RESEARCH REPORT
PORK SAFETY

Title: Sequencing of 10 potential molecular epidemiological targets from a collection of food safety relevant *Salmonella* spp. **- NPB #05-063**

Investigator: Dr. Scot E. Dowd

Institution: USDA ARS Livestock Issues Research Unit

Date Received: July 12, 2006

Abstract:

Salmonella research is a top priority for the industry, especially related to pre-harvest reduction of pathogens with potential public health significance. The pre-harvest food safety research for *Salmonella* described here fell under the category of control strategy and responded to the directive from the NPB "Development and evaluation of evolving molecular, and other, diagnostic tools and monitoring techniques that can be used in epidemiological investigations". The development and testing of molecular epidemiological tools for pre-harvest food safety applications continues to be an important and rapidly evolving issue. It is becoming vital to track contamination to its source during production, transport, and lairage. The ability to reliably distinguish between various isolates of the same serovar of *Salmonella* spp. will provide a valuable tool to the pork industry that would aid in tracking and identifying original sources of contamination. However, the development of reliable epidemiological methods for source tracking of *Salmonella* spp. represents a significant and real problem. One species of *Salmonella* (*Salmonella enterica*) has literally hundreds of very closely related subspecies or serotypes that must be distinguishable by any molecular method used. Many of the current attempts at the development of molecular epidemiological tools for *Salmonella* spp. are based upon methods that are applicable to other microorganisms. Because of the close genetic similarity of *Salmonella* spp. serotypes and isolates, most of these methods ultimately fail to provide reproducible results when used for *Salmonella* spp.

The current project will make available to food safety scientists a powerful comparative database that can be used to develop epidemiological methods that are specific to *Salmonella* spp. rather than attempting to adapt methods that have been used with other organisms. *Salmonella* spp. because of how closely related the serovars tend to be presents a significant challenge to the development and applicability of molecular epidemiological methods. Using the genome of *Salmonella* Typhimurium LT2 we have successfully aligned the entire genomes of 10 *Salmonella* spp. strains/serotypes. This data has been made available on the internet http://liru.ars.usda.gov/salmonella2. This research represented the first whole genome multiple alignment system available over the internet as well as one of the first base by base comparison of multiple bacterial genomes. During this National Pork Board funded research project we identified many regions of these genomes that show potential to serve as molecular epidemiology targets. It is this type of target region that

taken individually or in groups tend to have the strongest potential for use as powerful and sensitive molecular epidemiology source tracking tools. Based upon those comparative genome analyses we have identified 10 regions that putatively could allow for molecular typing but also have potential to be useful for very isolate specific source tracking. In order to evaluate their usefulness as epidemiology markers we sequenced these 10 hypervariable regions (molecular epidemiological target regions METR) of the Salmonella genome from a collection of 50 different *Salmonella* serovars. Following the collection of the genetic information we established prototype databases which allowed for comparative analysis and query of unknown sequences. Finally, we obtained 6 "test isolates" of known serotype and sequenced the METR of these to determine if we could correctly identify their serotypes using the prototype databases. Although a statistical model and additional database tools must be developed to allow for automated analyses of results, in each case the 6 test serotypes were correctly identified by comparison to the METR databases. These results indicated the potential for the successful establishment of a molecular typing method which can be used for *Salmonella* spp. In its current form this typing method may not have the ability to distinguish between individual strains of a particular serotype, but the results indicate that further elucidation and extension of the METR regions and expansion of the collection of serotypes and strains contained within the database do have the potential for highly sensitive differentiation of serotype strains. This pilot project has established a METR prototype sequence databases for putative molecular typing of *Salmonella* spp at the serovar level.

