C5.0 predicted model in this research was proven to be effective in *Salmonella enterica* serotyping, but it need to furtherly optimize in the case of increased sample size.

The biological analysis indicates that for extra- or intra-serotypes, the changed proteins usually belong to enzymes. So we suppose if there is a chance to develop some kind of strip holds the test chambers containing dehydrated media having chemically-defined compositions for each test to detect enzymatic activity, mostly related to our reported characters mentioned above like fermentation of carbohydrate or catabolism of proteins or amino acids by the inoculated organisms. By adding the bacterial suspension to rehydrate each of the wells and incubate the strips, metabolism produces a detecable color finally to monitor.

The multiple sample ANOVA t-tests is a widely used statistic tool with high efficiency and easy to spread. In this article, it was capable of distinguishing the five serotypes. For more Salmonella enteric isolates collected, the senior statistic methods including Decision Tree, R-Forest and Support Vector Machine (SVM) should be applied. And peptides fraction strategy seems to be a better option due to its high capability. The peptide markers for different serotypes will be optimized in the future by our developed peptidomic approach.

Conclusions

Food-borne diseases caused by Salmonella enteric serovars represent a serious public health problem worldwide. A quick identification for the pathogens is critical for controlling food pollution and disease spreading. So far, we have developed a peptidomic method to identify epidemiologically important *Salmonella enterica* subsp. enterica serovars and determined the genotypic similarity among *Salmonella enteric* isolates. Compared to the classical White-Kauffmann-Le Minor scheme and PFGE, the LC-MS/MS based peptidomic approach is of equal power for Salmonella enterica serotyping and similarity analysis but with more speediness. Taking 53 most variable serovar-related peptides by label-free quantitative peptidomics combined MS identification, a C5.0 predicted model with 4 predictor peptides was generated with the accuracy of 94.12%.

Abbreviations

LC-MS/MS: Liquid Chromatography with tandem mass spectrometry; PFGE: Pulsed Field Gel Electrophoresis; LFQ: label-free quantification.

DECLARATIONS

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and material: The proteomic data applied for further peptide markers developing and for the hierarchical clustering to differentiate similarity among Salmonella enteric isolates was included in Additional file 11. The dataset for Salmonella enteric serotypes modeling by SPSS Modeler was included in Additional file 9 and 10.

Competing interests: The authors declare no competing interests.

Funding: Not applicable

Authors' contributions: Wang X. performed the LC-MS/MS experiments, analyzed the data and wrote paper draft; Yang Y performed PFGE experiments; Li M collected the strains; Wang L, Zhang P analyzed data; Liang S. instructed experiments and revised paper. All authors read and

approved the final manuscript.

Acknowledgements: This work was financially supported by the grants from Projects of International Cooperation and Exchanges of NSFC (31961143005), Sichuan Science & Technology Program (2020YFH0094) and Health & Family Planning Commission of Sichuan Province (17ZD045).

References

[1] Grimont, P. A. D., F. X. Weill. Antigenic formulae of the *Salmonella* serovars. WHO Collaborating Centre for Reference and Research on *Salmonella*, Institut Pasteur, Paris, France, 2007.

[2] Chen B, Zhang D, Wang X, et al. Proteomics progresses in microbial physiology and clinical antimicrobial therapy. Eur J Clin Microbiol Infect Dis. 2017:36(3):403-413.

[3] Si T, Li B, Comi TJ, et al. Profiling of microbial colonies for high-throughput engineering of multistep enzymatic reactions via optically guided Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry. J Am Chem Soc. 2017 ,139(36):12466-12473.

[4] Chudejova K, Bohac M, Skalova A, et al. Validation of a novel automatic deposition of bacteria and yeasts on MALDI target for MALDI-TOF MS-based identification using MALDI Colonyst robot. PLoS One. 2017, 12(12): e0190038.

[5] Kassim A, Pflüger V, Premji Z, et al. Comparison of biomarker based Matrix Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry (MALDI-TOF MS) and conventional methods in the identification of clinically relevant bacteria and yeast. BMC Microbiol. 2017, 17(1):128

[6] Dieckmann R, Malorny B. Rapid screening of epidemiologically important *Salmonella enterica* subsp. enterica serovars by whole-cell matrix-assisted laser desorption ionization-time of flight mass spectrometry. Appl Environ Microbiol. 2011, 77(12):4136-4146.

[7] Pauker VI, Thoma BR, Grass G, et al. Improved Discrimination of Bacillus anthracis from Closely Related Species in the Bacillus cereus Sensu Lato Group Based on Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry. J Clin Microbiol. 2018, 56(5):e01900-e01917

[8] Fagerquist CK, Zaragoza WJ. Top-down and middle-down proteomic analysis of Shiga toxin using MALDI-TOF-TOF mass spectrometry. MethodsX. 2019, 6:815-826..

