

國立台灣大學獸醫專業學院獸醫學研究所

碩士論文

Graduate Institute of Veterinary Medicine

School of Veterinary Medicine

National Taiwan University

Master Thesis

比較 Schwarzengrund 血清型沙門氏桿菌 CRISPR、全

基因體定序與脈衝式電泳、MLST 分子分型方法

Evaluation of the Genotyping methods, PFGE, MLST,

CRISPR, and Whole Genome Sequence Analysis for

***Salmonella enterica* Serovar Schwarzengrund**

吳睿穎 撰

Rayean Wu

指導教授：周崇熙 博士

Advisor: Chung-Hsi Chou, Ph.D.

中華民國 106 年 1 月

January 2017



論文口試委員審定書



國立臺灣大學碩士學位論文

口試委員會審定書

比較 Schwarzengrund 血清型沙門氏菌 CRISPR、
全基因體序列分析與脈衝式電泳、MLST 分子分型方法

Evaluation of the Genotyping Methods, PFGE, MLST, CRISPR,
and Whole Genome Sequence Analysis for
Salmonella enterica Serotype Schwarzengrund

本論文係吳睿穎 (R03629013) 在國立臺灣大學獸醫學
系、所完成之碩士學位論文，於民國 106 年 1 月 9 日承下列
考試委員審查通過及口試及格，特此證明

口試委員：

(簽名)

張紹光 (指導教授)

系主任、所長

(簽名)

致 謝



加入了食品衛生實驗室有五年半了，從茫茫不懂事的大學生到今天碩士畢業，可以很高興的說著實學習不少。這段求學的旅程，最大的感想就是隨著知識灌輸越多、經驗累積越厚實，心態就越謙卑。學海無涯，一路走來無數人扶持、把手拉拔、循循善誘，今天要在短短一段文字致謝還真不夠。

在這裡先感謝周崇熙老師。大學細菌課時一句：『任何想要接觸實驗室的學生，都歡迎來敲我的門。』，在這樣的契機之下我加入了食品衛生實驗室。老師的大方和自由自在的風氣，讓我想要學什麼、做什麼、看什麼都可以。『學生就是要學老師沒學過的。』是老師的口頭禪，秉持著這樣的理念，他鼓勵我們大量閱讀論文、關心報章雜誌、參與研討會跟演講。從廣闊的接觸資訊來尋找研究的方向與解決問題的方法。跟隨老師的這許多年，大開眼界！

然後就是實驗室的同仁，有你們在這幾年實驗室除了是求學的殿堂更是個溫暖的歸所。從大學時期就進來我接受過許多人的指導：揚棋學長、嘉蘭學姊、蟲蟲、大A、小咪、子昇學長、健聰學長、金玉姐、UP、MOMO、阿大、品文學姊、宗承學長、庭維學長、小六、筱芙、峯哥、儂姐、唐維學長。學長姐們教導了實驗技術、給予不少經驗建議，在日常相處也都很照顧。大家聚在一起讓實驗室變成了很溫馨的場所！也感謝學弟妹們，欣霓、勝男、家輝、哲璿，有你們我也學著更督促自己，也時常反省自己，祝你們都能在學習過程找到自己的目標！

另外要再感謝潘世瑩、林之涵、高允凡、劉孟毅學長，我幾位大學的摯友，在我們各自追逐理想的同時，我們彼此扶持、學習跟成長。你們是難能可貴無法取代的一群朋友，謝謝你們在這條路上的陪伴。最後，我最衷心的感謝給我的家人和我的另一半孟蒔，你們是我心裡最大的支持。不論突破或挫折你們都是跟我分享的最好夥伴，你們無怨的支持和耐心的等候讓我有今天的成果。

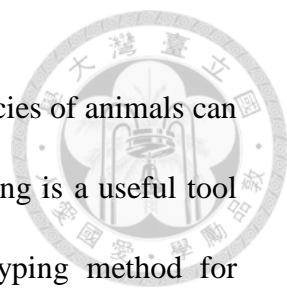
摘要



沙門氏菌是造成多起食物中毒事件的革蘭氏陰性菌。台灣近幾年研究家禽肉品發現在市售的肉雞有超過五成分離到的沙門氏菌都隸屬於血清型 Schwarzengrund。而在人類臨床案例，在 2004—2012 年期間 *S. Schwarzengrund* 為排名第 12 之常見血清型。本研究收集了 2000—2012 期間分離的 15 株來自雞、鴨、豬、寵物飼料與野鳥等流行病學不相關的菌株，外加一株來自同一次爆發病例的菌株共 16 株，想評估傳統基因分型工具脈衝式電泳 (PFGE) 與多基因座序列分析 (MLST) 的分型能力，但發現鑑定效果均不足。PFGE 方面，在使用沙門氏菌最常用的限制酶 *XbaI*，其 *D-value* 值僅 0.79。而 MLST 僅產生一種 Sequence Type 96 (ST96)，則完全無法辨別細菌。因此尋求其他基因分型方法並與傳統分型方式比較。其中 CRISPR 分析，其鑑定能力優於 MLST 和使用 *XbaI*-PFGE，能產生 8 型 CRISPR type，其 *D-value* 值為 0.87，雖不足以單獨運用但推論可以與 PFGE 併用增加分型能力。而全基因體序列 (WGS) 分析，可再進一步依據分析策略分為 MUMi、ND 與 SNV 親緣關係樹等三種不同分析方式。而 SNV、ND、MUMi 個別 *D-value* 值為 0.90、0.94 與 0.79。三種方法中僅 SNV 與 ND 的分析能力高於標準的 0.90。MUMi 對於核酸序列的差異最為敏感但鑑別力不足，因此，本研究中推論其不適合作為血清型內的分型工具。結論上，目前基因定序工具越來越優化，但不同親緣性分析方法的選擇，將對於親緣關係的決定產生影響。本研究發現 CRISPR、SNV 與 ND 分析都可以提供優於 *XbaI*-PFGE 的鑑別力，其中以 ND 分析效果最好。

關鍵字：Schwarzengrund 血清型沙門氏菌、基因分型、鑑別力、全基因體序列分析。

Abstract



Salmonella infections are a public health concern. Several species of animals can potentially transmit these pathogens to humans and molecular typing is a useful tool in epidemiological investigation. To evaluate a plausible genotyping method for *Salmonella* Schwarzengrund, one of the prevalent serotypes in Taiwan, 16 strains with 15 of them being epidemiologically unrelated were genotyped using different methods. Conventional typing methods (*Xba*I-PFGE and MLST) were found to be inappropriate, with discrimination indexes of 0.79 and 0, respectively. Only PFGE with combined results of multiple restriction enzymes (*Avr*II + *Sfi*I) can all unrelated strains be differentiated. For alternative typing schemes, clustered regularly interspaced palindromic repeats analysis generated eight types for the 15 strains, with a discrimination index of 0.87, which coordinated well with the *Xba*I-PFGE phylogenic results and increased the discrimination power. For the whole genome sequence-based analysis, the discrimination indexes of the three approaches utilized (a single nucleotide polymorphism tree, a nucleotide difference tree, and a maximum unique matches index tree) were 0.90, 0.93, and 0.79, respectively. Only the single nucleotide polymorphism tree and the nucleotide difference tree analyses provided sufficient discrimination power (discrimination index > 0.90). The maximum unique matches index tree obtained a lower discrimination index value, even though there is no requirement for a reference genome. In conclusion, the clustered regularly interspaced palindromic repeats analysis, the single nucleotide polymorphism tree, and the nucleotide difference tree all performed better than conventional methods, with the nucleotide difference tree being the best approach for *S. Schwarzengrund*.

phylogenic analysis.

Keywords: *Salmonella* Schwarzengrund, genotyping, discrimination power, whole genome sequence analysis

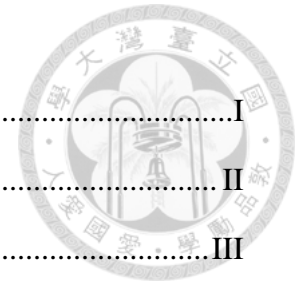


Abbreviation List



BWA	Burrows Wheeler Aligner
CDC	Center for Disease Control and Prevention
CRISPR	Clustered Regularly Interspaced Palindromic Repeats
GATK	Genome Analysis Tool Kit
MEGA 7	Molecular Evolutionary Genetics Analysis Version 7.0
MLST	Multilocus Sequence Typing
MUMi	Maximum Unique Matches index
ND	Nucleotide Difference
SNV	Single Nucleotide Variants
ST	Sequence Type
WGS	Whole Genome Sequence
UPGMA	Unweighted Pair Group Method

Contents



論文口試審定書.....	I
致謝.....	II
中文摘要.....	III
Abstract.....	IV
Abbreviation List.....	VI
Contents.....	VII
Table List.....	IX
Figure List.....	X
Chapter 1. Introduction.....	1
Chapter 2. Literature Review.....	3
2.1. <i>Salmonella</i> Serotype Schwarzengrund.....	3
2.1.1. <i>Salmonella</i>	3
2.1.2. <i>Salmonella</i> Schwarzengrund.....	4
2.2. Conventional Typing (PFGE and MLST).....	6
2.2.1. Pulsed-field gel electrophoresis.....	6
2.2.2. Multilocus sequence typing.....	8
2.3. CRISPR.....	10
2.4. WGS.....	13
2.4.1. From Sanger to high-throughput Sequencing.....	13
2.4.2. Genome Phylogenic Analysis Approaches.....	16
2.4.3. <i>In silico</i> Analysis.....	18
2.5. Discriminatory power of genotyping methods.....	19
Chapter 3. Materials and Methods.....	20
3.1. Strains Identification.....	20
3.2. DNA Extraction.....	22
3.3. Pulse Field Gel Electrophoresis.....	23
3.4. CRISPR.....	24
3.5. Whole Genome Sequence.....	25
Chapter 4. Results.....	28
4.1. Conventional Typing: PFGE and MLST.....	28

4.2. CRISPR	29
4.3. Whole Genome Sequence	30
4.3.1. SNV tree	30
4.3.2. Nucleotide difference tree	31
4.3.3. MUMi tree	31
4.3.4. <i>In silico</i> MLST and antimicrobial resistance gene detection	32
Chapter 5. Discussion	33
5.1. Conventional typing	33
5.2. CRISPR	35
5.3. WGS phylogenic analysis	37
5.4. <i>In silico</i> analysis	38
Chapter 6. Conclusions	39
References	40
Tables	46
Figures	56

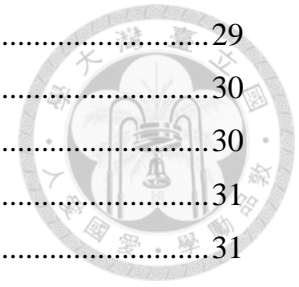


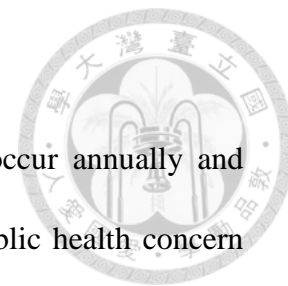
Table List

Tables	46
Table 1. Isolates of <i>Salmonella</i> Schwarzengrund	46
Table 2. Allele type and sequence type of all 16 strains of <i>S. Schwarzengrund</i>	47
Table 3. Primers used for CRISPR allele amplification and sequencing	48
Table 4. Distribution of CRISPR spacers among 16 <i>S. Schwarzengrund</i>	49
Table 5. Number of spacer difference between CTs	50
Table 6. WGS Read profile of 16 <i>S. Schwarzengrund</i>	51
Table 7. <i>In silico</i> MLST typing of 16 <i>S. Schwarzengrund</i>	52
Table 8. Resistance pattern of 16 <i>S. Schwarzengrund</i>	53
Table 9. Corresponding genes to identified SNVs	54
Table 10. The comparison of <i>D</i> -value among all subtyping methods	55

Figure List

Figures	56
Figure 1. PFGE dendrogram of 16 strains digested by <i>Xba</i> I	56
Figure 2. PFGE combined dendrogram of 16 strains digested by <i>Avr</i> II & <i>Sfi</i> I	57
Figure 3. CT correlation with <i>Xba</i> I-PFGE dendrogram	58
Figure 4. SNV tree of 16 strains with strain LT2 as reference	59
Figure 5. SNV tree of 16 strains with strain CVM19633 as reference	60
Figure 6. Nucleotide difference tree of 16 strains	61
Figure 7. MUMi tree of 16 strains	62

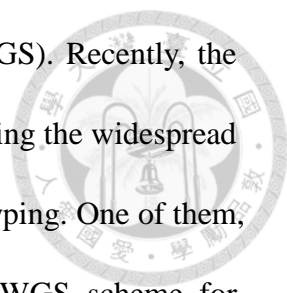
1. Introduction



Approximately 1.3 billion cases of *Salmonella* infections occur annually and result in more than three million deaths, representing a major public health concern (1). Of note, the serotype *Salmonella* Schwarzengrund has been implicated in invasive human infections (2). Recent reports from Taiwan showed that *S. Schwarzengrund* is account for around 50% of *Salmonella* isolates from broiler samples (3), and classified as the 12th most dominant serotype in Taiwan human cases from 2004-2012 (4). Additionally, *S. Schwarzengrund* has been isolated from other sources, such as pigs and pet food as well (5, 6). To track the transmission route across these sources and to elucidate the epidemiological relationships between them, an appropriate genotyping method must be developed.