Introduction:

The development and testing of molecular epidemiological tools for pre-harvest food safety applications is an important and rapidly evolving issue. It is vital to track contamination to its source during production, transport, or lairage. The ability to reliably distinguish between various isolates of the same serovar of *Salmonella* spp. will provide a valuable tool to the pork industry that would aid in source tracking contamination. As an example Salmonella contamination could be tracked back to a failed water supply on a farm. Epidemiological methods for *Salmonella* spp. present a significant and real problem. One species of *Salmonella* (*Salmonella enterica*) has literally hundreds of very closely related subspecies or serotypes that must be distinguishable by any molecular method used. The level of relatedness of these serotypes has made the functional value of traditional molecular epidemiological methods less than what they are with other microorganisms such as *Listeria* spp. for example. For this reason there is an abundance of research going on directed at applying methods that are used or developed for other genera of microorganism, to the molecular epidemiology of *Salmonella* spp. The failure of these methods is evident by lack of significant publication of methods and the call for the development and testing of reliable epidemiological methods for this molecularly complex food borne pathogen.

Molecular Epidemiological tools are based on measurement of phenotypic or genetic heterogeneity and relationships between strains can be established, because microbial mutation rates determine whether sufficient discrimination can be attained. Certain pathogenic strains can be responsible for disease outbreaks or increases in infection frequency. The possibility to identify hazardous isolates and their epidemiological properties and track these isolates to their source on the farm or in the environment is of clear importance to the pork industry as part of on-farm management and the development and identification of critical control points.

Microbial typing methods fall into two groups: techniques that determine phenotypic characteristics for example antibiotic resistance profiles and carbon source utilization profiles and approaches that focus on genetic determinants such as Restriction Fragment Length Polymorphism (RFLP) analyses.

An adequate typing system distinguishes identity versus non-identity, but the worth of a particular method is determined by several characteristics. The typing ability of a particular assay defines the proportion of strains that can be successfully analyzed with the typing technique that is applied. The optimal value would be equal to 100% of course. The reproducibility is determined by the ability of a typing system to assign the same type to a strain tested on independent separate occasions. Again, the optimal value would be 100%. A typing system is also influenced by the genetic stability of tester isolates: a type strain should show stable typing characteristics even after repeated laboratory processing. The discriminatory power of a typing system determines whether

epidemiologically non-related isolates within a single species or serotype (in the case of *Salmonella* spp.) will be recognized as phenotypically or genetically distinct. Within the farm setting, where local endemicity screening or outbreak tracking provides most of the microbial isolates to be analyzed, the costs, speed and throughput of a procedure are also important.

With all of these requirements for a molecular method it is not surprising that previous methods have fallen short or that it takes a combination of several methods to provide the necessary sensitivity and specificity required for *Salmonella* spp. In order to match laboratory data to pre-harvest phenomena, strict definition of pathogens characteristics and relatedness is required. Relatedness among bacterial isolates can only be determined in the context of extensive knowledge of the value of the typing data. Currently there is not yet a single typing strategy (with the theoretical exception of whole genome nucleic acid sequencing) that in 100% of cases enables adequate typing. At best comparative typing can be performed. For instance nucleic acid analyses can be used successfully to define genetic variability. Constant elements can be used taxonomically to define genus and species characteristics. The nucleic acid variability occurring among isolates and strains may be used for detailed tracking of microorganisms belonging to a given species.

New typing procedures are appearing month by month. These vary from procedures that identify mutations in specific genes to alternative approaches that assess changes occurring throughout an entire genome. The recent degenerate oligonucleotide primed PCR combines exquisite sensitivity to broad spectrum identification of polymorphic DNA regions, allowing characterization of pictogram quantities of DNA. Alternatively, the presence or absence of restriction site in well defined pieces of amplified DNA can be used to distinguish among strains. This procedure termed Cleaved Amplified Polymorphic Sequences, performed well, was thought to be highly reproducible rapid and utilized widely available technology. Another procedure uses RAPD screening to identify and develop into isolate specific DNA probes those genetic elements that can be either absent from or present in the genome of a strain of a species. Upon combination of various probes, multidigits +/- characteristics were established for different strains. The procedure performed well in comparison with others suited for typing. Amplification Fragment Length Polymorphism analysis has already surpassed the stages of initial development. AFLP shows good resolution and reproducibility at various taxonomic levels. The fact that no DNA sequence information is required prior to performing the experiments is another useful asset.

