Application of Amplicon Length Polymorphism to differentiate amongst closely related strains of bacteria

W. C. Rice

Conservation and Production Research Laboratory, United States Department of Agriculture, Agricultural Research Service, Bushland, TX 79012 USA

The incidence of food-borne illnesses from pathogenic bacteria (e.g. *Escherichia coli* and *Salmonella* spp) remains relatively constant despite use of various abatement practices for reducing them. Methods to rapidly detect and discriminate between closely related pathogens are required to identify sources of food-borne illnesses. Pulsed field gel electrophoresis (PFGE) is currently considered the 'gold standard' molecular method for subtyping pathogens by the Centers for Disease Control and Prevention (CDC). The PFGE method requires four days to complete, thus slowing identification of a disease source. Application of amplicon length polymorphism (ALP) represents a novel genome-based molecular method to detect, identify and confirm the source of an outbreak. ALP can discriminate amongst related strains of pathogens as well as or to a higher degree of resolution than PFGE while also reducing both the time and expense required to confirm the source responsible for the food-borne illness. The theory underlying ALP and its ability to differentiate amongst closely related pathogens such as *Escherichia coli* and *Salmonella* spp. will be described. In principle, ALP can be applied to differentiate any group of closely related strains of bacteria. Since nearly 1100 microbial and archaeal genomes have been completely sequenced, with several hundred other sequencing projects underway, the ALP method should have widespread applicability.

Keywords amplicon length polymorphism, bacterial source tracking, PCR, genetic index, DNA fingerprint

Disclaimer: Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.

1. Introduction

The zoonotic pathogen *Escherichia coli* O157:H7 and other enterohemorrhagic *E. coli* and the genus *Salmonella* comprises some of the most important pathogen groups involved in human food-borne illness. Due to their worldwide distribution, these agents pose serious threats to public health systems. *Escherichia coli* O157:H7 was first recognized following two outbreaks in Michigan and Oregon in 1982 (Riley et al 1983) and can cause diarrhea, hemorrhagic colitis, and hemolytic uremic syndrome. The illness was associated with eating hamburgers at the restaurants of one national chain. Thus, hemorrhagic colitis due to *E. coli* 0157:H7 is commonly referred to as hamburger disease and infections most frequently occur from raw or undercooked hamburger. *E. coli* can be found in the feces of most food animals including pork, poultry, and dairy and beef cattle (Hancock et al 1994, Chapman et al 1997). In the United States, close to 75,000 cases of O157:H7 infections are now estimated to occur annually (Mead et al. 1999). Outbreaks of human salmonellosis are linked to consumption of contaminated dairy, poultry and meat products. Domestic animals act as a reservoir for non-typhoid *Salmonella*, which are responsible for millions of infections and multiple deaths in the human population costing billions of dollars (range 4 to 23) yearly (Todd 1990). Typhoid fever also remains an important global health problem in developing countries with a conservative estimate of 16 million cases and 600,000 deaths occurring annually (Ivanoff 1998; Pang *et al*. 1998). *S. enterica* is subdivided into over 2,000 serovars, and *S. bongori*. The degree of host adaptation of *Salmonella* serotypes is known to vary widely, thus three categories of *Salmonella* host adaptation have been proposed to help reduce the ambiguity when describing *Salmonella* serotypes (Uzzau et al., 2000). Thus, the epidemiological relationships amongst various *Salmonella* serotypes are a key concern in monitoring the presence and spread of disease.