Conventional typing methods, including pulsed-field gel electrophoresis (PFGE) and multi-locus sequence typing (MLST), have been commonly used for *Salmonella* subtyping (7). However, the information obtained for *S. Schwarzengrund* using these subtyping schemes has been limited. Therefore, in the present study, both methods were performed first to establish a genotyping background.

In contrast, the clustered regularly interspaced palindromic repeats (CRISPR) analysis a newly developed genotyping method. The CRISPR analysis is based on short exogenous sequences called spacers, the polymorphic nature of this allele provides a historical record of foreign genomic elements of bacteria (8, 9). CRISPR has been utilized for *Mycobacterium tuberculosis* subtyping for years (10), and its feasibility has been demonstrated for subtyping *S. Newport*, *S. Typhimurium*, and *S. Virchow* (11-13).



Another promising method is whole genome sequencing (WGS). Recently, the cost and time required for WGS have decreased dramatically, enabling the widespread application of a number of WGS-related approaches for strain subtyping. One of them, single nucleotide variants (SNV), is the most commonly used WGS scheme for bacteria strain phylogenecity analysis. It utilizes a reference genome to identify nucleotide variations among samples and has been used in the investigation of several foodborne outbreaks (14, 15). Another WGS approach, the nucleotide difference (ND) analysis, uses a similar scheme, but was developed to overcome the bias created in different sequencing platforms by using k-mers rather than concatenated sequences (16). Finally, a third approach, the maximal unique matches index (MUMi), calculates the percentage of two genomes that are shared to define the genomic distance between strains; this allows for comparisons to be made without the requirement for a reference genome (17).

In this study, we compared CRISPR and several WGS approaches to determine their efficacy in discriminating between different strains of *S. Schwarzengrund*, and use this information to evaluate whether these methodologies can replace conventional typing techniques.

2. Literature Review



2.1. *Salmonella enterica* Serotype Schwarzengrund

2.1.1. *Salmonella*

Salmonella is a Gram-negative rod bacterium that belongs to *Enterobacteriaceae*, Genus *Salmonella*. Most *Salmonella* have flagella therefore are motile, except for *S. Pullorum* and *S. Gallinum*. Bacterium size is approximately 0.7-1.5 x 2.0-5.0 μm , and non-spore forming. *Salmonella* exists in natural environments and is capable of infecting both warm blood and cold blood hosts. The optimal growth temperature would be 37°C but can survive between 5-45°C. The Genus can be divided into 2 species, *S. bongori* and *S. enterica*, and the latter one can be further divided into six sub species: subsp. *enterica*, subsp. *salamae*, subsp. *arizonae*, subsp. *diarizonae*, subsp. *houtenae*, and subsp. *indica*. As a whole, *Salmonella* contains more than 2500 serotypes.

Salmonella is a great threat to public health. Every year it can cause approximately 1.3 billion of cases and result in more than 3 million deaths (1). According to the European Union (EU) annual surveillance report, the case rate of salmonellosis in 2012 is 21.9 cases per 100,000 population. The top 5 serotypes are *S. Enteritidis*, *S. Typhimurium*, *S. Typhimurium*, monophasic, *S. Infantis* and *S. Stanley*. For prevalence in Taiwan through year 2004-2013, the average cases occurring per year is near 3000, with the top leading 5 serotypes as *S. Enteritidis* (28.1%), *S. Typhimurium* (23.8%), *S. Stanley* (7.8%), *S. Newport* (6.8%), and *S. Albany* (3.7%) (4).

The bacteria are mainly transmitted through contaminated food like undercooked meat, eggs or un-pasteurized milk. Clinical symptoms include gastroenteritis and abdominal pain. In some cases invasive infection could occur, and could be life-threatening to immune-compromised patients or the young and elder (18).

2.1.2. *Salmonella* Schwarzengrund

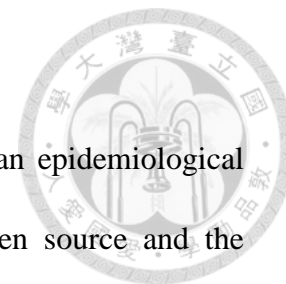
Salmonella Schwarzengrund may not be very common in Europe and America, but for countries in Asia this serotype has seen increased prevalence and importance. It also causes diarrhea, stomach cramps and fever in patients, most patients would recover after 4-7 days, but it is also one of the serotypes that may result in invasive infections, and without proper care and usage of antibiotics may result in severe illness.

For increase prevalence in Asia, in Japan the isolation rate has increased from 0% to 28.1% through year 2000-2003 (19); and in Thailand, isolation rate from chicken raised from 7.2% to 26% through year 2001-2002. The first published international epidemiological study on *S. Schwarzengrund* had suspected multidrug resistant strains to have originated from Thailand and spread to other countries like the USA and Denmark. These assumptions were based on results of the current gold standard for strain subtyping, pulsed-field gel electrophoresis (PFGE) (20).

S. Schwarzengrund has been a serotype of importance in Taiwan. Studies by Chou & Tsai through year 1998-1999 found that more than half (57.5%, 23/40) of the *Salmonella* isolates from broilers were identified as *S. Schwarzengrund* (21). And through 2006-2007, 90 out of 345 strains (second to *S. Albany*, 163 strains) of

Salmonella isolated from retailed local broilers and 50 out of 225 strains (followed by *S. Albany*, 45 strains) from retailed broilers were typed as *S. Schwarzengrund* (3). And the prevalence is not only high among poultry isolates, among human cases through 1998-2002, *S. Schwarzengrund* is in the top 5 most common serotypes for *Salmonella* infections as well (22). For more recent prevalence observations, *S. Schwarzengrund* still dominates the *Salmonella* strains isolated from broilers as it takes up to 39.3% of strains isolated (23). It is also the current 12th most seen serotype among human salmonellosis cases from 2004-2013 (4). Epidemiological observations indicate that *S. Schwarzengrund* is a serotype that needs to be noted in Taiwan.

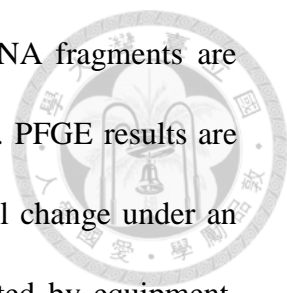
2.2. Conventional Typing methods



The basic unit of all biological diversity is the species. In an epidemiological study, the interests mostly lie with the identification of pathogen source and the transmission route of the disease. Time and geographic concordance of infection combined with genetic resemblance may suggest a common source. Therefore, several genotyping schemes has been developed for *Salmonella* DNA fingerprinting, such as PFGE, plasmid typing, phage typing, amplified fragment length polymorphism (AFLP) and multilocus sequence typing (MLST) etc. Serotyping is still widely used in *Salmonella* studies, but phenotyping schemes is generally believed to lack precision due to disadvantages like limited number of characteristics to be examined, and be misleading with various alternations of gene expression. As a result, current subtyping methods mostly adapt to genotyping schemes, with PFGE currently functioning as the gold standard tool for *Salmonella* discrimination.

2.2.1. Pulsed-field gel electrophoresis

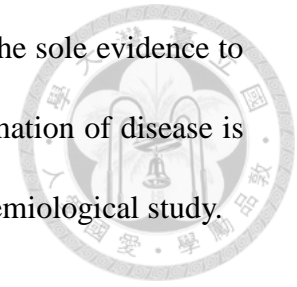
Centers for disease control and prevention (CDC) of the United States established the standard PFGE protocol in 1996. PFGE results were gathered from labs worldwide to create a database for different pathogens. Analysis of PFGE patterns were found to be useful in outbreak differentiation and gained international popularity at the time. PFGE uses different directions of electro fields to elongate electrophoresis duration, stabilize DNA movement allowing large DNA fragments to separate in one run (usually 19-21 hours per run). Combined with restriction enzymes that recognize distinct sequence, different bacteria genome will be cleaved into



fragments of variant sizes. By PFGE, these different sizes of DNA fragments are capable to be visualized and form a unique pattern for each strain. PFGE results are stable and are not easily affected by equipment and environmental change under an experienced practitioner. Unlike PCR, where PCR is easily affected by equipment, reagents, and the random nature of the procedure. This gives PFGE the advantage of communicational data among different labs, thus becoming the gold standard for strain differentiation to this date. The Taiwan CDC reference lab for *Salmonella* was established in 2004, with over 20000 clinical strains. Database includes more than 100 serotypes and over 3000 PFGE patterns. The software Bionumerics developed by Applied Maths, Belgium, currently performs the analysis of PFGE data. Bionumerics can process multiple data types like DNA sequences, antimicrobial drug resistance profile and PFGE fingerprint files to calculate evolutionary distance and establish phylogenetic relations.

Currently, the most popular restriction enzyme used for *Salmonella* PFGE is *XbaI*. Although studies have shown that PFGE alone can be used for serotyping, as certain serotypes correspond to similar patterns (24). It is against advise to imply that one restriction enzyme is the ideal choice for intra-serotype subtyping of all 2500 serotypes. The developers of PFGE claimed that bacteria strains with same PFGE patterns are considered to have common ancestry but should not be evaluated as proof of common source (25). Recently, several studies reported that strains with the same PFGE pattern might still come from different origins, and the combination of multiple restriction enzymes can increase the discriminatory power and accuracy of analysis (26). This indicates that genotyping results merely suggests the resemblance of strains

on the genetic level, and should not be, under any circumstances, the sole evidence to draw connection between pathogen strains. The background information of disease is still extremely important and should always be included in an epidemiological study.



2.2.2. Multilocus sequence typing

Another alternative typing method that is commonly utilized among practitioners is the MLST scheme. The MLST scheme was developed initially for typing *Neisseria meningitides* in 1998 (27). It is a PCR based subtyping method where differences in selected housekeeping genes sequence are detected and categorized accordingly. The basic concept is to search for a certain amount of housekeeping genes (7 genes in *Salmonella's* case) and identify their sequences. Because of the conserved nature of housekeeping genes, every variation in these genes may inflict difference among bacteria strains, even if the variation is just one nucleotide. Each gene is categorized to a specific allele type, and 7 allele types correspond to a Sequence Type (ST). As described earlier, housekeeping genes are prone not to change, thus MLST may not be an efficient tool to detect differences in short time periods (7). However, the MLST has other advantages over PFGE, such as identifying changes at the nucleotide level. This is something PFGE cannot accomplish as it only identifies sizes rather than the actual content of the DNA fragments; And MLST also being a lot cheaper and easier to operate than PFGE.

The 7 housekeeping genes for *Salmonella* MLST scheme are listed below (28):

aroC (chorismate synthase)

dnaN (DNA polymerase III beta unit)

hemD (uroporphyrinogen III co-synthase)

hisD (histidinol dehydrogenase)

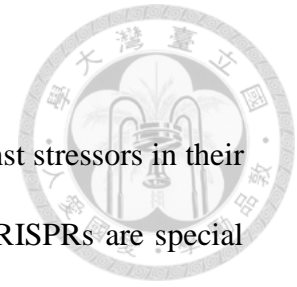
purE (phosphoribosylaminoimidazole carboxylase)

sucA (alpha ketolutarate dehydronase)

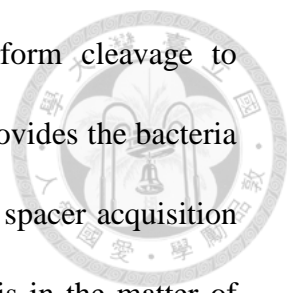
thrA (aspartokinase and homoserine dehydrogenase)



2.3. CRISPR



Bacteria and archaea have adapted to defend themselves against stressors in their environment. And that includes virus or other microbes attack. CRISPRs are special genomic elements found within 70% archaeal and 50% of bacterial genomes. Its structure contains conserved endogenous direct repeats (DR) with sequence size ranging from 20 - 50 base pairs and are interspaced by short similar size exogenous sequences called spacers (29). The CRISPR loci are mainly believed to function as a defense system against foreign genomic elements such as phages and plasmids through the interaction with a group of genes named CRISPR-associated (*cas*) proteins (30). The way CRISPR-*cas* system functions as a defensive mechanism is somewhat similar to a simplified antibody mechanism in our immune system. The process can be divided into three stages in general. The first stage, spacer acquisition, involves integrating genomic elements from foreign phages or plasmid into the CRISPR loci. The term protospacer is used to describe the corresponding sequence on a viral genome to a spacer. In several species the protospacer would have a proximal conservative sequence that appears to be a recognition motif for acquisition, referred as protospacer adjacent motif (PAM). *Cas* protein *cas1* and *cas2* are usually involved in this step. Spacers are integrated between DRs like books in bookshelves, and the whole CRISPR loci like a library. The next step, the main phase involving CRISPR expression, leads to a primary transcript of CRISPR loci. Pre-CRISPR RNA will be cleaved into small CRISPR RNAs (crRNAs) by endoribonuclease. The third phase and last phase, is where defense mechanism occurs. crRNAs can recognize intruding genomic elements that were introduced before with complementary sequence and

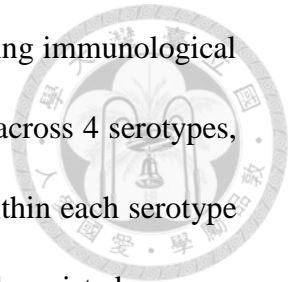


form complexes that allow *cas* proteins to recognize and perform cleavage to eliminate the virus or plasmid. The preservation of CRISPR loci provides the bacteria to have heritable defense (9). Although the specific mechanism of spacer acquisition remains unknown, studies have shown the acquisition of spacers is in the matter of time sequence, thus suggesting a historical trait can be implied, giving great advantage for evolutionary studies (8). Such subtyping methods are not new in practice, epidemiological studies on *Mycobacterium tuberculosis* has been performed under an alternative name spoligotyping for years (10), and studies on *Salmonella* to work as an alternative subtyping tool has been published (11).