Unfortunately, the development of molecular epidemiological methods including current attempts to develop reliable methods for *Salmonella* spp. follows a trial and error approach. A researcher may identify a logical approach and perform testing on a collection of isolates. Such trials may show initial success but their failure to provide needed sensitivity or selectivity may only be identified years later after they are widely practiced. In example, with Restriction Length Polymorphic Analysis of *Escherichia coli* O157, isolates can change their banding patterns after several serial passages in culture. This indicates that this method is not stable and is dependent upon the organism's genetic elasticity. With knowledge of the complete genetic sequence of two *Salmonella* spp. genomes (*typhi* and Typhimurium) and several partially completed *Salmonella* spp. genomes (eg. Typhimurium, Enterica, *paratyphi*, and *enteritidis*) it becomes possible to perform whole-genome comparison of these important *Salmonella* isolates, especially in relation to identification of specific markers that may be used as or as part of improved molecular epidemiological or detection targets. This whole genome comparison is the primary and most important step in the development of reliable molecular based epidemiological tool, yet until this time was impossible. Without in-depth knowledge and bioinformatics based comparison of the genomes of divergent isolates of *Salmonella* spp. and the use of this knowledge as a basis for the design of molecular epidemiological tools, current approaches (such as RFLP) are only shotgun attempts to bring about an epidemiological solution.

Currently entire genomes of at least 10 separate *Salmonella* spp. have been or are currently being sequenced. This facilitates and provides the information needed in the development of novel typing strategies based on, for instance, the presence or insertion elements, temperate phages and transposons. Essential ubiquitous genes can be identified and highly variable regions opening different evolutionary windows can be mapped. For instance

searches for intrinsically unstable stretches of tandemly repeated DNA can be undertaken. It has been demonstrated in various studies that repetition of a given short DNA sequence motif in tandem arrays generates a region that is liable to variation in the number of unit sequences present. As a result of the mechanism of slipped strand mispairing copies of the repeat unit can be formed or deleted during the process of DNA replication. Regions comprising repetitive motifs built from units that are relatively short are also known as variable number of tandem repeat loci. When PCR primers bordering the repeat region are designed amplification of the repeat region will reveal differences in repeat number between strains. This generates a figure by which strains can be discriminative in a digital fashion. This procedure allows high speed screening of genetic polymorphisms. Computerized screening of multiple full genome sequences will result in high resolution typing schemes.

As part of our NPB funded efforts we have performed whole genome comparison on a basic level to document the prevalence and relative location of every gene in each of the sequenced genomes of *Salmonella* spp. The identification of specific genetic markers such as surface proteins for instance, that show moderate polymorphism in centralized locations among these 10 isolates have the potential for development into methods and subsequently tested on collections of *Salmonella* spp. field isolates. It is vital to understand, that in-depth knowledge of the genome structure of *Salmonella* spp. is essential in all aspects of molecular microbiology and this knowledge is not limited to detection and molecular epidemiological method development. Thus, the data generated in this study will have broad and significant impact.

This project provides vital data needed in the development of new and reliable molecular epidemiological tools for *Salmonella* spp. Web based databases have previously been constructed that can be used by researchers to identify key markers among the 6 genomes represented that have potential for use as molecular epidemiological targets and tools. Rather than using non-specific methods designed for less complex species of microorganism in an attempt to find a molecular solution this project provides a solid foundation of bioinformatics data that can be used to logically design improved molecular methods specific for *Salmonella* spp. As part of this project we have sequenced 10 METR regions of 50 serovars of *Salmonella* spp, developed a set of databases, sequenced the 10 regions from an additional 6 test isolates to use as a challenge of the proposed method, and performed and initial test designed to evaluate if the use of the METR databases can effectively distinguish the serotypes of the 6 test isolates.

Objectives:

Specific objectives

1. To sequence ten (10) identified heterogeneous target regions (at least 3x coverage) from 40 *Salmonella enterica* serotypes
2. Develop specific PCR primers that will amplify these target regions.
3. Test the targets for their ability to sensitively and specifically identify or distinguish isolates of *S. enterica* from our culture collection.

Materials & Methods:

This project has sequenced 10 specific regions of 50 isolates of *Salmonella* spp. (primarily serotypes or species *enterica*) most of which are important from a food safety perspective to create the start of a molecular epidemiology database.