2. General methods to type bacterial DNA

A number of research methods have been developed to compare the biochemical, phenotypic, and genetic relationship amongst related bacterial strains, serotypes, and biotypes. Generally these methods are viewed as being too time consuming and lacking in sufficient resolution amongst related strains. Numerous DNA based methods each with their respective advantages and disadvantages have been developed to type and subtype various microbial organisms (reviewed, Olive and Bean 1999). Many of these methods include a combination of methods such as an initial PCR amplification procedure, an enzymatic manipulation of the reactants, followed by a gel based electrophoresis assay, staining of products, gel imaging and subsequent data analysis. For typing of bacterial strains, the following; Locus specific restriction fragment length polymorphism (RFLP), randomly amplified polymorphic DNA (RAPD), Cleavase I fragment length polymorphism (CFLP), repetitive element PCR (rep-PCR), terminal-RFLP (T-RFLP), amplified

fragment length polymorphism (AFLP), pulsed field gel electrophoresis (PFGE), and DNA sequencing have been used with varying time, cost, and efficacy factors. Validation of primers, probes, and methods for the identification and differentiation of pathogens isolated from the environment is essential. Clinical laboratory methods require the use of proven protocols performed under GLP-GMP standards. PFGE is solely a macrorestriction method in which the use of a restriction endonuclease is used to cleave the DNA molecule at the recognition site for that enzyme. Specialized gel electrophoresis equipment is required to analyze the DNA fragments, thus imposing a significant cost over conventional gel based analysis methods. Use of PFGE has been greatly facilitated by the incorporation of standard methods of analysis suggested by Tenover et al. (1995), which has led to its widespread adoption.

For the purpose of monitoring food safety, bacterial source tracking generally relies on the application of PFGE as recommended by the CDC PulseNet site for its participants (http://www.cdc.gov/pulsenet/). Thus, PFGE is generally considered as the 'Gold Standard' and is used by the CDC, FDA, and USDA to subtype unknown pathogenic isolates.

3. Genomic data for the evaluation of microbial strains

In order to better understand the pathogenesis and evolution of *E. coli* O157:H7 and to develop better methods of identification, several groups have completed the DNA sequence determination of two genomes of *E. coli* O157:H7; isolates O157:H7 EDL933 (Perna et al. 2001) and O157:H7B Sakai (Hayashi et al. 2001). The Sakai strain was implicated in a major enterohaemorrhagic *E. coli* O157:H7 outbreak in Japan (Watanabe 1996). A 4.1 megabase ancestral genomic backbone was implicated between the DNA sequence of the two O157:H7 genomes and that of the previously sequenced common laboratory strain of *E. coli* K12 MG1655 (Blattner et al 1997). Evidence for a common *E. coli* genomic backbone had already been suggested by multilocus sequencing of seven housekeeping genes of *E. coli* O157:H7 along with isolates belonging to serotypes O6, O26, O55, O56, O91, O104, O111, O113, O127, and O128 (Reid et al. 2000). Significant lateral gene flow was observed between the genomes of *E. coli* O157:H7 EDL933, O157:H7 Sakai and E. coli, K12 MG1655 and is attributed in part to prophages. Approximately 800 kilobases of DNA are unique to the O157:H7 serotypes and thus contain the gene function(s) that are responsible for severe infections caused by *E. coli* O157:H7 serotypes. These include putative virulence factors, alternative metabolic capacities, prophages, and other new functions.

Similarly the genomes of key *Salmonella* isolates were also sequenced to facilitate the comparison of pathogenic isolates of *S*. Typhi CT18 (Parkhill et al. 2001) and *S*. Typhimurium LT2 (McClelland et al. 2001). The S. Typhi CT18 genome is genetically degraded with respect to that of the previously sequenced *Escherichia coli* genomes (Blattner, et al. 1997; Perna, et al. 2001; Hayashi, et al. 2001) and may help to explain the host restricted nature of *S*. Typhi CT18. In all, 601 genes (13.1%) in 82 blocks are unique to CT18 relative to LT2, while 479 genes (10.9%) in 80 blocks are unique to LT2 relative to CT18. Thus, much of the diversity of *Salmonella* genomes is located in gene clusters spread throughout the different genomes (McClelland et al., 2001; Parkhill et al., 2001).

The overall implication based on DNA sequence analysis of the above five mentioned genomes is that each group of zoonotic agent has an underlying genomic architecture. This is somewhat similar to the classical genetic observation of synteny or physical co-localization of genetic loci that has been observed in cereal crops. These ancestral genomic backbones form a structural architecture that may initially be exploited by an 'in silico' analysis based on a bioinformatics approach.