Different species contain different numbers or sizes of CRISPR loci, and the evolution of these loci vary also. The genus *Salmonella* possesses two CRISPR loci, with size and spacer content specific to different serovars (11). Several schemes of typing utilizing CRISPR were recommended for *Salmonella*. First, the determination of CRISPR loci size can function as an initial screen easy to perform. If higher discrimination is required, using Sanger sequencing to identify spacer content can provide better insight. Serotypes were found to correlate with certain spacer structures, indicating CRISPR analysis could function as a serotyping tool (11). An inter-serovar subtyping scheme has been introduced by studying the serovar Newport, and proved to be an effective subtyping tool by combining highly variable loci like virulence genes and CRISPRs as parameters (12). Studies on typing serovar Virchow had also proved to be efficient in outbreak differentiation (13).

Even though spacer content may provide additional precision to intra species subtyping, an alternative study on the evolution of *Salmonella* CRISPR-cas system

reached the result that CRISPR may no longer function as an ongoing immunological defense system in *Salmonella*. By assessing more than 600 strains across 4 serotypes, CRISPR spacers were found to be rather conservative, variation within each serotype occurs mostly by loss of spacers or gaining duplicate of originally existed spacers. Thus the plausibility of choosing CRISPR as subtyping tool based on presuming its high evolution rate remains to be debated (31).



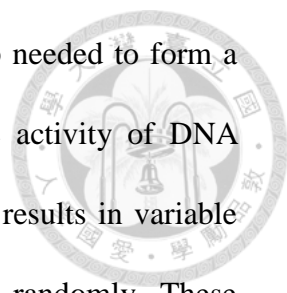
2.4. WGS



Whole genome sequence (WGS) is a powerful tool for genomic investigations and could be applied to evolutionary and epidemiological studies (32). For the past few years there has been a significant decrease in both process time and cost of WGS, allowing it to gain popularity in routine outbreak investigations around the world. In 2011, WGS was first to have functioned as the tool to identify the *Escherichia coli* O104: H4 outbreak, putting its first mark of genotyping on the map (33). It also shows practical application in identifying drug resistance genes or virulence factors. Its usage in genotyping has been proved to be extraordinary in several other *Salmonella* studies as WGS typing successfully distinguished 7 *S. Enteritidis* outbreaks in the USA through year 2001-2014, while gold standard PFGE could only identify four types. Combined databases have allowed various usages of epidemiological studies, when adapted with global positioning system (GPS), allows global pathogen tracing (34). There are variable approaches for WGS genotyping, with the most popular one being single nucleotide polymorphism (SNV) calling; however, with different target samples alternative bioinformatics approaches are required.

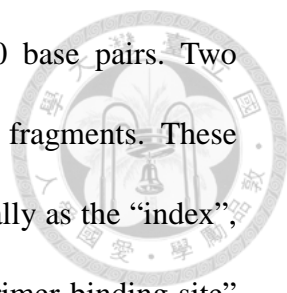
2.4.1. From Sanger to high-throughput Sequencing

Genome sequencing technology has always been the key to understand genome content, and was constantly being improved over the last few decades. The first sequencing method was introduced in 1977 by Frederick Sanger, which utilized the combination of normal deoxynucleosidetriphosphates (dNTPs) and modified di-deoxynucleosidetriphosphates (ddNTPs) during the amplification of genomic



element. ddNTPs act as terminators, as they lack the 3'-OH group needed to form a phosphodiester bond with the following dNTP, thus ceasing the activity of DNA polymerase when added to the chain of DNA in production. This results in variable sizes of synthesized DNA since the ddNTPs are incorporated randomly. These ddNTPs may be labeled radioactively or with fluorescence to help in reading. Through electrophoresis, different sizes of PCR products are separated and the labeled ddNTPs would indicate the nucleotide (A, T, C or G) corresponding to its position. The Sanger method was improved in the 1990's with the introduction of capillary array electrophoresis and detection system, with current technology, could sequence up to 384 of 600-1000 nt length sequences a time. The Sanger method sequencing errors mostly suffers from errors during amplification, natural variance and sample contamination, but altogether is still highly accurate. For the past decade several other strategies of genome sequencing has been developed and available for researchers around the globe. These methods outperform the Sanger method to about 100-1000 times of sequence amount, and simultaneously reducing the cost to 0.5-1% (35). Being the breakthrough of genome sequencing, high-throughput sequencing is also referred as "Next generation sequencing".

In our study we used the sequencing platform by Illumina. The Illumina MiSeq system adapts a sequencing-by-synthesis plot, where the addition of each nucleotide base is detected and identified before the next base is incorporated. This allows the sequence to be obtained without electrophoresis and in real time as well. A sequencing scheme has been adapted to compare results with other platforms of sequencers (36). The procedure is described in the following. The genome of interest



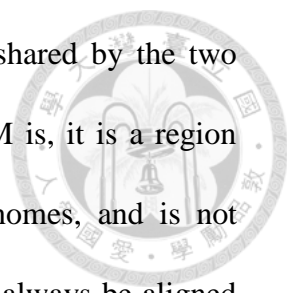
first undergoes sonication and breaks into fragments of 100-200 base pairs. Two different adapters are then added to both 5' and 3' end of the fragments. These adapters contain three sequences combined that functions individually as the “index”, sequence to identify forward or reverse strains, the “sequencing primer binding site” and the “complementary sequence to immobile oligonucleotides”, which such oligonucleotides are fixed to the slide where sequencing are performed. Our fragments that bear the adapters are added to the slide coated with 2 different oligonucleotides and come to complement with them. A polymerase synthesizes the hybridized complementary DNA and the original template is washed away. An isothermal amplification is performed and the fragments bend into an arch form due to complement of both adaptors to the slide, thus creating thousands of identical templates that are gathered proximately, this is called bridge amplification. Now since both forward and reverse templates both exist on the slide, in order to obtain uniform sequence results, the reverse strains are cleaved and the forward templates remain. Fluorescent dNTPs are then added to into the flow and for each nucleotide added a signal would be released and recorded, this allows massive parallel sequencing. After the sequencing of forward templates is finished, the read product is washed away, and bridge amplification is performed again, this time cleaving the forward templates and leaving the reverse templates attached and to undergo sequencing.

After the sequencing process is finished, bioinformatic analysis is introduced to assemble the several Gb of sequence data obtained from forward and reverse templates into a whole genome (37). Identical reads are stacked together (the amount of reads is sometime also referred as depth), and similar sequences are overlapped,

forming a consistent genome sequence. The finished product is then aligned with a known reference genome to check for variances, this is known as “Re-sequencing”. In contrast, another scheme for genome assembly would be “*de novo* sequencing”, where the species genome of interest has not been discovered before, or has no similar reference material available. Since the main purpose of our study is to identify the difference within our 16 strains of interest, the re-sequencing scheme is sufficient for our needs.

2.4.2. Genome Phylogenic Analysis Approaches

Diversity is characterized in the genome level by two major criteria: (1) proportion of un-shared sequences and (2) divergence of the remaining, common DNA. On these bases, two major approaches to analyze genome variance have been developed. The average nucleotide identity (ANI), which detects conservation of core genome, and the DNA content, which calculates the proportion of DNA shared by two genomes. ANI starts by assessing a list of orthologs and calculates the percentage of identical nucleotides of all the orthologs found. Recently, fixed length DNA fragments of the first genome is used to blast against the second genome, and fragments that meet the identity threshold is kept to derive the ANI (38). DNA content calculates distance by assessing the proportion of common genes. By doing so, a list of orthologs is created and the estimation of proximity of two strains by ratio is calculated. Now a question arises, since two methods could both be used as general evaluation, but one being based on gene acquirement and loss, the other based on variances among orthologs, how could the two be correlated together? A new method is developed for genomic distance calculation. This distance is based on the number

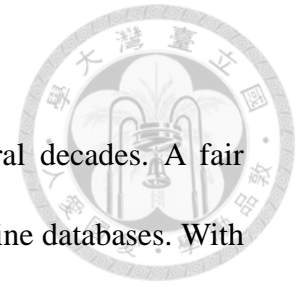


of maximal unique matches (MUM) of a given minimal length shared by the two genomes being compared (17). A more clear explanation of MUM is, it is a region that matches exactly between genomes, exists only once in genomes, and is not contained in a longer such region. The idea is these MUM should always be aligned in the true alignment. The MUM index is a tool to assess distance between two genomes. The index is located between 0 (for very similar) and 1 (for very distant). MUMi was found to correlate better with ANI analysis and less with DNA content analysis. The MUMi distance has been designed to be most sensitive in the range of differences between closely related strains.

Nucleotide difference and SNV calling are both based on calling difference at the nucleotide level. While MUMi focus on comparing how alike two strains can be, ND and SNV analyze the regions that two strains aren't alike. For SNV analysis, reads are mapped against the reference genome and positions with nucleotide variations are identified and extracted to generate a pseudogene consisting the nucleotides of interest. The SNV tree adapts bootstrapping to construct evolutionary relations. The ND tree analysis is designed to overcome data bias between different sequencing platforms (Illumina, Roche & Life technologies). Like SNV analysis, ND also utilizes a reference genome for reads to map against it. The difference is that ND tree will split reference genome and reads into constructed, smaller continuous sequences of k bases called k-mers (in ND analysis k is set to 17, as default settings). K-mers from reads are then mapped against the reference genome to create un-gapped alignments. When mapping is finished, variances are called and calculated, number of differences between each strain is used to draw phylogenic relations (16).

2.4.3. *In silico* Analysis

The study of bacterial genome has been ongoing for several decades. A fair amount of genes has been discovered and organized on several online databases. With the cost of WGS decreasing yearly, bioinformatic tools are also upgraded to process WGS data. In this study we choose to detect MLST and antimicrobial resistance genes directly from WGS data. Such applications are not new to the investigation of infectious diseases, as serotyping and MLST has successfully proved to be plausible to predict identical results as conventional typing methods in a national scale retrospective study of *Listeria* in Australia (15). And antimicrobial resistance genes detected from clinical samples of patients experiencing bladder infections (39). Such results indicate that significant time and resources could be spared.



2.5. Discriminatory power of genotyping methods

To determine the discrimination power of a certain typing method, one could adapt the formula as below:

$$(D) = 1 - \frac{1}{N(N-1)} \sum_{j=1}^s x_j(x_j - 1)$$

D-value Indicates the discriminatory power, *N* the number of unrelated strains tested, *S* the number of different types, and x_j the number of strain in the j^{th} type. *D*-value Can reach the maximum value of 1.0, meaning all strain can be differentiated; whereas minimum value of 0.0, as incapable of distinguishing any strains apart. For a typing method to be reliable, generally a *D*-value value of 0.90 is required (40).

3. Materials and Methods



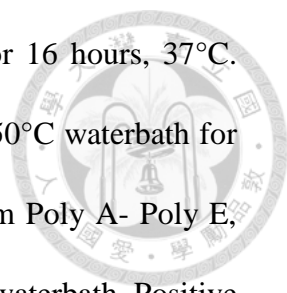
3.1. Strains Identification

This study collected 16 strains of *S. Schwarzengrund* with 15 of them epidemiologically being unrelated throughout year 2000 to 2012 from various sources such as chickens, pet food, stray dogs, wild birds, turkeys, ducks and pigs. Strains are labeled SS01-SS16 with SS16 from the same outbreak of SS15 to serve as a control strain. List of strains, source and isolation year in this study is shown in Table 1. All strains were analyzed with PFGE and MLST prior to this study; the restriction enzymes used for PFGE fingerprinting include *Xba*I, *Avr*II, and *Sfi*I. MLST results indicate all 16 strains to be ST96, detailed allele types is listed in Table 2.

