RESEARCH REPORT
PORK SAFETY

**Title:** Bioinformatics Based Genome Comparison of Six Salmonella spp. to Provide a Foundation for the Development of Reliable Molecular Epidemiological Methods.
**NPB #03-132**

**Investigator:** Dr. Scot E. Dowd

**Institution:** USDA-ARS Livestock Issues Research Unit, Lubbock, TX

II. Abstract:

*Salmonella* research is the 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* proposed here falls under the category of control strategy. The research proposed here is highly relevant and applied in nature. Without the groundwork, foundation and benefit of comprehensive genetic and bioinformatic analyses many shotgun approaches to molecular epidemiology based upon methods that are applicable to other microorganisms for instance will ultimately fail when used for *Salmonella* spp. Using the genome of Salmonella typhimurium LT2 as a backbone we have successfully aligned the entire genomes of between 6 and 10 additional Salmonella spp. strains/serotypes. This data has been made available on the internet http://199.133.147.108/salmonella/salmonella.html by clicking on the Salmonella multiple alignment viewers. Additional genetic information is still be generated for several of these strains of Salmonella and we will continue to curate this information and perform whole genome comparisons, generate publicly available analysis tools, and provide this information over the internet to the research community. In addition during our own analysis we have identified many regions of these genomes that show potential as molecular epidemiology targets. It is this type of region that taken individually or in groups tend to have the strongest potential for use as powerful and sensitive molecular epidemiology tools. Based upon our current analysis we may have identified regions that would allow for molecular serotyping at the very least which would be useful for source tracking. This project makes 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 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. This project is the primary and quite simply the only logical foundation step that will ever result in the development of a method that fulfills all the requirements of a superior molecular epidemiological method for *Salmonella* spp.

III. 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. Epidemiological methods for *Salmonella* spp. present and 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 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 that is attempting to apply methods that are used or developed for other genera of microorganism to the problem of *Salmonella* spp. molecular epidemiology. 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 measure of phenotypic or genetic heterogeneity, 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 Polymophism (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, multidigit +/- 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. often 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 4 partially completed *Salmonella* spp. genomes (*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 10 separate *Salmonella* spp. have been or are being sequenced and annotated. 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.

The whole genome comparison performed here serves on a basic level to document the prevalence and relative location of every gene in each of the sequenced *Salmonella* spp. The identification of specific genetic markers such as surface proteins for instance, that show moderate polymorphism in centralized locations among these 6 isolates can such whole genome comparative analyses, be identified as having 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 are not limited to detection and molecular epidemiological method development. Thus, the data generated in this study will have broad and significant impact. For instance the prevalence and phylogenetic relationship of virulence factors among these isolates can provide evidence that such a marker was recently added to a genome through lateral genetic transfer (LGT) providing enhanced pathogenic capabilities. This gives us a look at the evolution of this pathogen which can help us better understand and control it. On the other hand, if the putative virulence factors are found in some clinical isolates but not others this may suggest that these factors may not be required for virulence. This type of information could be important in the design of vaccines. This project represents the first and most logical step in the development of a reliable molecular epidemiological methods specifically applicable to *Salmonella* spp. typing for preharvest food safety.

This project will provides vital data needed in the development of new and reliable molecular epidemiological tools for *Salmonella* spp.. Several web based databases and analysis tools have 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.

## IV. Objectives:

**Overall Objective**

Perform whole-genome comparison of two *Salmonella* spp. genomes (*typhi* and *typhimurium*) and 4 partially completed Salmonella spp. genomes (*typhimurium*, *enterica*, *paratyphi*, and *enteritidis*) in order to identify specific markers that may be used as or as part of improved pre-harvest molecular epidemiological and detection targets.

**Specific objectives**

1. Perform whole genome comparison of 6 completely and partially sequenced *Salmonella* spp. genomes using a step by step sequential approach.
2. Build a database that can be used to catalogue and data mine this wealth of comparative information.
3. Specifically look at key genetic markers such as surface proteins (eg flagellin) and develop linkage maps and web based alignment views of these potential targets.