PCR primer design

Regions from the 10 genomes that show potential as targets for molecular epidemiological tools (METR) were identified using the NPB funded whole genome comparison tool at http://liru.ars.usda.gov/salmonella2. Using the original multiple alignment files we will extract consensus sequences and design primers that will amplify universally on the known serotypes and amplify across the hypervariable regions of the targets. Several software packages were utilized to perform this work including PrimerSelect from DNAstar, Multialign from DNAstar and EditSeq from DNAstar. Primers that did not have ambiguous bases and have high selection scores were chosen and synthesized using the IDTdna (idtDNA, CA) oligonucleotide synthesis service.

PCR

PCR was performed using standard methods. PCR of regions less than 1kb was performed using Qiagen's ProofStart DNA Polymerase (Qiagen Inc., CA) which provides high fidelity and error correction. PCR products were purified using Qiagen PCR purification kits.

Sequencing

PCR product sequencing was performed using standard methodologies on an Applied Biosystems API prism 9800 sequencer using BIG-DYE termination v3.1 technology and manufacturers protocols. The forward and reverse primers for the PCR reaction were used as appropriate to obtain at least 3x coverage of the METR. Assemblies of the sequenced regions were compiled with SEQMAN II from DNAstar.

Public database entry of sequencing products

Sequences from each METR were entered into a local database and then each individual METR (containing sequence information from each of the 50 serotypes) was formatted into BLAST databases.

Testing of targets for specificity and sensitivity

Once the sequence databases were developed, we challenged the methodology using the designed PCR primers combined with sequencing and Computer database homology comparison of 6 test isolates. These test isolates were known serotypes of Salmonella enterica that were not sequenced initially as part of the METR project. Using the sequence data from these test isolates we individually BLASTed each of the METR regions of the test isolates against the full BLAST databases. Individual results are obtained for each METR region. The top 10 hits are collected and from these the serotype with the highest ranks compiled from each of the databases was determined. The serotype with the highest cumulative rank scores are then compared to the expected serotype of the test isolate to evaluate the use of PCR_CDHC (polymerase chain reaction and computer database homology comparison) as a method for typing serovars of *Salmonella enterica*.

Results:

1. To sequence ten (10) identified heterogeneous target regions (at least 3x coverage) from 40 *Salmonella enterica* serotypes Typhimurium and Enteritidis.
2. Develop specific PCR primers that will amplify these target regions.
3. Test the targets for their ability to sensitively and specifically identify or distinguish isolates of *S. enterica* from our culture collection.

We have finished the stated objectives of this research project. We are also exceeding the stated goals by obtaining sequencing METR data from 50 isolates rather than 40 isolates as noted in objective 1. This does not include the additional 6 test isolates we also obtained sequence data from during the testing of this methodology. We have successfully developed a sequence archive as well as a set of METR BLAST databases to maintain these sequences and use as a methodology test bed. As required for Objective 2, we have also developed a set of 10 primer pairs that can be used as part of PCR-CDHC (polymerase chain reaction and computer database homology comparison) methods when screening unknown isolates in the future. This database will act as a dedicated resource for housing this genetic data as well as an integrated analysis tool for future epidemiological analyses. As per Objective 3, we have obtained 6 test isolates, known isolates from 3 serovars that can be used in simulation experiments. These isolates have been screened with the 10 primer pairs and the resultant amplicons sequenced (1x coverage only). These test isolates are being used to assist in ongoing developing a model for serotype identification utilizing the target populations. Following this follow-up project, which was **not** one of the original objectives but instead a logical follow-up to the completed study proposed, we will submit all of the sequences to a public database (www.ncbi.nlm.nih.gov) and publish our results and findings providing access to sequences, primers, and identification models.

Discussion:
Salmonella spp., because of how closely related the serovars tend to be and because so many serotypes exist, presents a significant challenge to the development and applicability of molecular epidemiological methods. This project was performed as a primary step to provide advances in the logical development of a molecular method would allow for specific methods that are able to distinguish between different isolates within the same subserovar. To develop an easy to use molecular source tracking method and make it available to the industry would be a great benefit. The product of this project was anticipated to be a powerful comparative sequence and genome comparison database which could be used to refine epidemiological methods that are specific to *Salmonella* spp..