Strains were revived from microbanks and stored at -80°C freezer after 12 hours cultivation, 37°C, in Brain-Heart infusion broth (BHI, Difco, East Rutherford, NJ, USA). Revived bacteria were then cultivated in Trypticase soy broth (TSB, Difco), 37°C, for an additional 12 hours before any further procedures. All strains were confirmed again as serotype *S. Schwarzengrund* through White-Kauffmann-Le Minor serotyping scheme (41).

Coagulation test of somatic O antigen was performed with mixture of 3-5 µL bacteria and anti-serum, positive results form coagulation on slides of mixture. All strains were confirmed with positive results of Poly A-I and Vi and further with Poly B, group O: 4(B), factor 1-4-12-27 antisera.

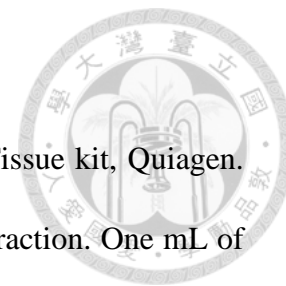
Flagella H antigen identification is performed in 2 phases. And 0.85% of NaCl solution was prepared and mixed with 37% formalin to obtain 0.6% formalin NaCl



solution. For phase I, bacteria is cultivated in 3.5 mL of TSB for 16 hours, 37°C. Broth is then mixed with 3.5 mL 0.6% formalin NaCl solution in 50°C waterbath for an hour to fix flagella. Retrieve 0.5 mL of *Salmonella* H antiserum Poly A- Poly E, and mix with same amount fixed bacteria broth for 1 hour in 50°C waterbath. Positive results should form cotton-like coagulation within tube in an hour. The expected H1 result for *S. Schwarzengrund* is factor d, as all strains performed.

In phase II we prepared 0.35% agar of TSB, retrieved 3 mL and mixed with 0.5 mL of *Salmonella* H Antiserum Single Factor d. A plastic tube was carefully inserted into the middle of the semi-solid agar. A single colony of bacteria was then inoculated within the plastic tube just beneath the surface, and underwent 37°C cultivation to wait for bacteria to grow out of the tube. Tube was checked every 2 hours until bacteria reaches 2-3 mm beneath the surface of the exterior agar, then inoculated in 3.5 mL TSB to undergo another 6-8 hour 37°C cultivation. Then repeat procedures in phase I. The expected phase II result should be factor 1, 7, as all strains performed.

3.2. DNA Extraction



DNA extraction was performed with DNeasy® Blood and Tissue kit, Qiagen. All strains were cultivated in TSB, 37°C, for 12 hours prior to extraction. One mL of broth was retrieved and centrifuged at 5,000 xg for ten minutes, then discard the supernatant. The bacteria pellets were then treated with 200 µL of 10 mg/mL lysozyme for 30 minutes in 37°C waterbath, before following the procedures of DNeasy® Blood and Tissue kit.

180 µL of ATL was added into the mix with 4 µL of RNase A and incubated at room temperature for 10minutes. Added 20 µL of proteinase K and incubated in 55°C waterbath for 1.5 hours. Added 200 µL of AL buffer and 200 µL 100% EtOH, mixed thoroughly and gently. Then Added solution to column and centrifuge at 6,000 xg for 2 minutes. Renewed collection tube and add 500 µL of AW1 buffer and centrifuge at 6,000 xg for 2 minutes. Renewed collection tube and added 500µL of AW2 buffer and centrifuge at 20,000 xg for 4 minutes. Replaced collection tube with 1.5 mL eppendorf; added 50 µL ddH₂O into column and centrifuged at 6,000xg for 2 minutes. All extracted DNA qualities were checked with Nanodrop 1000. Qualify standards are set to 1.8-2.0 for OD_{260/280} value, and 1.8-2.2 for OD_{260/230}. Extracted DNA was stored at -20°C.

3.3. Pulse Field Gel Electrophoresis

PFGE were performed prior to this study. Protocol follows the standard operation procedure of PulseNet USA for *Salmonella*. Restriction enzyme *Xba*I, *Avr*II and *Sfi*I were used to perform restriction digestion.

Results were analyzed by software Bionumerics; phylogenic tree was inferred using the unweighted pair group method with arithmetic mean (UPGMA) method. Band tolerance was set to 1.5% and optimization at 1%. Discrimination power was evaluated using the Discrimination index *D*-value (40).

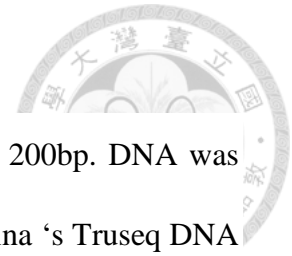


3.4. CRISPR



Two CRISPR alleles (CRISPR1 and CRISPR2) exist in *Salmonella*; both were targeted with primers designed in previous study (42), and is listed in Table 3. PCR amplification was performed with *Taq* master mix kit. A 25 μL system contained 12.5 μL of *Taq* PCR 2X master mix, 9.5 μL of PCR grade water, 1.0 μL of DNA template, 1.0 μL of forward primer (final concentration 1.0 μM) and 1.0 μL of reverse primer (final concentration 1.0 μM). A single PCR cycling was used for all primers and is listed as following: initial denaturing for 10 minutes at 95°C; 45 cycles, each cycle containing 1 minute at 95°C, 1 minute 30 seconds at 55°C and 1 minute 30 seconds at 72°C; final extension step of 10 min at 72°C. PCR products underwent electrophoresis at the condition using 1.5% agarose gel with cyber green dye, 100V, runtime 45 minutes. PCR product was then sent for sequencing (PURIGO biotechnology Ltd, Taipei, Taiwan). Analysis of CRISPR1 and CRISPR2 was conducted using CRISPR-finder (<http://crispr.u-psud.fr/server/>) (43), A final CRISPR type (CT) was given to each strain accordingly to the composition of its spacers. Discrimination power was evaluated using the Discrimination index *D*-value.

3.5. Whole Genome Sequence



Target DNA was sonicated to the size ranging from 180 to 200bp. DNA was then end-repaired, A-tailed and adaptor-ligated following the Illumina 's Truseq DNA preparation protocol, and the product DNA library were validated by Agilent 4200 TapeStation (D1000 screen tape) to check the library fragments at the recommend size. Each library was barcoded and sequenced on a NextSeq500 as paired-end 150 reads (PE150). All reads files were uploaded. Low quality reads (passing filter and $<Q20$) or adaptor-contaminations (>6 Bases aligned with adaptor sequence) were removed, and then the qualified reads were performed to Base-calling. The qualified reads data then went through a genomic alignment against Ensemble database using BWA to get basic sequence information. To further variants analysis, Genome Analysis Toolkit (GATK) created variants calling and annotations. BaseRecalibrator were used to detect systematic errors in base quality scores, and UnifiedGenotyper were used to call SNVs and indels on a per-locus basis. All tools were set as standard settings. Qualified SNVs were selected once they met the threshold of reaching a minimum coverage of 15 reads.

SNV tree was inferred using the UPGMA method. Two reference strains were used to create to results, *S. Schwarzengrund* CVM19633 (Accession No. PRJNA19459) and *S. Typhimurium* LT2 (Accession No. PRJNA241). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) is shown above the branches. The evolutionary distances were computed using the UPGMA method. Positions containing gaps and missing data

were eliminated. Evolution analyses were conducted in the Molecular Evolutionary Genetics Analysis program version 7 (MEGA7).

For phylogenetic analysis, the Nucleotide difference tree (ND tree) was generated with known *S. Schwarzengrund* strain CVM19633 as reference and using the pipeline tool on the Center for Genomic Epidemiology (<http://www.cge.cbs.dtu.dk/services/NDtree/>) with default settings. When all reads had been mapped, the significance of the base called at each position was evaluated by calculating Z score from X (the most common nucleotide of that position) and Y (number of other nucleotides). The Z-score is calculated as $Z=(X-Y)/\sqrt{X+Y}$. The value 1.96 was used for Z threshold to a p-value of 0.001. Further requirement of $X>10*Y$ is proceeded.

MUMi analysis can be derived using the following formula: $MUMi = 1 - (L_{mum}/L_{av})$, where L_{mum} is the sum of the length of all nonoverlapping MUMs and L_{av} is the average length of the two genomes to be compared. MUMi values close to 0 represent very similar sequences, while values close to 1 are gained for very distant genomes. In brief, each pair of genome sequences were detected twice (reference vs. query, and the reciprocal order) for lists of shared MUMs using Mummer3 software version 3.23 (<http://mummer.sourceforge.net/>) (44) with the following parameters: -mum, -b, -c, and -l 19. Mummer3 results were parsed for nonoverlapping MUMs, and then an average MUMi value was calculated for each pair of genomes. All MUMi values from 16 genomes were then outputted as a distance matrix file for the use of constructing a Neighbor-joining tree using the MEGA version 7 (<http://www.megasoftware.net/>) (45).

For *in silico* analysis, web-tools developed by Center for Genomic Epidemiology, Resfinder 2.1 (<https://cge.cbs.dtu.dk/services/ResFinder/>) (46) was used for antimicrobial resistant genes detection, and MLST 1.8 (<https://cge.cbs.dtu.dk/services/MLST/>) used for MLST typing. Both web-tools use blastN for gene detection. Paired-end reads were submitted and analysis parameters were set at default settings.

4. Results



The evaluation data consists of a set of 16 strains of *Salmonella* Schwarzengrund from 15 epidemiological unrelated origins. The performance of differentiation power is calculated with the discrimination index.

4.1. Conventional Typing: PFGE and MLST

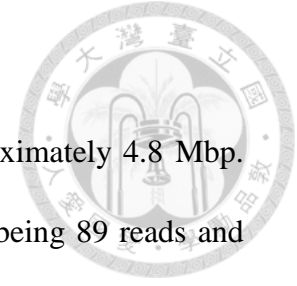
The PFGE method with digestion by the restriction enzyme *Xba*I could not fully distinguish all 15 strains of *S. Schwarzengrund* (Fig. 1), with a *D*-value of 0.79. However, after changing to two alternative restriction enzymes (*Avr*II and *Sfi*I) and combining respective results all strains were distinguishable (Fig. 2), and the internal controls SS15 and SS16 were closely grouped, as expected. Conversely, the MLST method failed to distinguish the strains from each other and identified all of them as ST 96 (Table 2).

4.2. CRISPR

The PCR product size of CRISPR1 and CRISPR2 loci vary (but not with great difference) between the 16 strains of *S. Schwarzengrund* of interest, ranging from approximately 900 to 1,100 base pairs. PCR products are then sent for sequencing and submitted to crispr-finder to identify CRISPR structures. A total of 14 spacers in CRISPR1 and 17 in CRISPR2 were found respectively based on the sequencing results. Strains were given CTs according to the content of the spacers (Table 4). Among the 16 strains, five CRISPR1 allele patterns and three CRISPR2 allele patterns were found, creating eight unique CTs with a *D*-value of 0.87.



4.3. Whole Genome Sequence



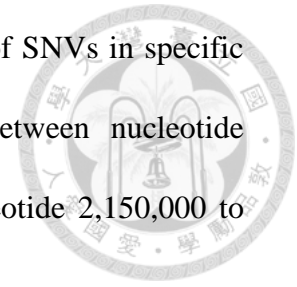
The actual size of the *S. Schwarzengrund* genome is approximately 4.8 Mbp. The average depth of reads was 184, with the lowest (SS03-Ck) being 89 reads and the highest (SS14-Dk) being 262 reads. The reference genome (CVM19633) coverage was above 90% for every tested strain (Table 6.).

4.3.1. SNV tree

SNV tree was conducted 2 times with different reference strains, *S. Typhimurium* strain LT2 and *S. Schwarzengrund* strain CVM19633. When using LT2 (*S. Typhimurium*) as the reference genome, the SNV analysis identified more than 30,000 SNVs and all strains were indistinguishable from one another on the phylogenetic tree (Fig. 4). Conversely, when we used CVM19633 (*S. Schwarzengrund*) as a reference genome, the SNV analysis identified 122 qualified SNVs. However, when constructing the phylogenetic tree, the internal controls SS15 and SS16 failed to cluster, and SS04 was calculated as having a higher genetic similarity with SS16 (Fig. 5). As a result, strains that have relatively fewer genetic differences than SS15 and SS16 should be defined as genomically indistinguishable. By this definition, SS09, SS10, SS11, and SS14 were considered indistinguishable, as well as the following pairs: SS12 and SS13, SS01 and SS02, SS03 and SS05, SS07 and SS08, and SS04, SS15 and SS16, with an acquired *D*-value of 0.90.

A total of 122 SNVs were traced back to each location on the whole genome and only two SNVs were located on known genes: the heme lyase subunit of *NrfE* (*ccmF*) and ribosome recycling factor (*rrf*), with neither of the SNVs resulting in any

amino acid coding changes (Table. 9). Furthermore, a clustering of SNVs in specific regions were noted; specifically, 72 SNVs were identified between nucleotide 610,000 to 620,000 and 32 SNVs were identified between nucleotide 2,150,000 to 2,165,000.