[9] Fagerquist CK, Sultan O. A new calibrant for matrix-assisted laser desorption/ionization time-of-flight-time-of-flight post-source decay tandem mass spectrometry of non-digested proteins for top-down proteomic analysis. Rapid Commun Mass Spectrom. 2012, 26(10):1241-1248.

[10] Fagerquist CK, Garbus BR, Miller WG, et al. Rapid identification of protein biomarkers of Escherichia coli O157:H7 by matrix-assisted laser desorption ionization-time-of-flight-time-of-flight mass spectrometry and top-down proteomics. Anal Chem. 2010, 82(7):2717-2725.

[11] Fagerquist CK, Zaragoza WJ, Carter MQ. Top-Down proteomic identification of Shiga Toxin 1 and 2 from pathogenic Escherichia coli using MALDI-TOF-TOF tandem mass spectrometry. microorganisms. 2019, 7(11):488

[12] Blosser SJ, Drake SK, Andrasko JL, et al. Multicenter Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry study for identification of clinically relevant Nocardia spp. J Clin Microbiol. 2016, 54(5):1251-1258.

[13] Tammen H, Hess R. Data preprocessing, visualization, and statistical analyses of nontargeted peptidomics data from MALDI-MS. Methods Mol Biol. 2018, 1719:187-196.

[14] Sriram R, Sahni AK, Dudhat VL, et al. Matrix-assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS) for rapid identification of Mycobacterium abscessus. Med J Armed Forces India. 2018, 74(1):22-27

[15] Nomura F. Proteome-based bacterial identification using matrix-assisted laser desorption

ionization-time of flight mass spectrometry (MALDI-TOF MS): A revolutionary shift in clinical diagnostic microbiology. Biochim Biophys Acta. 2015, 1854(6):528-537.
[16] Thouvenot P, Vales G, Bracq-Dieye H, et al. MALDI-TOF mass spectrometry-based identification of Listeria species in surveillance: A prospective study. J Microbiol Methods. 2018, 144:29-32.
[17] Stein M, Tran V, Nichol KA., et al. Evaluation of three MALDI-TOF mass spectrometry libraries for the identification of filamentous fungi in three clinical microbiology laboratories in Manitoba, Canada. Mycoses. 2018, 61(10):743-753.
[18] Jung RH, Kim M, Bhatt B, et al. Identification of pathogenic bacteria from public libraries via proteomics analysis. Int J Environ Res Public Health. 2019, 16(6): E912.
[19] Sauget M, Nicolas-Chanoine MH, Cabrolier N, et al. Matrix-assisted laser desorption ionization-time of flight mass spectrometry assigns Escherichia coli to the phylogroups A, B1, B2 and D. Int J Med Microbiol. 2014, 304(8):977-983.
[20] Kleinteich J, Puddick J, Wood SA, et al. Toxic cyanobacteria in svalbard: chemical diversity of microcystins detected using a liquid chromatography mass spectrometry precursor ion screening method. Toxins (Basel). 2018, 10(4): E147.
[21] Ullberg M, Lüthje P, Mölling P, et al. Broad-range detection of microorganisms directly from bronchoalveolar lavage specimens by PCR/electrospray ionization-mass spectrometry. PLoS One. 2017, 12(1): e0170033.
[22]Vetor R, Murray CK, Mende K, et al. The use of PCR/Electrospray Ionization-Time-of-Flight-Mass Spectrometry (PCR/ESI-TOF-MS) to detect bacterial and fungal colonization in healthy military service members. BMC Infect Dis. 2016, 22(16):338.
[23] Shah V, Lassman ME, Chen Y, et al. Achieving efficient digestion faster with Flash Digest: potential alternative to multi-step detergent assisted in-solution digestion in quantitative proteomics experiments. Rapid Commun Mass Spectrom. 2017, 31(2):193-199.
[24] Russo R, Valletta M, Rega C, et al. Reliable identification of lactic acid bacteria by targeted and untargeted high-resolution tandem mass spectrometry. Food Chem. 2019, 1(285):111-118.
[25] Yuan G, Bai Y, Zhang Y, et al. Data Mining Mycobacterium tuberculosis Pathogenic Gene Transcription Factors and Their Regulatory Network Nodes. Int J Genomics. 2018:3079730.
[26] Sigurdardottir AK, Jonsdottir H, Benediktsson R. Outcomes of educational interventions in type 2 diabetes: WEKA data-mining analysis. Patient Educ Couns. 2007;67(1-2):21-31.
[27] Wiśniewski JR, Zougman A, Nagaraj N, et al. Universal sample preparation method for proteome analysis. Nat Methods 2009, 6 (5): 359-362.
[28]Rappsilber J, Mann M, Ishihama Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat Protoc. 2007, 2(8):1896-1906.
[29] Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 2008, 26 (12):1367-1372.
[30] Bielow C, Mastrobuoni G, Kempa S. Proteomics quality control: quality control software for MaxQuant results. J Proteome Res. 2016 ;15(3):777-787.
[31] Stefka T, Tikira T, Pavel S, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nature Methods. 2016,13:731–740.
[32] Michael S. B, Joshua B, Vagisha S. An Automated pipeline to monitor system performance in