4. Bioinformatics Approach – the ALP method

ALP is a molecular method based on the use of information contained in microbial genomic DNA sequences of related strains (Rice, 2008; Rice, 2009). Approximately 1000 prokaryotic microbial genomes along with additional 100 or so archaeal genomes have been completely sequenced (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html). Complete DNA sequence data has now been obtained for 18 *E. coli* O157:H7 strains. Nearly 50 additional *E. coli* strains and approximately 40 *Salmonella* strains belonging to different serovars have also been sequenced.

4.1 Central Concept Concerning ALP Method

The precise genomic locations of the PCR primers used in the ALP method are known with respect to the indexed strain(s) (Rice, 2008; Rice, 2009); this is not the case with respect to all other PCR primers currently used in various typing methods. Thus, it is possible to create a "genetic index" of an unknown strain based on the observation of specific ALPs. For strains in which a large amount of phenotypic and genotypic data is available like *E. coli* K12 MG1655 and *S*. Typhi CT18, this can be used to anchor a global alignment, and they can serve as index strains. Due to a strain's overall genetic composition, it is possible to determine to some degree, if it is more 'O157:H7 like' or is it more 'K12 like' at a given genomic location. For *Salmonella,* a similar genetic index can be created, and the genetic relationship of an unknown isolate to the index strain can be determined. This is possible based on a strongly inferred ancestral genomic backbone identified from the overall global DNA alignment of closely related *E. coli* and *Salmonella*

spp. Determination of this genetic index is not possible with other bacterial typing methods. Thus, ALP facilitates rapid identification of genomic regions of high homology that are interspersed with regions of either: a) low homology, b) deletions, c) insertions or d) genetic rearrangements. Thus, the ALP method differs from all of the above previously mentioned DNA typing methods. For regions of low homology, a size difference may or may not be observed. However, this region is amenable to other molecular analyses such as restriction fragment length polymorphism, amplicon Tm prediction, high resolution melting curve analysis, and multi-locus specific sequencing, which may provide additional levels of information in which to differentiate strains at these specific locations.

4.2 Fast-scan analysis

Fast-scan analysis with the maximum score option selected (Align Plus v4.1 software, Sci-ed Software, Durham, NC) is based on the rapid search lookup table methods popularized by BLAST programs (Altschul, et al., 1990). Fast-scan is modified to allow for determination of homologous genomic regions with multiple gaps and frame shifts and is much faster than the Needleman-Wunsch method, which uses exhaustive base-by-base comparison. Application of fast-scan analysis is then conducted on large genomic sections (up to two megabases in length) in order to identify large homologous genomic regions. In this way large DNA blocks (15 to 45 kb) with a high degree of homology (≥ 97%) can be located within this two megabase region. This approach is then performed on two prokaryotic microbial genomes at a time and can be extended along the length of the genome in a pair wise manner using related microbial genomes as desired to identify all possible genomic regions amenable to global alignment methods. For a three genome evaluation of *E. coli* strains, the *E. coli* K12 MG1655 genome was aligned against the O157:H7 EDL933 genome and the O157:H7 EDL933 was aligned against the O157:H7B Sakai genome (Rice, 2009. For *Salmonella* spp. fast-scan analysis was conducted on *S*. Typhi CT18 (accession AL513382), and *S.* Typhimurium LT2 (accession AE00468) as indicated (Rice, 2008). Results obtained with fast-scan analysis will indicate the number and size of DNA homologous regions that are suitable for analysis in global DNA alignment settings.