4.3.2. Nucleotide difference tree

The ND analysis calculates nucleotide differences using the reference genome CVM19633 and was used to generate the phylogenic tree in Figure 6. The 16 strains were differentiated into 10 subtypes with a *D*-value of 0.93, with some strains (SS03 and SS05; SS07 and SS08; SS09 and SS14; SS10 and SS11; SS04, SS15 and SS16) remaining indistinguishable.

4.3.3. MUMi tree

The MUMi tree calculates the percentage of the genome sequence that two bacteria strains share in common. The higher proportion of the genomes that are identical, the closer they are on a phylogenic tree. The phylogenic tree draws out the relative genomic distance between strains, and in this study the MUMi tree is the only WGS approach to detect differences between all strains (Fig. 7). However, the genomic distance between our internal controls SS15 and SS16 was approximately 0.02%, strains with a distance less than this should be defined as indistinguishable. Therefore, strains SS06, SS07, SS08, SS09, SS10, SS11 and SS14 were indistinguishable from each other, as well as SS03 and SS05; SS01 and SS02, which resulted in a final *D*-value of 0.79.

4.3.4. *In silico* MLST and Antimicrobial Resistance Gene Detection

The *in silico* MLST typing method recognized all 16 strains as ST 96 (Table. 7), which agreed with the results obtained from the actual MLST typing. The results of the *in silico* antimicrobial resistance gene detection agreed with the majority of the results from the disk diffusion test, with the exception of the quinolones (Table 8), as well as the chloramphenicol resistance for SS01, SS02 and SS03. The close resemblance of these results to those of conventional methods suggests that the web-tools (46, 47) could be an alternative tool.

5. Discussion

5.1. Conventional typing



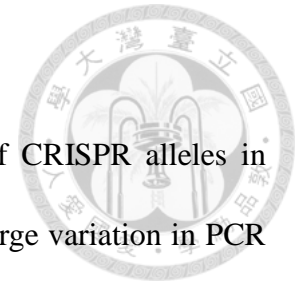
PFGE is regarded as the gold standard for bacterial genotyping, and *Xba*I is the restriction enzyme that is primarily used to differentiate *Salmonella* strains in the PulseNet database. In the current study, we found that the *D*-value for *Xba*I-PFGE didn't reach 0.90, indicating that there may be false or incomplete epidemiological relationships determined when solely using this typing method. Our findings are in agreement with a previous study on *S. Schwarzengrund* subtyping, in which 23% of the studied strains gave rise to only three *Xba*I-PFGE patterns, among 581 multi-drug resistant strains (20). Similarly, Chen and colleagues reported that 66% of *S. Schwarzengrund* isolates taken from chicken retail meat in Taiwan had three dominant *Xba*I-PFGE patterns (48). The lack of divergent patterns suggests that *Xba*I-PFGE has insufficient power to discriminate between *S. Schwarzengrund* strains. The results of our study pertaining to *Avr*II and *Sfi*I, indicate that using a combination of the two restriction enzymes can improve the discrimination power. However, the practice of utilizing multiple restriction enzymes is expensive and time consuming. Additionally, the amount of international comparative information is limited, making this method less practical.

Within the MLST database (<http://mlst.warwick.ac.uk/mlst/dbs/Senterica>) for *Salmonella*, four STs (ST 96, ST 241, ST 322, and ST 848) cover all 18 published *S. Schwarzengrund* strains. Among them, 13 strains isolated from different locations and times were classified as ST96. Similarly, in our study all 16 strains were also

identified as ST 96, indicating that MLST provides insufficient discrimination power.



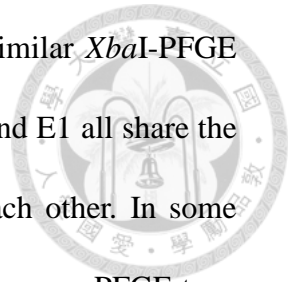
5.2. CRISPR



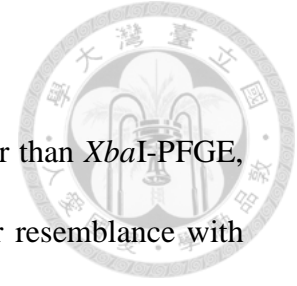
A previous study by Shariat *et al.* reported that the sizes of CRISPR alleles in studied *S. Newport* strains ranges from 300 to 1,700 bp, and the large variation in PCR products can differentiate between strains (11, 12). The large variation in PCR product sizes has been reported in other serotypes as well (11). However, in our study, the size of the PCR products at the CRISPR1 and CRISPR2 loci had a much smaller range: from approximately 1,000 to 1,100 bp, which limited the discrimination between different strains of *S. Schwarzengrund*. In a result, 5 different CRISPR1 alleles (labeled A-E) and 3 different CRISPR2 alleles (labeled 1-3) were identified, and the combination created 8 CTs in total. According to the results, this method provides a better resolution in strain differentiation compared to PFGE, as *D*-value of CRISPR typing is 0.87. Since no novel spacers were detected within *S. Schwarzengrund*, all spacers in this study were already existent in the CRISPR database. Spacer variance between strains occurs through the gaining and losing of existing spacers; such phenomena have been noted in other serotypes of *Salmonella* as well (31).

It has been studied that closely related *Salmonella* strains were found to share more identical spacers with one another (9, 11). In our study, the sums of spacer difference among the 16 tested *S. Schwarzengrund* strains are listed in Fig. 3. Genomic differences between CTs were evaluated by sum of spacer differences in CRISPR1 and CRISPR2 alleles. For example, CT pair of A1 and B1 was considered genomically more closely related than the pair of A1 and C2, for the former contained two spacers in difference while the later contained eight. The distribution of CTs was further compared against the *Xba*I-PFGE phylogenic tree (Fig. 3). The two methods

coordinated well with each other: most strains with identical or similar *Xba*I-PFGE patterns had the same or closely-related CTs, such as CTs A1, B1 and E1 all share the same PFGE-*Xba*I pattern with only 2 spacer variations among each other. In some cases, we were able to further discriminate between strains with the same PFGE type by using the CTs. This observation is in agreement with a previous study (12) and supports the possibility of combining the two subtyping methods to increase the discrimination power for *S. Schwarzengrund*.



5.3. WGS phylogenic analysis



The SNV tree has D -value of 0.90, though performing better than *Xba*I-PFGE, failed to cluster SS15 and SS16 together, as SS04 showed higher resemblance with SS16. The ND analysis achieved similar results, while SS15 and SS16 were clustered together; SS04 is still indistinguishable from the control pair strains with the acquired D -value for ND analysis is 0.93. The ND analysis performed better than the SNV tree because the alignment process in ND is based on the mapping of k -mers, whereas the SNV analysis requires concatenated sequences for alignment. However, the strains used in this study are all epidemiologically unrelated, even though both D -values reached 0.90 neither the SNV or ND analysis was able to fully differentiate among all strains. A report by Kwong *et al.* emphasized the importance of utilizing reference strains with high resemblance to the tested strains, such as strains of the same serotype, in WGS phylogenic studies to improve the accuracy of the analysis (15). Our study supports this claim as using a reference genome of a completely different serotype (*S. Typhimurium*) for *S. Schwarzengrund* genotyping has been proved fugitive. However, in our study we showed that even using a reference genome of the same serotype still might not be enough for complete differentiation. The reference strain CVM19633 used in the *in silico* analysis for antimicrobial resistant genes detection was found to have great difference in drug resistance profile with tested strains. It is possible that the large difference in genetic contents prevented CVM19633 from being an ideal reference genome for *S. Schwarzengrund* differentiation in this study.

Regarding the MUMi analysis, to the best of the authors' knowledge, this is the

first report to compare MUMi with other WGS-based genotyping schemes, and provided a poor result. According to a study by Deloger *et al.*, the utilization of MUMi was originally for the taxonomic designation of closely related species rather than for evolutionary-phylogenetic studies (17). Therefore, in the present study, MUMi was the least appropriate approach for intra-serotype subtyping, as we believe that the difference detected between internal controls was overly sensitive.

5.4. *In silico* Analysis

The web-tools (46, 49) developed by Center of Genomic Epidemiology are open source for anyone to use and capable of processing data outputted from different sequencing platforms. This brings convenience for researchers across the globe to compare NGS data.

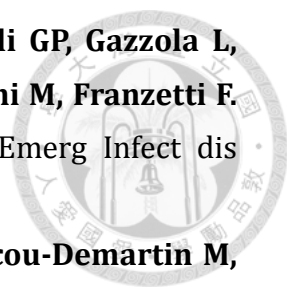
The MLST scheme is popular in the use of subtyping *Salmonella*, and in recent years utilized for serotyping as well (50). In our study, the 7 house keeping genes of the 16 strains of *S. Schwarzengrund* were all perfectly predicted as the same results with the conventional method. All strains have unanimous results as ST 96. The length of time required for each strain's analysis is significantly less than the time conventional PCR requires. As for Antimicrobial gene detection, *in silico* analysis (46) also performed accurately to previous results, with the majority of results in line with conventional antimicrobial tests. Only Phenicol's weren't predicted in SS01, SS02 and SS03 and quinolone drugs in all strains. The web-tool was specific that it could only detect anti-fluoroquinolones genes, but still failed to predict any strains with drug resistance to ciprofloxacin. *In silico* analysis screens up to 1411 antimicrobial genes, and process time required is significantly less than conventional tests.

6. Conclusion

In summary, we provided the respective D -value for each genotyping method (Table. 10). *Xba*I-PFGE provided fair differentiation; however, using a combination of the restriction enzymes *Avr*II and *Sfi*I allowed for the differentiation of all strains, although the method is less practical. CRISPR functions as a better subtyping scheme when compared to *Xba*I-PFGE, but still needed additional typing schemes to increase the discrimination power. The SNV analysis is the most commonly used, and generates a relatively high discriminatory power ($D = 0.93$) for intra-serotype subtyping. ND contains a k-mer alignment scheme, which gives it an advantage for more detailed screening and provided the best discriminatory power in the present study. MUMi was found unsuitable for *Salmonella* intra-serotype phylogenetic studies. With WGS technology becoming more and more accessible, it is still important to note that alternative bioinformatic analysis may deeply affect the assumed phylogenetic relations, and utilizing the optic approach is critical for epidemiological studies.

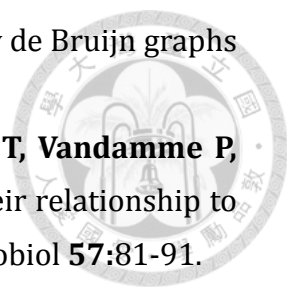
References

1. **Coburn B, Grassl GA, Finlay BB.** 2007. *Salmonella*, the host and disease: a brief review. *Immunol Cell Biol* **85**:112-118.
2. **Vugia DJ, Samuel M, Farley MM, Marcus R, Shiferaw B, Shallow S, Smith K, Angulo FJ.** 2004. Invasive *Salmonella* infections in the United States, FoodNet, 1996-1999: incidence, serotype distribution, and outcome. *Clin Infect Dis* **38 Suppl 3**:S149-156.
3. **Lin CC, Guo JW, Chang CC, Wang YC, Shien JH, Yeh KS, Cheng TH.** 2008. *Salmonella* Serovars Isolated from Marketing Broilers and simulated native chickens: prevalence and drug resistance Taiwan Vet J **34**:217-225..
4. **Chiou CS, Liao YS, Liao CH, Tsao CS, Kuo JC.** 2015. Salmonellosis surveillance and epidemiological trend in Taiwan. *Taiwan Epidemiology Bulletin* **31**.
5. **Kuo HC, Lauderdale TL, Lo DY, Chen CL, Chen PC, Liang SY, Kuo JC, Liao YS, Liao CH, Tsao CS, Chiou CS.** 2014. An association of genotypes and antimicrobial resistance patterns among *Salmonella* isolates from pigs and humans in Taiwan. *PLoS One* **9**:e95772.
6. **Prevention Center for Disease Control.** 2008. Multistate outbreak of human *Salmonella* infections caused by contaminated dry dog food--United States, 2006-2007. *MMWR Morb Mortal Wkly Rep* **57**:521-524.
7. **Harbottle H, White DG, McDermott PF, Walker RD, Zhao S.** 2006. Comparison of multilocus sequence typing, pulsed-field gel electrophoresis, and antimicrobial susceptibility typing for characterization of *Salmonella enterica* serotype Newport isolates. *J Clin Microbiol* **44**:2449-2457.
8. **Allard MW, Luo Y, Strain E, Li C, Keys CE, Son I, Stones R, Musser SM, Brown EW.** 2012. High resolution clustering of *Salmonella enterica* serovar Montevideo strains using a next-generation sequencing approach. *BMC Genomics* **13**:32.
9. **Fricke WF, Mammel MK, McDermott PF, Tartera C, White DG, Leclerc JE, Ravel J, Cebula TA.** 2011. Comparative genomics of 28 *Salmonella enterica* isolates: evidence for CRISPR-mediated adaptive sublineage evolution. *J Bacteriol* **193**:3556-3568.