Liquid Chromatography–Tandem Mass Spectrometry proteomic experiments. J. Proteome Res.2016, 15124763-4769.
[33] Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015;27(2):130-135.
[34] Zhou M, Chen Y, Liu J, Huang G. A predicting model of bone marrow malignant infiltration in 18F-FDG PET/CT images with increased diffuse bone marrow FDG uptake. J Cancer. 2018;9(10):1737-1744.
[35] Aron M B, Yu B, Han L, et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 2017, 4(45): D200–D203.

FIGURE AND TABLE LEGEND

Figure 1. The part of LC-MS/MS spectra of Salmonella enterica serovar-identifying peptide markers in training group.

Figure 2. Hierarchical cluster analysis with 53 peptide markers in training group. 25 isolates from 5 serotypes were divided into 5 clusters without overlap.

Figure 3. The decision tree for the prediction of Salmonella enteric serotypes. The model based on the C5.0(A) and QUEST(B) method.

Figure 4. The injection mode for the peptidomic analysis of Salmonella enterica serotypes. Each sample was performed in triplicate randomly. And the QC samples were analyzed by every five unknowns.

Figure 5. Hierarchical clustering to differentiate similarity among Salmonella enteric isolates. The cluster analysis by LC-MS/MS (A) and PFGE (B) In the training group. There was no evident similarity between Salmonella enteric strains.

Figure 6. Hierarchical clustering to identify similarity among Salmonella enteric isolates. The cluster analysis by LC-MS/MS (A) and PFGE (B) in testing group. The strains No.1211 to 1219 were certificated as the gene-closed strains in both ways.

Table 1. Salmonella enterica subsp. enterica strains (n =42) used in this study.

Table 2. Representative 53 peptide markers detected from 5 serotypes in training groups.

Table 3. The EC numbers and enzyme names of the proteins derived from changed peptides.

Additional files

Additional file 1: The GO enrichment analysis of changed proteins in five Salmonella serotypes(.TIF).

Additional file 2: The GO enrichment analysis of changed proteins in each of four Salmonella serotypes(.TIF).

Additional file 3: The changed proteins derived from 53 peptides markers(.xlsx).

Additional file 4: The annotation of peptide sequences in NCBI's Conserved Domain Database(.xlsx).

Additional file 5: The changed proteins and functional information in Derby(.xlsx).

Additional file 6: The changed proteins and functional information in Enteritidis(.xlsx).

Additional file 7: The changed proteins and functional information in Rissen(.xlsx).

Additional file 8: The changed proteins and functional information in Typhimurium(.xlsx).

Additional file 9: The dataset in training group for serotyping model training by SPSS Modeler(.xlsx).

Additional file 10: The dataset in testing group for the model assessment by SPSS Modeler(.xlsx).
Additional file 11: The quantified data of 7339 peptides by shotgun proteomics based on LC-MSMS(.csv).