4.3 Global DNA alignment

The location of large genomic blocks determined from fast-scan analysis is then used to organize a global alignment of desired target regions of the respective genomes. The CLUSTAL W program module contained within the Align Plus v4.1 software (Sci-ed Software, Durham, NC) was used to perform global alignments. Global DNA alignments of the genomes of *E. coli* strains K12 MG1655; accession L48811, O157:H7 EDL933; accession AE005174, and O157:H7B Sakai; accession BA000007, were performed as previously described (Rice, 2009). In this way major regions (ranging in size of 100-600 kb) of high homology (≥ 97%) were identified. Three way global DNA alignments were then conducted using either the K12 MG1655 or the O157:H7 EDL933 genome as the reference genome dependent upon which two genomic regions shared the greatest homology. This method identified multiple deletion sites contained within a genome (either the K12 or the two O157 genomes) such that it was possible to amplify a single unique PCR generated diagnostic fragment per site with respect to either the K12 or the two O157 genomes. Similarly global DNA alignment of *S*. Typhi CT18 (accession AL513382), and *S.* Typhimurium LT2 (accession AE00468) genomic DNA sequences was performed using approximately 1-2 megabase blocks of DNA. Thus, major regions (several hundred kb) of high homology (≥ 97%) were also located, and suitable genomic deletions within a genome (either the CT18 or the LT2 genome) were subsequently identified. DNA regions surrounding identified DNA deletions can then be examined for the presence of suitable PCR primers.

4.4 ALP primer design

Homologous genomic regions flanking small deletions identified by the Global Alignment program were evaluated using the Primer Design v4.2 program (Sci-ed Software, Durham, NC). An overall amplicon length of less than two kb and with a genomic deletion generally within the range of three to seven hundred bp was the upper target size limit for amplicon design. A critical success to the application of any PCR based assay is the availability of highly specific and well matched PCR primers based on Tm. Thus, fairly restrictive design criteria were used for the selection of ALP based PCR primers. Criteria used for primer construction were: length (20 bp), % GC (50-60), Tm °C (55-80), 3' dimers (< 3), dimers-any (< 7), stability (> 2.0 kcals 5' vs. 3'), runs (< 3), and repeats (< 3). This approach resulted in numerous ALP PCR primer sets spanning the *E. coli* and *Salmonella* genomes that produced amplicons of different size (length polymorphism) with respect to a specific reference microbial genome. Primers meeting the above criteria were then subjected to Primer Mix analysis (Primer Design v4.2 software) performed against other unrelated microbial genomes deposited in GenBank to confirm specificity for either *E. coli* or *Salmonella* strains. Selected ALP PCR primer sets were then used to generate computer derived amplicons for each genomic site for 'in silico' analysis. These analyses included restriction fragment length polymorphism, amplicon Tm prediction, and high resolution melting curve analysis.

5. Application of ALP to reference *E. coli* and beef cattle fecal isolates

Reference strains for which complete genomic DNA sequence is available are O157:H7 EDL933 (accession no. AE005174) and the O157:H7B Sakai (accession no. BA000007) from the National Food Safety and Toxicology Center at Michigan State University (East Lansing, MI), while the *E. coli* K12 MG1655 (accession no. U00096) strain was obtained from the *E. coli* Genetic Stock Center at Yale University (New Haven, CT). Reference strains of *E. coli* for which DNA genomic sequence information was not available were obtained from the ATCC (strain no. 35150, 35218, and 25922), the Texas Veterinary Research Lab (VDL O157:H7) and USDA, ARS NRRC laboratory (B6914 and B6914 GFP; Fratamico, et al., (1997)), and SEA 13B-88 Odwalla apple juice isolate, and O157:H7 Beltsville (USDA, ARS, Beltsville). Experimental *E. coli* strains were isolated from fecal swabs obtained from beef cattle. Fecal samples were obtained over four sampling dates and three antibiotic regimes. Primary isolation was on SMAC plates (both white and pink colonies were recovered) followed by re-isolation on *E. coli* CHROMagar (resulting in both blue and red colonies) in order to confirm strain purity. Isolates were stored as frozen milks until later analysis.