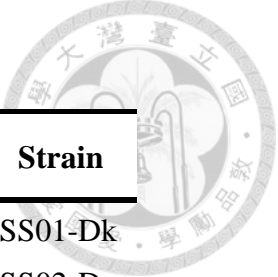
- 
10. **Gori A BA, Marchetti G, Esposti AD, Catozzi L, Nardi GP, Gazzola L, Ferrario G, ven Embden JDA, ven Soolingen D, Moroni M, Franzetti F.** 2005. Spoligotyping and *Mycobacterium tuberculosis*. *Emerg Infect dis* **11**:1242-1248.
 11. **Fabre L, Zhang J, Guigon G, Le Hello S, Guibert V, Accou-Demartin M, de Romans S, Lim C, Roux C, Passet V, Diancourt L, Guibourdenche M, Issenhuth-Jeanjean S, Achtman M, Brisse S, Sola C, Weill FX.** 2012. CRISPR typing and subtyping for improved laboratory surveillance of *Salmonella* infections. *PLoS One* **7**:e36995.
 12. **Shariat N, Kirchner MK, Sandt CH, Trees E, Barrangou R, Dudley EG.** 2013. Subtyping of *Salmonella enterica* serovar Newport outbreak isolates by CRISPR-MVLST and determination of the relationship between CRISPR-MVLST and PFGE results. *J Clin Microbiol* **51**:2328-2336.
 13. **Bachmann NL, Petty NK, Ben Zakour NL, Szubert JM, Savill J, Beatson SA.** 2014. Genome analysis and CRISPR typing of *Salmonella enterica* serovar Virchow. *BMC Genomics* **15**:389.
 14. **Hommais F, Pereira S, Acquaviva C, Escobar-Paramo P, Denamur E.** 2005. Single-nucleotide polymorphism phylotyping of *Escherichia coli*. *Appl Environ Microbiol* **71**:4784-4792.
 15. **Kwong JC, Mercoulia K, Tomita T, Easton M, Li HY, Bulach DM, Stinear TP, Seemann T, Howden BP.** 2016. Prospective whole-genome sequencing enhances national surveillance of *Listeria monocytogenes*. *J Clin Microbiol* **54**:333-342.
 16. **Leekitcharoenphon P, Nielsen EM, Kaas RS, Lund O, Aarestrup FM.** 2014. Evaluation of whole genome sequencing for outbreak detection of *Salmonella enterica*. *PLoS One* **9**:e87991.
 17. **Deloger M, El Karoui M, Petit MA.** 2009. A genomic distance based on MUM indicates discontinuity between most bacterial species and genera. *J Bacteriol* **191**:91-99.
 18. **Dhanoa A, Fatt QK.** 2009. Non-typhoidal *Salmonella* bacteraemia: epidemiology, clinical characteristics and its' association with severe immunosuppression. *Ann Clin Microbiol Antimicrob* **8**:15.
 19. **Asai T, Murakami K, Ozawa M, Koike R, Ishikawa H.** 2009. Relationships between multidrug-resistant *Salmonella enterica* Serovar

- Schwarzengrund and both broiler chickens and retail chicken meats in Japan. *Jpn J Infect Dis* **62**:198-200.
20. **Aarestrup FM, Hendriksen RS, Lockett J, Gay K, Teates K, McDermott PF, White DG, Hasman H, Sorensen G, Bangtrakulnonth A, Pornreongwong S, Pulsrikarn C, Angulo FJ, Gerner-Smidt P.** 2007. International spread of multidrug-resistant *Salmonella* Schwarzengrund in food products. *Emerg Infect Dis* **13**:726-731.
21. **Chou CH and Tsai HJ.** 2001. The prevalence and antimicrobial susceptibility of *Salmonella* and *Campylobacter* in meat-type chickens in Taiwan. *Taiwan Vet J* **27**:27-38.
22. **Lauderdale TL, Aarestrup FM, Chen PC, Lai JF, Wang HY, Shiau YR, Huang IW, Hung CL.** 2006. Multidrug resistance among different serotypes of clinical *Salmonella* isolates in Taiwan. *Diagn Microbiol Infect Dis* **55**:149-155.
23. **Chen MH, Wang SW, Hwang WZ, Tsai SJ, Hsieh YC, Chiou CS, Tsen HY.** 2010. Contamination of *Salmonella* Schwarzengrund cells in chicken meat from traditional marketplaces in Taiwan and comparison of their antibiograms with those of the human isolates. *Poult Sci* **89**:359-365.
24. **Zou W, Lin WJ, Foley SL, Chen CH, Nayak R, Chen JJ.** 2010. Evaluation of pulsed-field gel electrophoresis profiles for identification of *Salmonella* serotypes. *J Clin Microbiol* **48**:3122-3126.
25. **Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV.** 2001. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerg Infect Dis* **7**:382-389.
26. **Zheng J, Keys CE, Zhao S, Ahmed R, Meng J, Brown EW.** 2011. Simultaneous analysis of multiple enzymes increases accuracy of pulsed-field gel electrophoresis in assigning genetic relationships among homogeneous *Salmonella* strains. *J Clin Microbiol* **49**:85-94.
27. **Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG.** 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* **95**:3140-3145.
28. **Kidgell C, Reichard U, Wain J, Linz B, Torpdahl M, Dougan G, Achtman**

- M. 2002. *Salmonella* Typhi, the causative agent of typhoid fever, is approximately 50,000 years old. *Infect Genet Evol* **2**:39-45.
29. **Bolotin A, Quinquis B, Sorokin A, Ehrlich SD.** 2005. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**:2551-2561.
30. **Bhaya D, Davison M, Barrangou R.** 2011. CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annu Rev Genet* **45**:273-297.
31. **Shariat N, Timme RE, Pettengill JB, Barrangou R, Dudley EG.** 2015. Characterization and evolution of *Salmonella* CRISPR-Cas systems. *Microbiology* **161**:374-386.
32. **Fournier PE, Dubourg G, Raoult D.** 2014. Clinical detection and characterization of bacterial pathogens in the genomics era. *Genome Med* **6**:114.
33. **Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, Fitzgerald M, Godfrey P, Haas BJ, Murphy CI, Russ C, Sykes S, Walker BJ, Wortman JR, Young S, Zeng Q, Abouelleil A, Bochicchio J, Chauvin S, Desmet T, Gujja S, McCowan C, Montmayeur A, Steelman S, Frimodt-Moller J, Petersen AM, Struve C, Krogfelt KA, Bingen E, Weill FX, Lander ES, Nusbaum C, Birren BW, Hung DT, Hanage WP.** 2012. Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011. *Proc Natl Acad Sci U S A* **109**:3065-3070.
34. **Hoffmann M, Luo Y, Monday SR, Gonzalez-Escalona N, Ottesen AR, Muruvanda T, Wang C, Kastanis G, Keys C, Janies D, Senturk IF, Catalyurek UV, Wang H, Hammack TS, Wolfgang WJ, Schoonmaker-Bopp D, Chu A, Myers R, Haendiges J, Evans PS, Meng J, Strain EA, Allard MW, Brown EW.** 2016. Tracing origins of the *Salmonella* Bareilly strain causing a food-borne outbreak in the United States. *J Infect Dis* **213**:502-508.
35. **Kircher M, Kelso J.** 2010. High-throughput DNA sequencing--concepts and limitations. *Bioessays* **32**:524-536.
36. **Kaas RS, Leekitcharoenphon P, Aarestrup FM, Lund O.** 2014. Solving the problem of comparing whole bacterial genomes across different sequencing platforms. *PLoS One* **9**:e104984.

- 
37. **Compeau PE, Pevzner PA, Tesler G.** 2011. How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol* **29**:987-991.
38. **Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM.** 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* **57**:81-91.
39. **Hasman H, Saputra D, Sicheritz-Ponten T, Lund O, Svendsen CA, Frimodt-Moller N, Aarestrup FM.** 2014. Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. *J Clin Microbiol* **52**:139-146.
40. **Hunter PR, Gaston MA.** 1988. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. *J Clin Microbiol* **26**:2465-2466.
41. **Grimont PAD and Weill FX.** 2007. Antigenic formulae of *Salmonella* serovars, (9th ed). Paris: WHO Collaborating Center for Reference and Research on *Salmonella*.
42. **Liu F, Barrangou R, Gerner-Smidt P, Ribot EM, Knabel SJ, Dudley EG.** 2011. Novel virulence gene and clustered regularly interspaced short palindromic repeat (CRISPR) multilocus sequence typing scheme for subtyping of the major serovars of *Salmonella enterica* subsp. *enterica*. *Appl Environ Microbiol* **77**:1946-1956.
43. **Grisa I, Vergnaud G, Pourcel C.** 2007. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**:172.
44. **Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL.** 2004. Versatile and open software for comparing large genomes. *Genome Biol* **5**:R12.
45. **Kumar S, Stecher G, Tamura K.** 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* **33**:1870-1874.
46. **Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV.** 2012. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* **67**:2640-2644.
47. **Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Ponten T, Ussery DW, Aarestrup FM, Lund O.** 2012.

- Multilocus sequence typing of total-genome-sequenced bacteria. *J Clin Microbiol* **50**:1355-1361.
48. **Chen MH, Hwang WZ, Wang SW, Shih YC, Tsen HY.** 2011. Pulsed field gel electrophoresis (PFGE) analysis for multidrug resistant *Salmonella enterica* serovar Schwarzengrund isolates collected in six years (2000-2005) from retail chicken meat in Taiwan. *Food Microbiol* **28**:399-405.
49. **Bartual SG, Seifert H, Hippler C, Luzon MA, Wisplinghoff H, Rodriguez-Valera F.** 2005. Development of a multilocus sequence typing scheme for characterization of clinical isolates of *Acinetobacter baumannii*. *J Clin Microbiol* **43**:4382-4390.
50. **Achtman M, Wain J, Weill FX, Nair S, Zhou Z, Sangal V, Krauland MG, Hale JL, Harbottle H, Uesbeck A, Dougan G, Harrison LH, Brisse S, Group SEMS.** 2012. Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. *PLoS Pathog* **8**:e1002776.

Table 1. Isolates of *Salmonella* Schwarzengrund

NO.	Source	Year	Strain
1	Duck	2000	SS01-Dk
2	Stray dog	2003	SS02-Dg
3	Broiler breeder farm	2008	SS03-Ck
4	Pet food	2008	SS04-Pf
5	Broiler breeder farm	2009	SS05-Ck
6	Broiler	2010	SS06-Ck
7	Wild bird (Crested Goshawk)	2011	SS07-Cg
8	Wild bird (Moorhen)	2011	SS08-Mh
9	Turkey	2012	SS09-Tk
10	Pig Farm	2012	SS10-Pg
11	Pig Farm	2012	SS11-Pg
12	Turkey farm	2012	SS12-Tk
13	Turkey farm	2012	SS13-Tk
14	Duck	2012	SS14-Dk
15	Broiler breeder farm	2008	SS15-Ck
16	Broiler breeder farm	2008	SS16-Ck

Table 2. MLST results that were performed prior this study. Allele type and sequence type of all 16 strains of *S. Schwarzengrund*

Gene	<i>thrA</i>	<i>purE</i>	<i>hisD</i>	<i>aroC</i>	<i>hemD</i>	<i>sucA</i>	<i>dnaN</i>
Allele type	3	41	49	43	43	15	47
Sequence type	96						

Table 3. Primers used for CRISPR allele amplification and sequencing

Primer	Sequence (5' to 3')
CRISPR1-F	GATGTAGTGCGGATAATGCA
CRISPR1-R	GATGATATGGCAACAGGTTT
CRISPR2-F	ACCAGCCATTACTGGTACAC
CRISPR2-R	ATTGTTGCGATTATGTTGGT

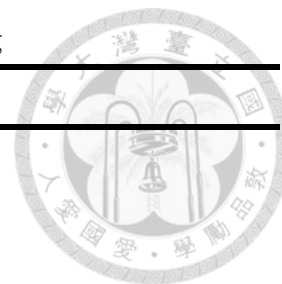


Table 4. CRISPR1 and CRISPR2 allele spacer distributions among the 16 *S. Schwarzengrund* strains tested.