## V. Materials & Methods:

This project took the genetic information of 6 partially and fully sequenced Salmonella genomes from national databases, compared and catalogued this data and prepared user friendly web based structures for visualizing comparisons of this data.

*Bioinformatics Software Programs that were utilized*

DNAstar Lasergene package:

Lasergene provides multiple analysis tools for DNA and protein sequences including sequence alignment, contig assembly, gene discovery, primer design, restriction mapping, and protein structure prediction. Lasergene also integrates BLAST and Entrez searching for easy sequence downloading and comparisons.

*Wisconsin Package*
The Wisconsin Package is an integrated bioinformatics package featuring a comprehensive collection of DNA-, RNA-, and protein-sequence-analysis tools. Its functions span the entire range of sequence-analysis tasks from sequencing-run assembly, sequence editing, database searching, and pattern finding to nucleotide-chain folding.

*BLAST*
For the construction of genome alignments, a piece by piece and all-against-all comparison of the genetic sequences encoded in the complete *Salmonella* genomes was performed. Ten Kilobase (10KB) segments of the Salmonella enterica typhimurium LT2 genome were extracted and BLASTed against each of the other Salmonella genomes (partial and complete) as well as against the entire database using the appropriate BLAST programs found at the National Center for Biotechnology Information website (http://www.ncbi.nlm.nih.gov/blast). This provided the basic information upon which we built a database directed toward the development of epidemiological methods specific to *Salmonella* spp. The information provided by this basic database search and alignment tells us in essence where in comparison to the LT2 genome each of the other genomes encodes the same genetic sequences or more importantly divergent or nonhomologous sequences. We have completed a linkage map connecting each of the genomes. This same genetic information was then extracted from the database and used in building databases and other bioinformatics methods all directed toward identification of targets within the genome that can be used for epidemiological purposes.

*Database design*.
A novel database structure will ultimately be developed for translation and mining of this huge dataset compiled during this analyses. These databases were built upon MySql and were built so that it is useful to web based data mining. We must be flexible and open minded in the design of the database because as we compile data our needs will change and become more directed and focused. We were also aware and developed this database so that it is comprehensive and user friendly. The database interface were designed in HTML and published on the USDA-ARS-LIRU website http://199.133.147.108 .

*Local gene comparisons*:
Genes, such as the flagellin gene, from the fully sequenced *Salmonella typhimurium* LT2 genome including insertion sites 100bp upstream and 100bp downstream of the O-islands were entered into a local installation of Genome BLAST and search performed against the other 9 other *Salmonella* genomes. Upstream and downstream insertion sites of the genes in the other genomes were determined in this manner and the corresponding sequence extracted from them using EditSeq (DNAstar). The corresponding genes from all 6 genomes will then aligned using the Wilbur-Lipman pair-wise alignment method for large nucleotide sequence comparisons and Multialign (DNAstar) or Multiple sequence alignment features of the Wisconsin Package version 10. Alignments were evaluated and edited manually. Following the alignment those sequences, within the corresponding genes of the 10 genomes, where no homology existed were analyzed again using BLASTn to determine the significance of their genome rearrangements.

*Whole-Genome Comparison:*
While we were performing section by section comparison it was also useful to perform whole genome alignments to help us visualize on a larger scale the dataset we are working with. DNA alignment of whole genomes of the various combinations of the 10 genomes was accomplished using the MUMmer 2 programs (http://www.tigr.org/softlab). MUMmer is a system for aligning whole genome sequences. Using an efficient data structure called a suffix tree, the system was able rapidly to align sequences containing millions of nucleotides. Alignment levels were set to 10bp which increases computation time but also increases resolution. Graphs generated by the software were rendered into publication quality dotplots using Adobe illustrator.