5.1 DNA Extraction

Three to five colonies were picked using a toothpick and suspended in 100 µL 1 X TE. The cell suspension (25 µL) was added to 150 µL of InstaGene Matrix (Bio-Rad, Hercules, CA), mixed well and incubated at 56 °C for 15 min. Samples were vortexed for 10 s and incubated at 95 °C for 8 min. Samples were vortexed again for 10 s and then centrifuged for 5 minutes at 4100 rpm. The supernatant was transferred to another plate and stored at -20°C. The matrix was carefully avoided to prevent PCR inhibition.

5.2 PCR

Primer sequences along with their genomic locations and amplicon designations were previously reported (Rice, 2008). Amplification was carried out in a 25 µL reaction volume containing 1.5 mM $MgCl_2$, 200 µM each dNTP, 1 unit of JumpStart *Taq* DNA polymerase (Sigma-Aldrich, St. Louis, MO), 400 nM each primer, and 2.5 µL template. Thermal cycling conditions for each reaction were as follows: initial denaturation at 95°C for 2 min; 30 cycles of denaturation at 95°C for 30 s, annealing (see Rice 2008 for temperatures) for 30 s, extension at 72°C for 1 min; and a final extension at 72°C for 7.5 min.

5.3 Amplicon analysis

Amplicons were separated using a 1.75% agarose gel with the size standard GeneRuler 100bp Plus DNA Ladder (Fermentas, Glen Burnie, MD). Amplicons generated from the primer sets (G, L, OB-O; B, OB-F; R, OB-A; E, O-C; F, O) were run separately and pooled. Gels were stained using SYBR Safe (Invitrogen, Carlsbad, CA) and imaged using the Kodak Image Station 4000MM (Carestream Health, Rochester, NY). Samples were also analyzed using the Experion Automated Electrophoresis Station (Bio-Rad, Hercules, CA) according manufacturer's recommendations for using the 1K DNA analysis chip (Bio-Rad, Hercules, CA).

6. Dataset analysis

Banding patterns for each isolate were recorded as a categorical code: i.e. the 'in silico' base pair value was used if the observed base pair value was within the predicted value range (Rice, 2009). Missing bands were coded as 0 and novel bands (outside a range of +/- 10%) were co-coded as the observed molecular weight value determined by size analysis based on agarose gel electrophoresis. All processing and analysis of the dataset was performed using Bionumerics software program version 5.0 (Applied Maths, Austin, TX). A similarity matrix was calculated using the Multi-state coefficient categorical, and cluster analysis of the matrix data was performed using the unweighted pair-group method, arithmetic average (UPGMA) (Sneath and Sokal 1973). The similarity matrices derived for the *E. coli* reference set and unknown *E. coli* isolates was based on data obtained from use of 13 ALP-PCR primer sets (B, D, E, F, G, L, O, R, OB-A, OB-F, OB-I, and O-E, Rice, 2009). Use of the Multi-state coefficient categorical provided the highest degree of differentiation of reference *E. coli* strains rather than using other coefficients (Jacquard's or Dice) based on binary datasets (data not shown).

7. Results

ALP-PCR primer sets previously described (Rice 2009) were capable of successfully resolving or indexing a set of known *E. coli* isolates. Primer sets B and OB-F can function in a multiplex assay to reduce assay time and expense. Data from these two primer sets also serve to illustrate some of the data types that are observed when performing ALP. Agarose gel electrophoresis analysis of APL-PCR primer sets B and OB-F generated the expected fragments for sequenced strains MG1655, EDL933, and Sakai (Fig 1). Of the eleven laboratory reference strains, eight of the strains

contained an O157:H7 backbone at the B and OB-F primer binding site whereas K12 MG1655 contained a K12 backbone at this site. However, ATCC 35218 (serotype unknown) and 25922 (O6) isolates lacked both the B and OB-F derived O157:H7 amplicons (437 and 558 bp respectively). Both strains were positive for a B K12 amplicon (522 bp), while the OB-F amplicon (314 bp) was absent. One interpretation would the lack of an O157:H7 backbone at the B and OB-F ALP defined sites and a missing OB-F (314 bp) K12 backbone. This may be due to the loss of one or both PCR binding sites associated with these specific ALP primer binding sites. This type of genome plasticity was clearly evident and somewhat abundant with both the *E. coli* DECA and *Salmonella* SARB sets of reference strains (Rice, 2009, 2008). A number of indexed ALP-PCR primer sets successfully detected unexpected polymorphisms within these reference strain sets thus, improving the level of discrimination at which related strains can be resolved. For the SARB reference strains, many of these sites were attributed to mobile genetic elements (Rice, 2008).