Strains	SS01-Dk	SS02-Dg	SS03-Ck	SS04-Pf	SS05-Ck	SS06-Ck	SS07-Cg	SS08-Mh	SS09-Tk	SS10-Pg	SS11-Pg	SS12-Tk	SS13-Tk	SS14-Dk	SS15-Ck	SS16-Ck
CRISPR1	1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	2	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	3	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	4	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	5	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	6	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	7	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	8	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	9	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	10	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	11	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	12		■	■	■	■	■	■	■	■	■	■	■	■	■	■
	13		■	■	■	■	■	■	■	■	■	■	■	■	■	■
	14												■	■	■	■
CRISPR2	1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	2	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	3	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	4	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	5	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	6	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	7	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	8	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	9	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	10	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	11	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	12	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	13	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	14	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	15	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	16							■								
	17							■								
CT	A1	B1	B1	B1	B1	B1	C2	C1	B3	B3	B3	D1	D1	E3	E1	E1



Table 5. Number of spacer difference between CTs

A1	0							
B1	2	0						
B3	3	1	0					
C1	4	2	3	0				
C2	8	6	7	4	0			
D1	4	4	5	6	10	0		
E1	2	2	2	4	8	2	0	
E3	3	3	3	5	9	3	1	0
CT types	A1	B1	B3	C1	C2	D1	E1	E3



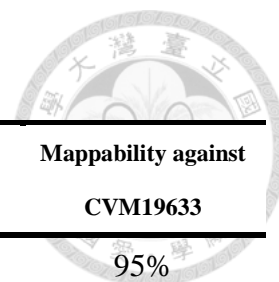


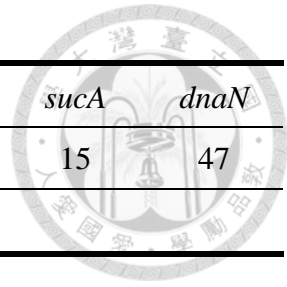
Table 6. WGS Read profile of 16 strains

Strain	Total reads after QT*	Total base after QT	Average depth	Mappability against CVM19633
SS01-Dk	11036,194	994,927,166	207	95%
SS02-Dg	10,198,540	913,056,876	190	94%
SS03-Ck	5,080,908	431,045,461	89	95%
SS04-Pf	12,439,150	1,124,485,899	234	99%
SS05-Ck	7,601,882	694,841,031	144	94%
SS06-Ck	7,829,392	702,478,211	146	90%
SS07-Cg	13,204,098	1,196,303,851	249	94%
SS08-Mh	10,688,548	965,227,959	201	92%
SS09-Tk	9,537,362	854,887,312	178	92%
SS10-Pg	6,562,338	593,724,599	123	92%
SS11-Pg	12,395,336	1,138,982,249	237	93%
SS12-Tk	11,928,540	1,065,982,249	222	93%
SS13-Tk	6,494,510	593,079,712	123	96%
SS14-Dk	13,768,596	1260,763,095	262	94%
SS15-Ck	7,005,766	620,739,212	129	98%
SS16-Ck	11,210,632	1,014,713,610	211	98%

*QT (Quality trimming)

Table 7. *In silico* MLST typing of 16 *S. Schwarzengrund*

Gene	<i>thrA</i>	<i>purE</i>	<i>hisD</i>	<i>aroC</i>	<i>hemD</i>	<i>sucA</i>	<i>dnaN</i>
Allele type	3	41	49	43	43	15	47
Sequence type	96						



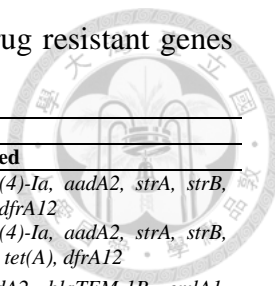


Table 8. Results of drug resistance testing and the *in silico* detection for drug resistant genes for the 16 *S. Schwarzengrund* strains.

Strains	Disk diffusion test*	<i>In silico</i>
		Genes existed
SS01-Dk	AMP, <u>C</u> **, CN, K, <u>NA</u> , OT, TE, S, SXT	<i>aadA1, aph(3')-Ia, aac(3)-Iva, aph(4)-Ia, aadA2, strA, strB, blaTEM-1B, sul1, sul2, sul3, tet(A), dfrA12</i>
SS02-Dg	AMP, C, CN, K, <u>NA</u> , OT, TE, S, SXT	<i>aadA1, aph(3')-Ia, aac(3)-Iva, aph(4)-Ia, aadA2, strA, strB, blaTEM-1B, cmlA1, sul1, sul2, sul3, tet(A), dfrA12</i>
SS03-Ck	AMP, <u>C</u> , CN, <u>NA</u>	<i>aph(4)-Ia, aac(3)-Iva, aadA1, aadA2, blaTEM-1B, cmlA1, sul1, sul3, tet(A), dfrA12</i>
SS04-Pf	C, <u>KF</u> , <u>NA</u> , S	<i>aadA1, aadA2, strA, strB, sul2, sul3, cmlA1</i>
SS05-Ck	AMP, C, CN, <u>NA</u> , OT, TE, S, SRT	<i>aph(4)-Ia, aadA1, aac(3)-Iva, aadA2, blaTEM-1B, cmlA1, tet(A), dfrA12</i>
SS06-Ck	AMP, C, CN, <u>NA</u> , OT, TE, S, SXT	<i>aac(3)-Iva, aph(4)-Ia, aadA2, aadA1, blaTEM-1B, cmlA1, floR, sul2, tet(A), dfrA12</i>
SS07-Cg	AMP, C, CN, K, KF, <u>NA</u> , OT, TE, S, SXT	<i>aadA1, aph(3')-Ia, aac(3)-Iva, aadA2, strA, strB, blaTEM-1B, cmlA1, Sul1, sul2, sul3, tet(A), dfrA12</i>
SS08-Mh	AMP, C, CN, K, KF, <u>NA</u> , OT, TE, S, SXT	<i>aadA1, aph(3')-Ia, aac(3)-Iva, aadA2, strA, strB, blaTEM-1B, cmlA1, sul1, sul2, sul3, tet(A), dfrA12</i>
SS09-Tk	C, <u>CIP</u> , CN, <u>E</u> , KF, <u>NA</u> , OT, TE, S, SXT	<i>aph(4)-Ia, aac(3)-Iva, aadA1, aadA2, cmlA1, floR, sul1, sul3, tet(A), dfrA12</i>
SS10-Pg	<u>AMP</u> , C, <u>CIP</u> , CN, <u>E</u> , KF, <u>NA</u> , S, SXT	<i>aph(4)-Ia, aac(3)-Iva, aadA1, aadA2, cmlA1, floR, cmlA1, floR, dfrA12</i>
SS11-Pg	<u>AMP</u> , C, <u>CIP</u> , CN, <u>E</u> , KF, <u>NA</u> , S, SXT	<i>aadA1, aph(4)-Ia, aac(3)-Iva, aadA2, cmlA1, floR, sul1, sul3, dfrA12</i>
SS12-Tk	C, <u>NA</u> , OT, TE, S, SXT	<i>aadA1, aadA2, cmlA1, sul1, tet(A), dfrA12</i>
SS13-Tk	AMP, C, CN, <u>NA</u> , OT, TE, S, SXT	<i>aadA1, aac(3)-Iva, aph(4)-Ia, aadA2, strA, strB, blaTEM-1B, cmlA1, floR, sul1, sul2, sul3, dfrA12</i>
SS14-Dk	C, <u>CIP</u> , CN, <u>E</u> , <u>NA</u> , OT, TE, S, SXT	<i>aadA1, aph(4)-Ia, aac(3)-Iva, aadA2, cmlA1, floR, sul1, sul3, tet(A), dfrA12</i>
SS15-Ck	C, <u>NA</u> , S	<i>aadA1, aph(4)-Ia, aac(3)-Iva, aadA2, cmlA1, sul2, sul3</i>
SS16-Ck	C, <u>NA</u> , S	<i>aadA1, aph(4)-Ia, aac(3)-Iva, aadA2, cmlA1, sul2, sul3</i>

*AMP, Ampicillin; C, Chloramphenicol; CIP, Ciprofloxacin; CN, Gentamicin; E, Enrofloxacin; K, Kanamycin; KF, Cephalothin; NA, Nalidixic acid; OT, Oxytetracycline; TE, Tetracycline; S, Streptomycin; SXT, Sulfamethaxole/Trimethoprim

**Drugs that are underlined were not detected by the *in silico* analysis.

Table 9. Corresponding genes to identified SNVs

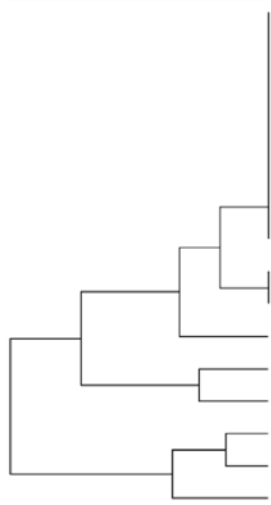
	<i>No.</i>	Corresponding genes and significance of change	Strains
<i>SNVs in genes</i>	2	<i>ccmF</i> , synonymous coding	SS06, SS15
		<i>rrf</i> , downstream of ORF	SS01
<i>SNVs not in genes</i>	120	-	

Table 10. The comparison of *D*-value among all subtyping methods

Genotyping Methods	PFGE	MLST	CRISPR	WGS		
				MUMi	ND	SNV
<i>D</i>-value	0.79	0	0.87	0.79	0.94	0.90

Genetic similarity (%)

75 80 85 90 95 100



Strains

SS01-Dk (Duck, 2000)
SS04-Pf (Pet food, 2008)
SS02-Dg (Dog, 2003)
SS15-Ck (Broiler breeder, 2008)
SS16-Ck (Broiler breeder, 2008)
SS03-Ck (Broiler breeder, 2008)
SS05-Ck (Broiler breeder, 2009)
SS06-Ck (Broiler, 2010)
SS07-Cg (Crested Goshawk, 2011)
SS08-Mh (Moorhen, 2011)
SS14-Dk (Duck, 2012)
SS12-Tk (Turkey, 2012)
SS13-Tk (Turkey, 2012)
SS11-Pg (Pig, 2012)
SS09-Tk (Turkey, 2012)
SS10-Pg (Pig, 2012)



Fig 1. Strain isolation information, *Xba*I-PFGE patterns, phylogenetic tree of the 16 *S. Schwarzengrund* strains tested. The phylogenetic tree was obtained using the Dice coefficient of similarity and the UPGMA method. Band tolerance was set to 1.5% and optimization at 1%. The degree of genetic similarity is shown by percentage according to the scale.

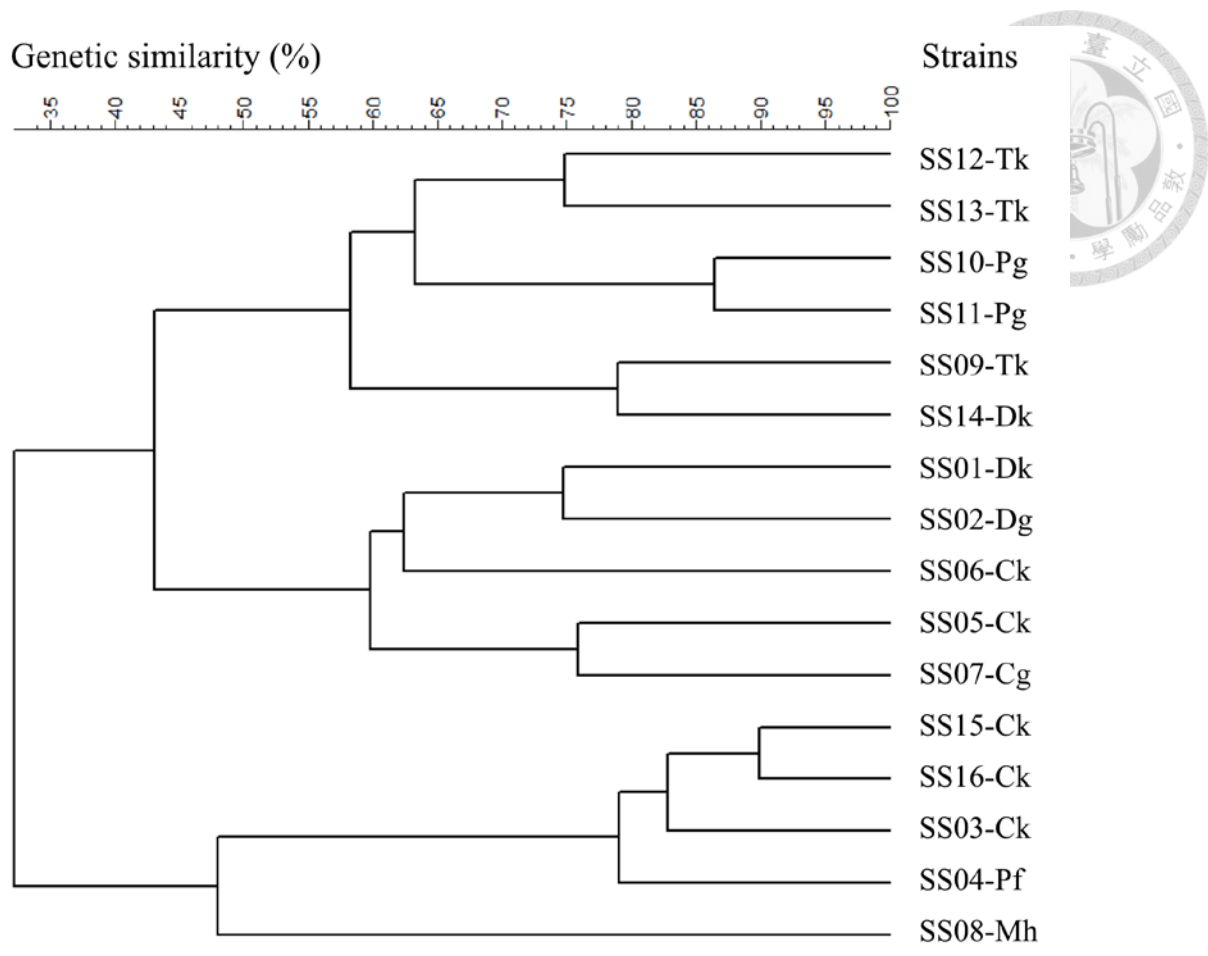


Fig 2. Combined dendrogram of 16 *S. Schwarzengrund* strains digested by *AvrII* and *SfiI*. The phylogenetic tree was obtained using the Dice coefficient of similarity and the UPGMA method. Band tolerance was set to 1.5% and optimization at 1%. The degree of genetic similarity is shown by percentage according to the scale.