Figures

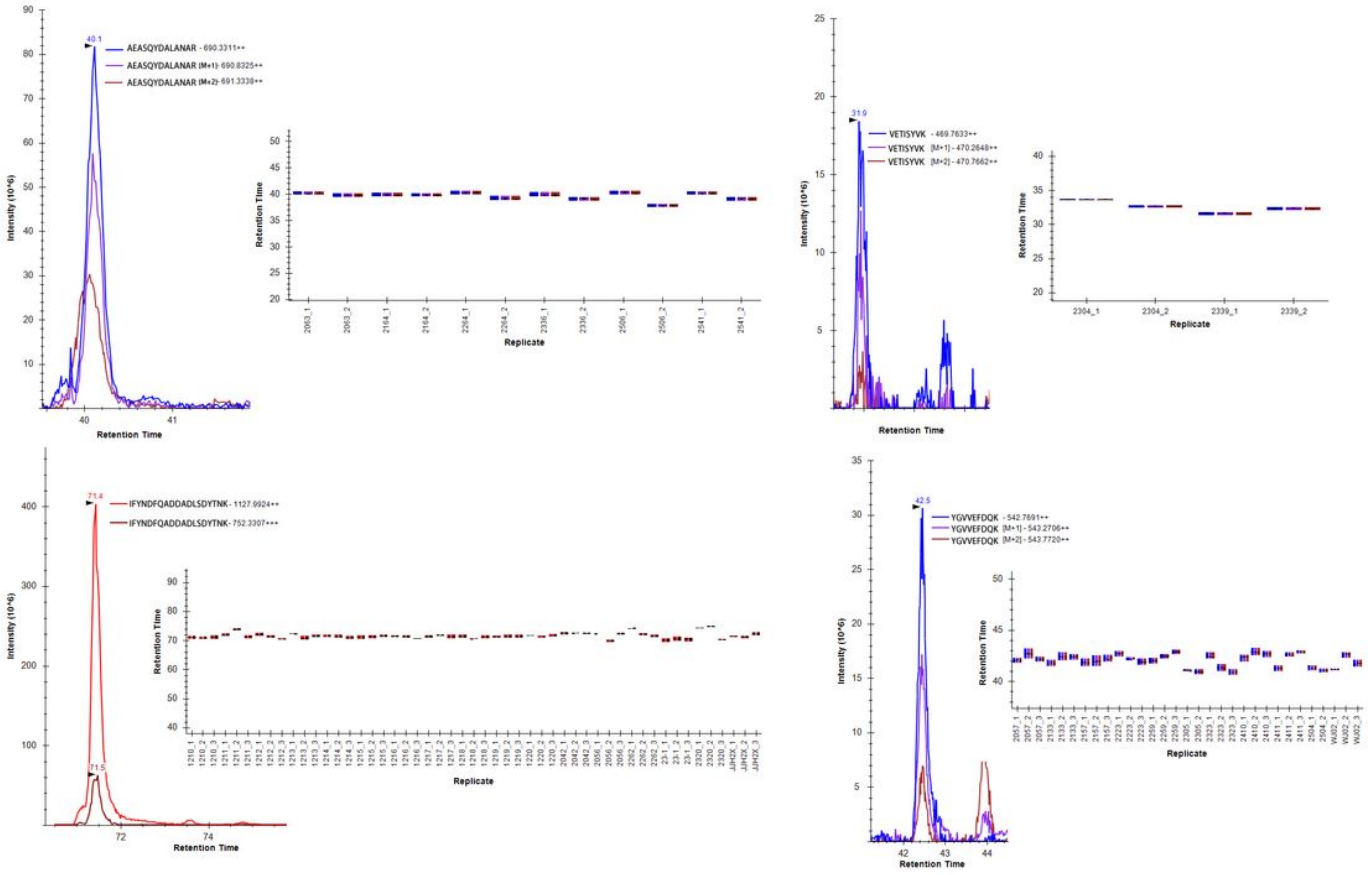

Figure 1

The part of LC-MS/MS spectra of Salmonella enterica serovar-identifying peptide markers in training group.