The specific goals of this project were to identify 10 potential genetic target regions that are both common in *Salmonella* enterica serovars but also hypervariable from a sequence standpoint. In our previously funded Pork Board project we proposed whole genome comparisons of 6 *Salmonella* spp. to provide a foundation for the development of epidemiological markers. We exceeded these previously funded goals by comparing the genomes of 10 *Salmonella* spp rather than the proposed 6 genomes. We also developed an online data storage and visualization system that was published at http://liru.ars.usda.gov/salmonella2/salmonella.html . This background research provided the data that allowed for the identification of 10 regions that are both common (all genomes have a similar genetic region) and hypervariable (the sequence is slightly difference among most or all of the serovars evaluated). This means that these molecular epidemiological target regions (METR) within the genome of *Salmonella enterica* serovars are encoded in the genomes of all isolates but also that these METR have evolutionary differences in the actual genetic sequences.

The ultimate goal for this project was to attempt the development of a molecular method for serotyping and source tracking of Salmonella spp. using both the commonalities and the differences of newly identified genetic targets. We have utilized data from the previously funded Pork Board project to identify 10 potential genetic targets as part of the current project. We have finished sequencing of these 10 targets within 50 known isolates of *S. enterica* that are of various serotypes, evaluated their potential to provide reliable source tracking, and prepared a database to validation. We have exceeded the stated objectives by obtaining sequencing data for 50 isolates rather than the proposed 40 isolates. A web enabled database noted above will also provide access to primer sets developed as part of this project for future epidemiological or sero/genotyping studies. Preliminary evaluation of the databases using the test isolates that were sequenced 1x, indicates that, through the use of these 10 markers, there is the ability to discriminate the serovar of *Salmonella* spp. isolates. Further testing with additional isolates and test serovars is required to fully evaluate the potential of this methodology. The development of a comparison model for interpretation of the results is also necessary before this genotyping method can be validated. Finally, the inclusion of these targets into national archive databases (Genbank) will allow future improvements in the methodologies under development, which we will publish in a peer reviewed journal in the coming months. All manuscripts prepared from this data will acknowledge NPB project 05-063 and will be forwarded to the pork board following publication.

Lay Interpretation:
The development and testing of molecular epidemiological tools for pre-harvest food safety applications is an important and rapidly evolving issue. This means that it is vital to track bacterial contamination of our food to its source, whether that contamination originated during production, transport, holding/lairage or even within processing plants, delivery, stores, or at home. As an example *Salmonella* contamination found in meat products could be tracked back to a failed water supply on a farm if a successful molecular tracking method could be developed. By identifying sources we can prevent further contamination of our food originating from that source. As part of this project our goal was to take steps toward the development of improved molecular epidemiology tools using new whole genome informatics. Specifically our goals were: 1) To sequence ten (10) identified heterogeneous target regions (at least 3x coverage) from 40 *Salmonella enterica* serotypes 2) Develop specific PCR primers that will amplify these target regions in all serotypes.3) Test the targets for their ability to sensitively

and specifically identify or distinguish isolates of *S. enterica* from our culture collection. We successfully completed all of these objectives and exceeded the goals by including an additional 10 serotypes in the sequencing objective. We have developed a set of primers that are able to amplify target regions in all of the serotypes we have tested. We have also sequenced a set of test isolates that have allowed evaluation of a computer database genetic homology comparison method using the data generated in this study. Our results show that we can identify the *S. enterica* serotypes test isolates correctly, which is a valuable tool in itself. With the increase cost of serotyping a genotyping method would be extremely valuable to the industry. We are now using a neural network approach to develop a model that will allow evaluation of complicated set of data that is generated during the homology comparison step in this process. Thus, we have developed a method that shows great potential as a genotyping method for *S. enterica* serovars and has potential as a method for source tracking if it becomes possible to continue to expand and improve the genetic databases and models we have initiated. For more information contact Dr. Scot E. Dowd Ph.D. USDA-ARS-LIRU 1604 E. FM 1294 Lubbock, TX 79403. 806-746-5356. sdowd@lbk.ars.usda.gov