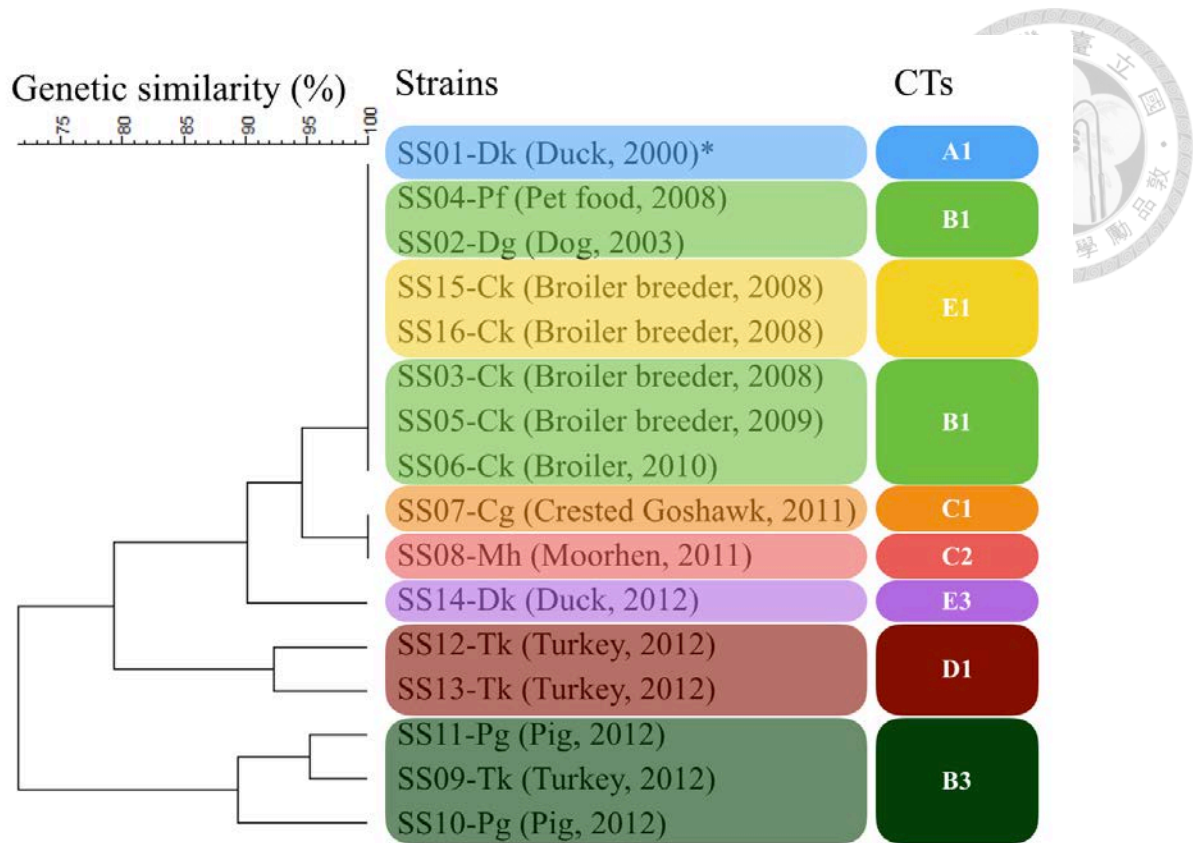


Fig 3. Distribution of CTs of 16 strains among *Xba*I-PFGE dendrogram.



Fig 4. SNV tree of 16 *S. Schwarzengrund* strains with *S. Typhimurium* strain LT2 as reference. The evolutionary distance units represent the number of base substitutions per site.

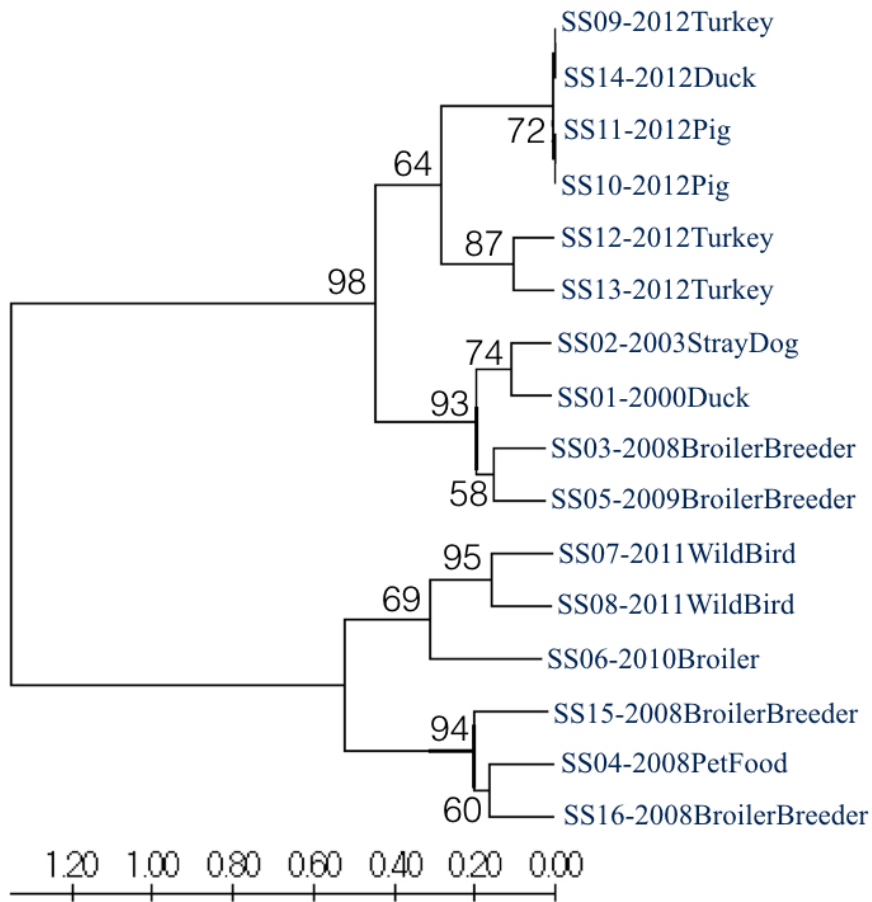


Fig 5. SNV tree of 16 *S. Schwarzengrund* strains with *S. Schwarzengrund* strain CVM19633 as reference. The evolutionary distance units represent the number of base substitutions per site.

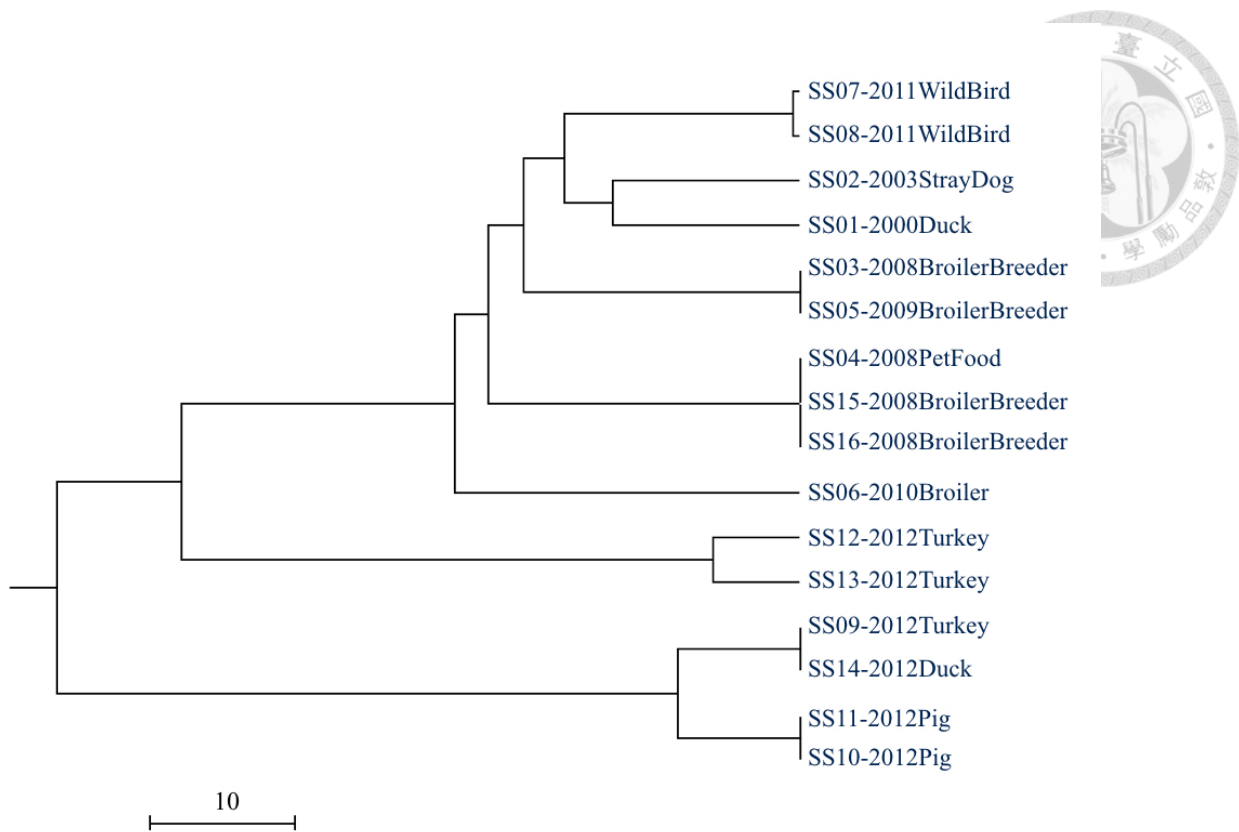


Fig 6. Nucleotide difference tree of 16 *S. Schwarzengrund* strains. The evolutionary distance units represent the number of differences between nucleotides.

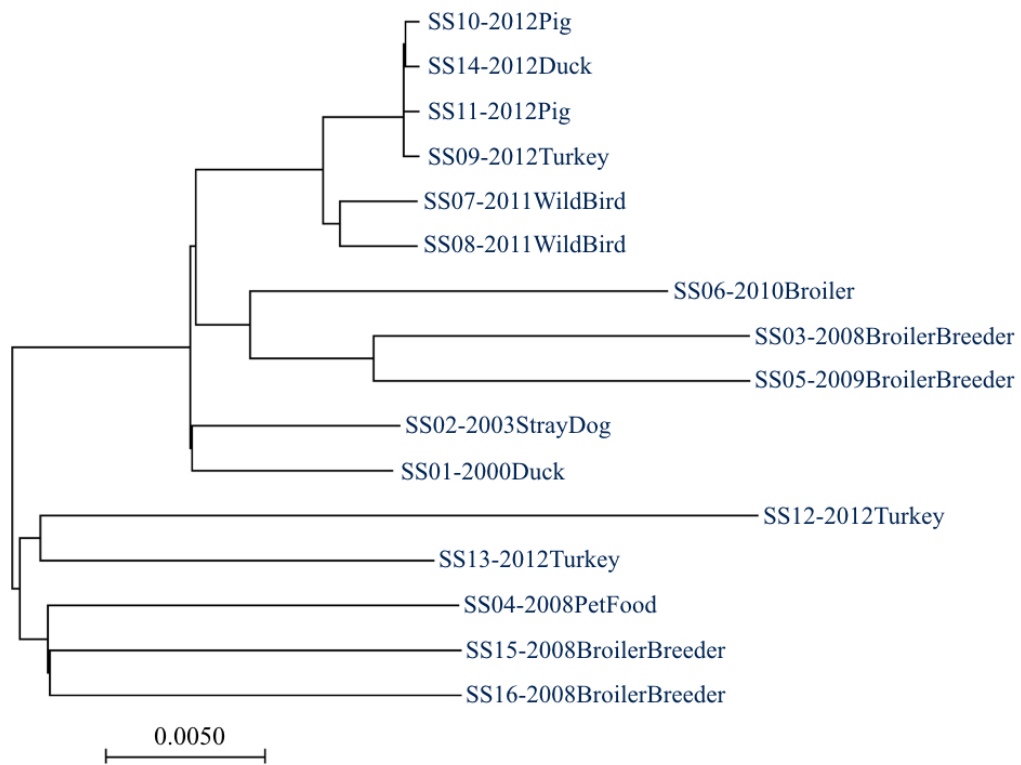


Fig 7. MUMi tree of 16 *S. Schwarzengrund* strains. The evolutionary distance units represent the percentage of total genomic differences.