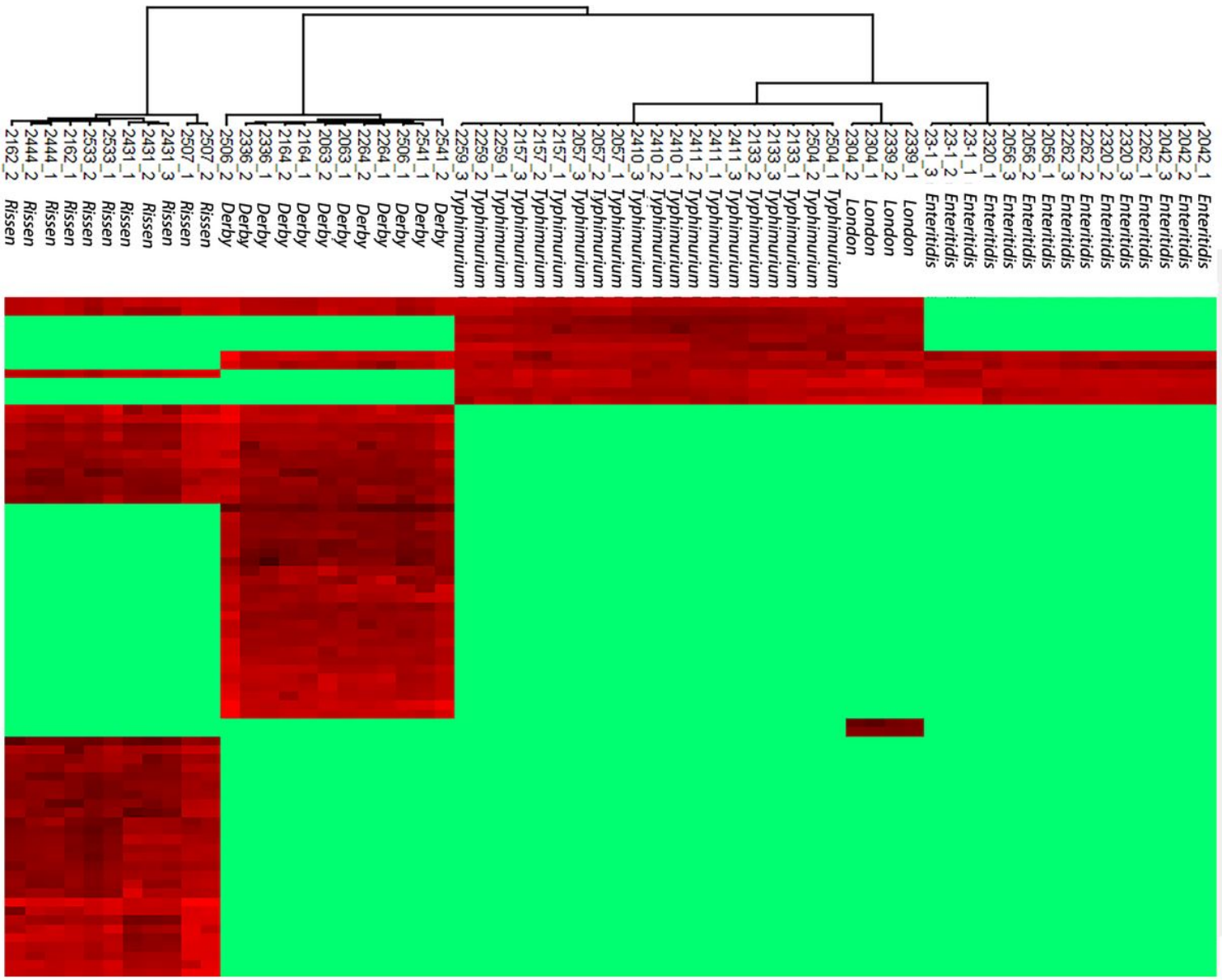

Figure 2

Hierarchical cluster analysis with 53 peptide markers in training group. 25 isolates from 5 serotypes were divided into 5 clusters without overlap.