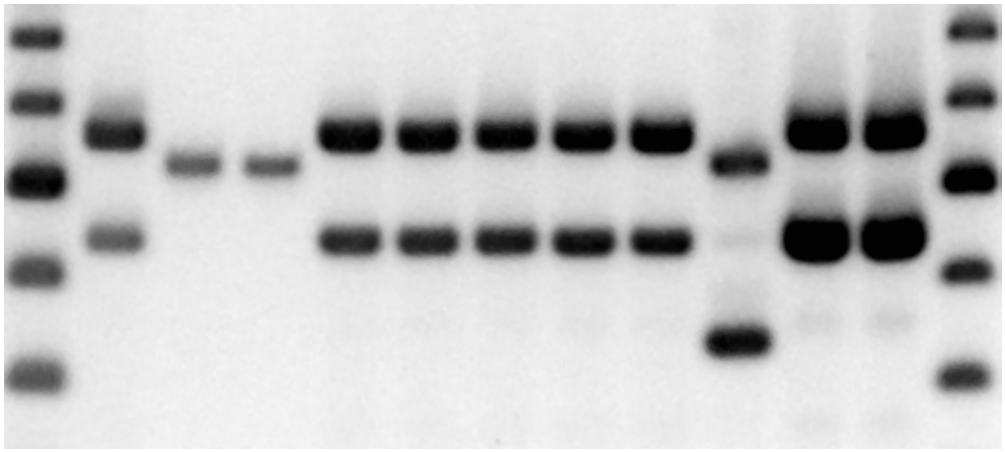

Fig 1 Amplicons generated by primer sets B, and OB-F for *E. coli* reference strains. Lanes: M, 1, ATCC 35150; 2, ATCC 35218; 3, ATCC 25922; 4, VDL O157:H7; 5, B6914; 6, B6914 GFP; 7, SEA 13B-88; 8, Beltsville O157:H7; 9, MG-1655; 10, EDL-933; and 11, SAKAI; M (MW; 700, 600, 500, 400, and 300 bp).

The overall genetic relationship of these reference strains is revealed by hierarchical cluster analysis using all of the primers previously listed. Cluster analysis of the ALP generated banding patterns revealed that MG1655 served as an anchor (Fig. 2) while the O157:H7 strains EDL933 and Sakai co-clustered with the known O157:H7 isolates. ATCC 25922 (serogroup O6) and ATCC 35218 (a non O157:H7 strain but of unknown serotype) grouped with K12 MG1655 whereas VDL K-99 formed its own out group. Three isolates were not completely resolved using these 13 primer sets. Interestingly B6914 GFP, a genetically engineered strain (green fluorescent protein) derived from B6914 groups more closely with two other strains than it does with its parent strain. Evidently during the transformation event there were genetic rearrangements other than just the acquisition of the plasmid encoding GFP. In this analysis it appears more closely linked to O157:H7 isolates obtained from an animal source and an apple juice source. In this case the use of additional primer sets described in Rice (2009) could be used to attempt to resolve these strains.

For clinical laboratory settings it is very desirable to streamline and automate as many of the data acquisition steps as possible. Data acquisition steps involved in ALP analysis are amenable to automation. ALP data can be acquired through the use of equipment that separates DNA amplicons using DNA chips based on Caliper DNA 1K and 12K chip designs (Caliper Life Sciences, Inc, Hopkinton, MA) instead of agarose gel electrophoresis. Reference strains MG1655 and EDL922 along with ATCC 25922 were analyzed with ALP primer sets R – OB-A, B – OB-F, and F – O (Fig. 3). Correct banding patterns are clearly visible along with very accurate DNA concentration levels. Advantages of a chip based analysis are that very low amounts of amplicon are required for analysis, a DNA imager is not required, gel images and associated data are automatically processed during the acquisition of the data generating image files in a variety of formats and Excel spreadsheets with mobility, molecular weight and concentration for immediate application. Many different ALP-PCR primer set reactions can be combined and run on these chips as long as there is sufficient size resolution between the amplicons, thus potentially reducing error while reducing the time required for the construction of datasets for downstream analysis.