*Other Programs and Methods Summary*
A wide variety of programs and methods are available to help us with this daunting task. Pairwise sequence comparison, multiple sequence alignments, consensus sequence determination from multiple sequence alignments, restriction mapping, virtual PCR, PCR primer design, linkage mapping and many other tasks were accomplished using the two software packages described above as well as various other software packages available as freeware. This data were transformed into usable information for the database in an ongoing fashion.

## VI. Results:

**Overall objective:** The progress toward reaching the overall object has exceeded expectations and the objectives set out in the original proposal. At this point we have completed alignments of the available data from genomes of 6 Salmonella species and are beginning to incorporate data from additional *Salmonella* genome sequencing projects that have started since our proposal was funded.

Specific Objectives:

1. We have using a systematic approach separated the genome of *Salmonella typhimurium* LT2 into 220 segments and used these segments to organize and extract sequence data from up to 9 additional *Salmonella* spp. genomes. Because some of the additional genomes are still in the process of being sequenced and annotated this approach provided a unique and amazing opportunity to evaluate these unpublished genomes. Thus we have in the past 12 months met and fully exceeded the first specific objective.
2. Our database has grown beyond expectation. The requirements for data storage exceeded our capacity and thanks to the generous support of the Cropping systems Research Laboratory we have added additional servers to house the database. We developed several cutting edge mining tools that have web based implementations. These tools are found on the bioinformatics page of our website http://199.133.147.108
3. Our comparative genomes approach has evaluated dozens of specific genes within this data and have uncovered several candidates that may allow for molecular “serotyping” that could be as specific as current antibody and phage related methods. We are focusing now on looking at large regions of the 10 *Salmonella* spp. genomes for regions of heterogeneity that can be used for development of even more sensitive molecular screening methods directed toward epidemiological tool development. One of the tools this study has helped to develop is a virulence gene targeted microarray containing close to 2500 genes.

## VII. Discussion:

Our results are quite impressive and meet and fully exceed the proposed objects especially from a bioinformatics standpoint. As stated previously we have completed alignment of the available data from 10 fully or partially sequenced *Salmonella* spp. genomes. We have put together a Gene Ontology website that originally focused on the 3 fully sequenced *Salmonella* and will also include the additional 7 sequenced *Salmonella* as their annotations become available. Links to these and other food safety genome and bioinformatics resources can be found on the bioinformatics link of our website http://199.133.147.108 . The web visualization of the ontology data provides a valuable comparative resource for genetic data beyond what was originally proposed. As part of this project we have also completed development a novel web based tool to allow for visualization of our database of genome alignments. This database can be accessed at the following url: http://199.133.147.108/salmonella/salmonella.html . These results will allow scientists to glean comparative information from these web based tools to allow for rapid development of improved methodologies related to epidemiology. The data from this experiment has led to the development of a virulence gene targeted microarray that may provide valuable epidemiological data that is both reproducible and sensitive. The final tool we are developing that is again over and above the

objectives of this project is comparative genome mapping data. Using our cMAP technology that automates the formatting and databasing of genetic maps we will soon have a complex interoperable genetic mapping infrastructure.

VIII. Lay Interpretation:

*Salmonella* research is the top priority for the pig industry especially related to pre-harvest reduction of pathogens with potential public health significance. The pre-harvest food safety research for *Salmonella* proposed here falls under the category of control strategy. The research proposed here is highly relevant and applied in nature. Without the groundwork, foundation and benefit of comprehensive genetic bioinformatics based analyses; many shotgun approaches based upon methods that are applicable to other microorganisms for instance, will ultimately fail when used for *Salmonella* spp. This project makes available to the industry and food safety scientists a powerful comparative sequence and genome comparison databases that can be used to develop epidemiological methods that are specific to *Salmonella* spp. rather than attempting to adapt methods that have proven to be of use 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. This project represents the primary and quite simply the only logical foundation step that will ever result in the development of a method that fulfills all the requirements of a superior molecular epidemiological method for *Salmonella* spp.