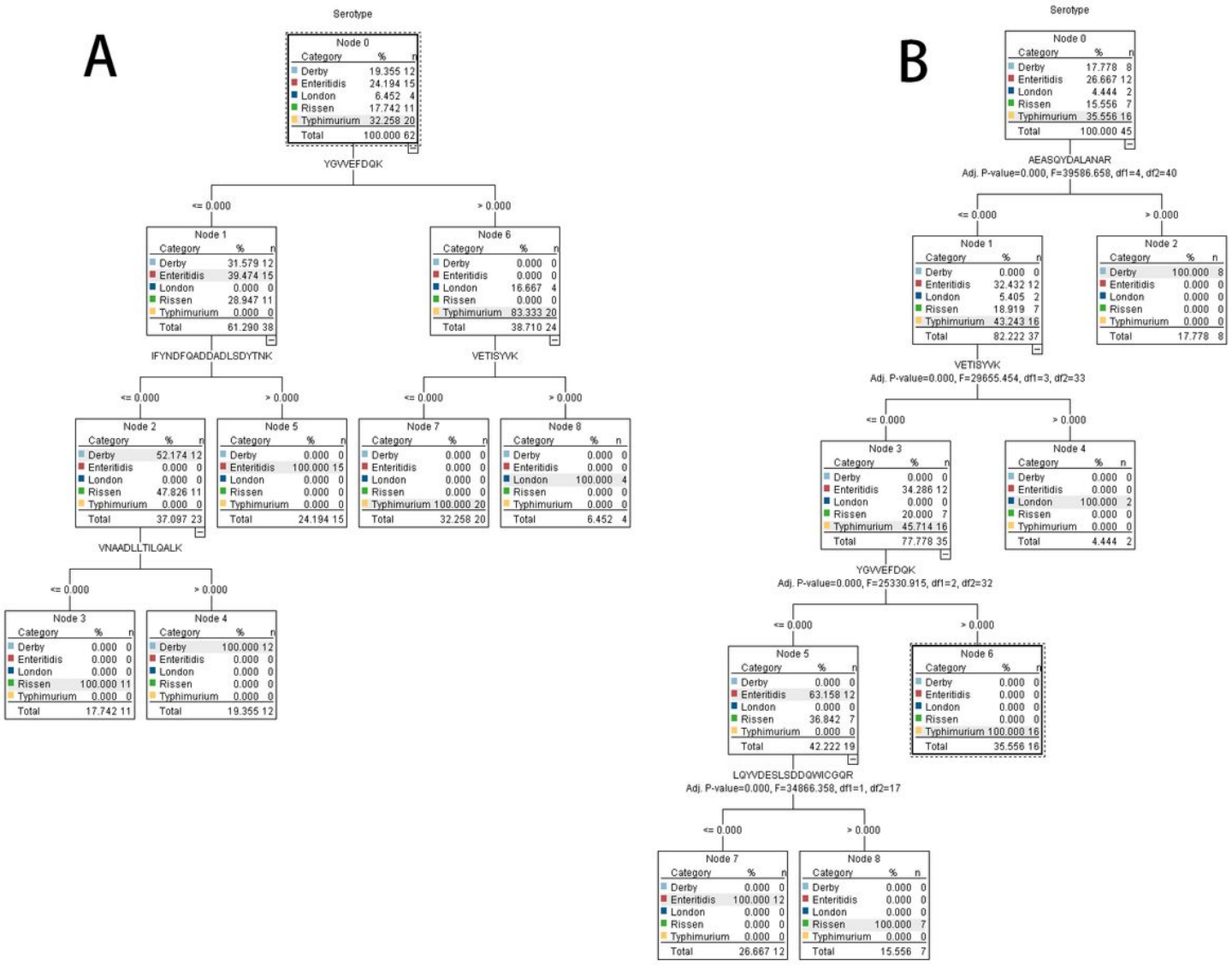

Figure 3

The decision tree for the prediction of Salmonella enteric serotypes. The model based on the C5.0(A) and QUEST(B) method.

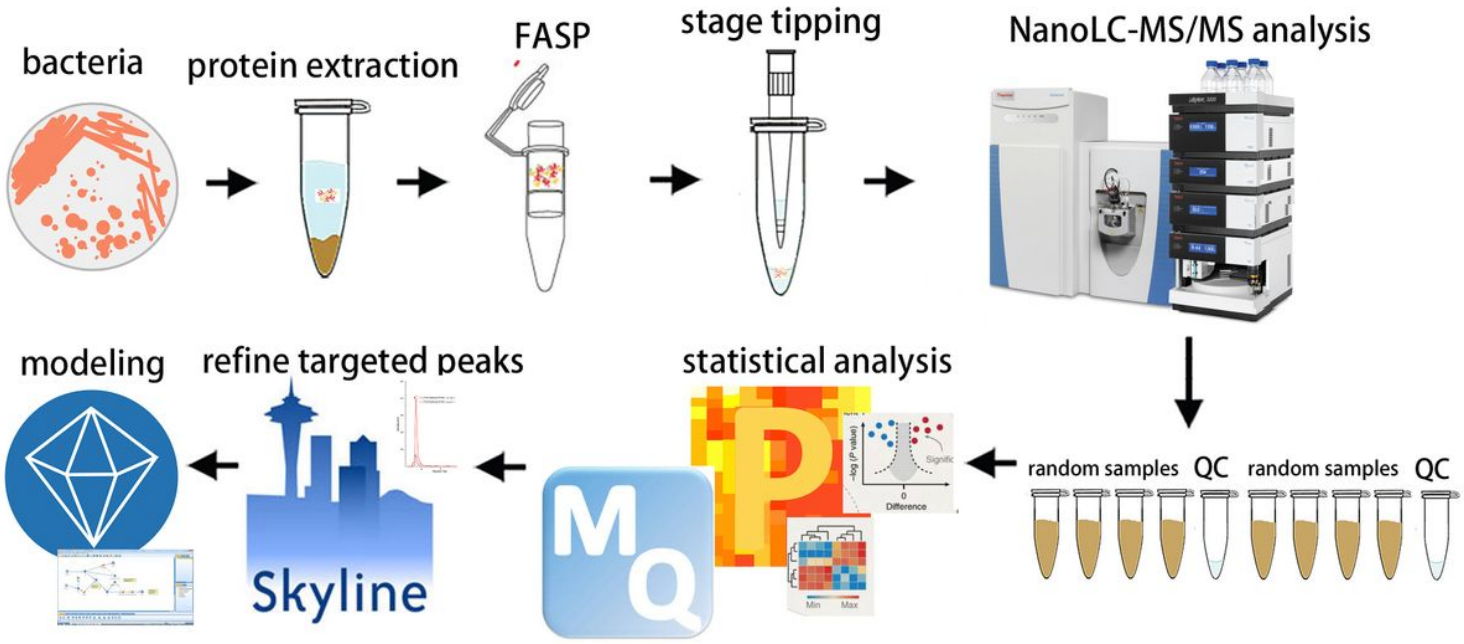

Figure 4

The injection mode for the peptidomic analysis of Salmonella enterica serotypes. Each sample was performed in triplicate randomly. And the QC samples were analyzed by every five unknowns.

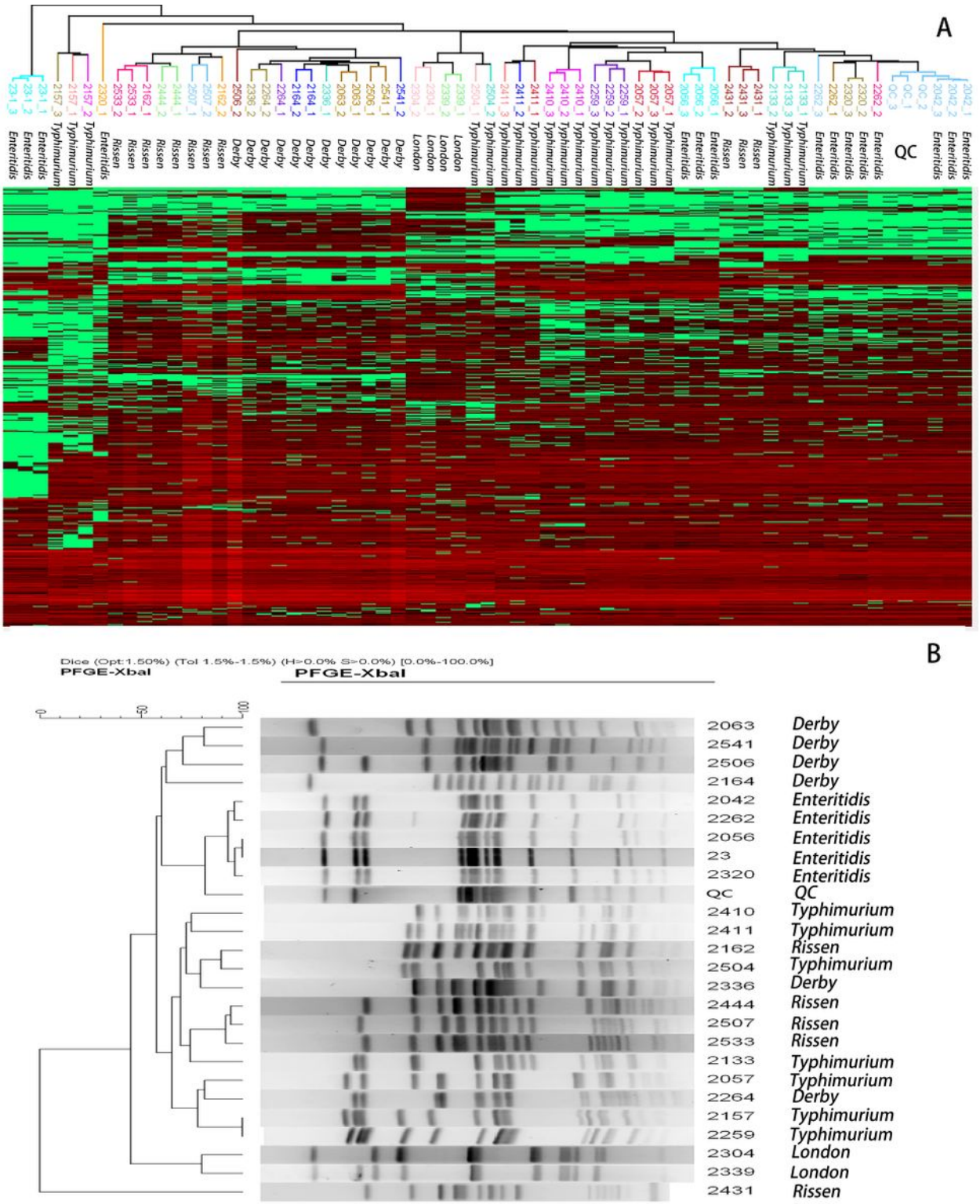

Figure 5

Hierarchical clustering to differentiate similarity among Salmonella enteric isolates. The cluster analysis by LC-MS/MS (A) and PFGE (B) In the training group. There was no evident similarity between Salmonella enteric strains.

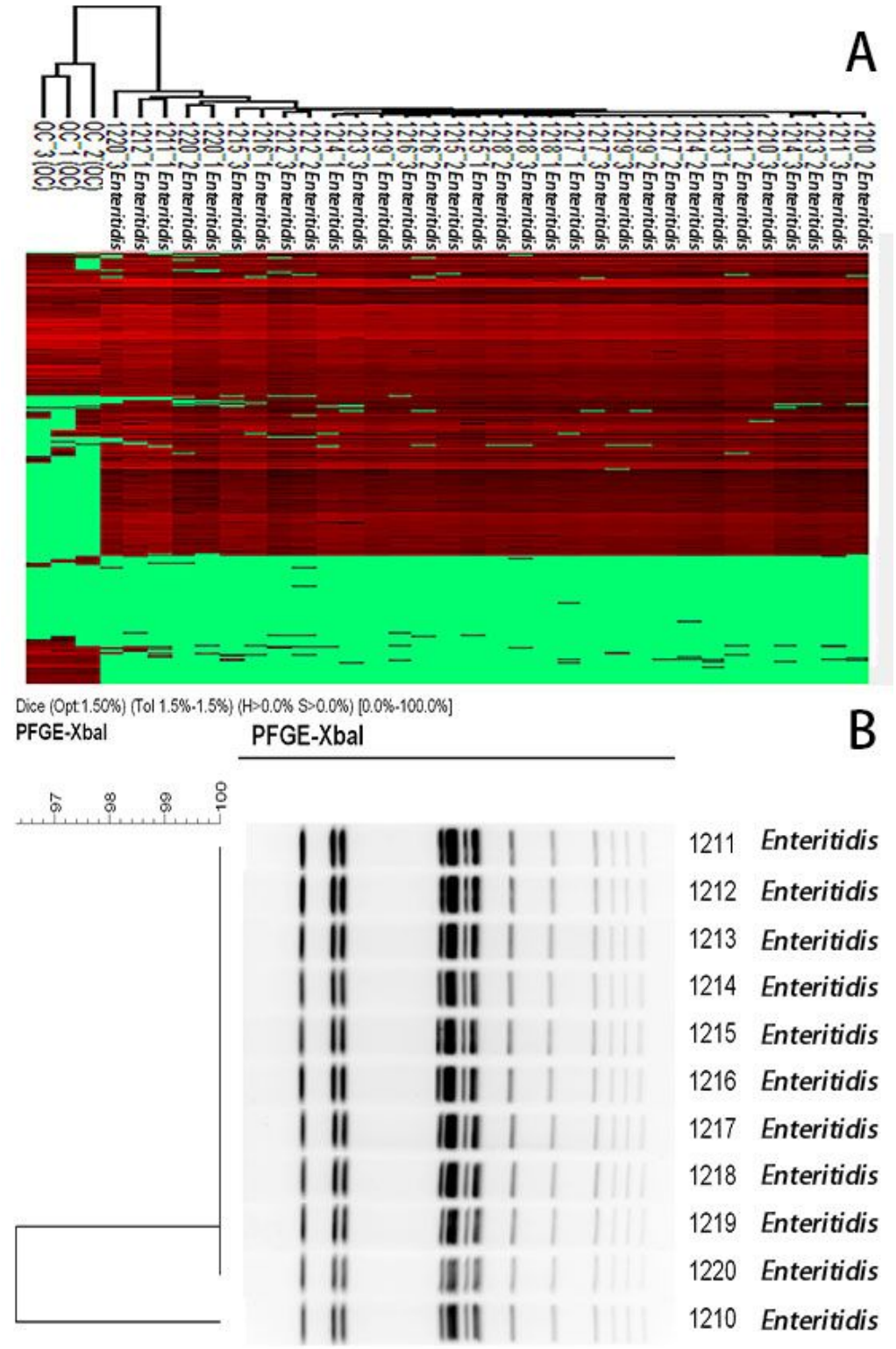

Figure 6

Hierarchical clustering to identify similarity among Salmonella enteric isolates. The cluster analysis by LC-MS/MS (A) and PFGE (B) in testing group. The strains No.1211 to 1219 were certificated as the gene-closed strains in both ways.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1..tif
- Additionalfile10..xlsx
- Additionalfile11..csv
- Additionalfile2..tif
- Additionalfile3..xlsx
- Additionalfile4..xlsx
- Additionalfile5..xlsx
- Additionalfile6..xlsx
- Additionalfile7..xlsx
- Additionalfile8..xlsx
- Additionalfile9..xlsx