Cluster analysis of an unknown set of *E. coli* strains obtained from beef cattle feces revealed three main clusters with an overall similarity of approximately 30% (Fig 4). Both MG1655 and EDL933 each served to anchor clusters (at a similarity of approximately 54% each) while a major cluster of 19 isolates grouped with a similarity of approximately 65%. ATCC 25922 (O6 serotype) was the most distantly related isolate based on this method (< 25% similarity). Several observations can be made at this point. Isolates from dates D1 and D2, the first two sampling dates separated by

45 days, formed their own clusters and this was supported by group separation analysis with 100% correct rates of assignment (data not shown). Isolates from dates D3 and D4 were observed to somewhat co-cluster and this effect seemed to be influenced by antibiotic treatment (M=Micotil vs. OT=oxytetracycline). Isolates SM10476, SM10481 appear to be identical even though SM10476 had a pink phenotype on SMAC plates while SM10481 had a white phenotype. The same can be said for isolates SM 10477 and SM10483. Isolates SM10497 and SM10503 are either siblings arising through the selection (enrichment procedure) or could possibly represent a dominant member of the *E. coli* population.

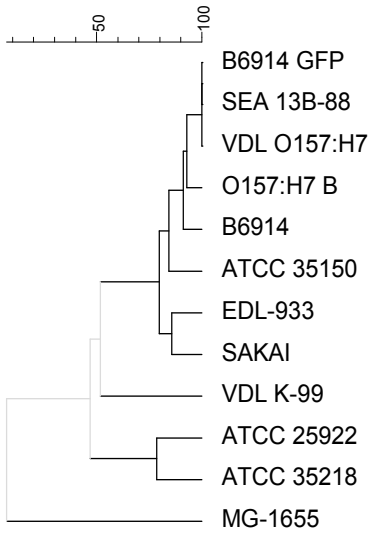

Fig 2 A UPGMA derived dendrogram from cluster analysis of a similarity matrix calculated using the multistate coefficient categorical. Dark solid lines indicate the location of significant clusters within the overall dendrogram. The similarity matrix is based on ALP analysis of *E. coli* reference strains obtained from the ATCC, the Texas Veterinary Research Lab and various USDA, ARS laboratories (see materials and methods). From left to right is the dendrogram and *E. coli* strain identifier.

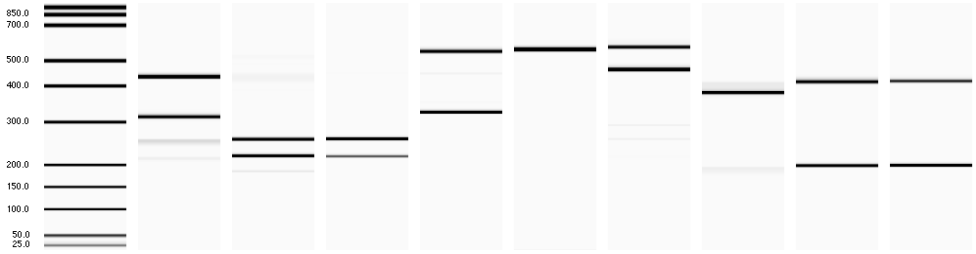

Fig 3 Lanes 1, MW ladder; 2-4 amplicons R and OB-A, strains (MG1655, ATCC 25922, EDL933); 5-7 amplicons B and OB-F, strains (MG1655, ATCC 25922, EDL933); 8-10, amplicons F and O, strains (MG1655, ATCC 25922, EDL933)

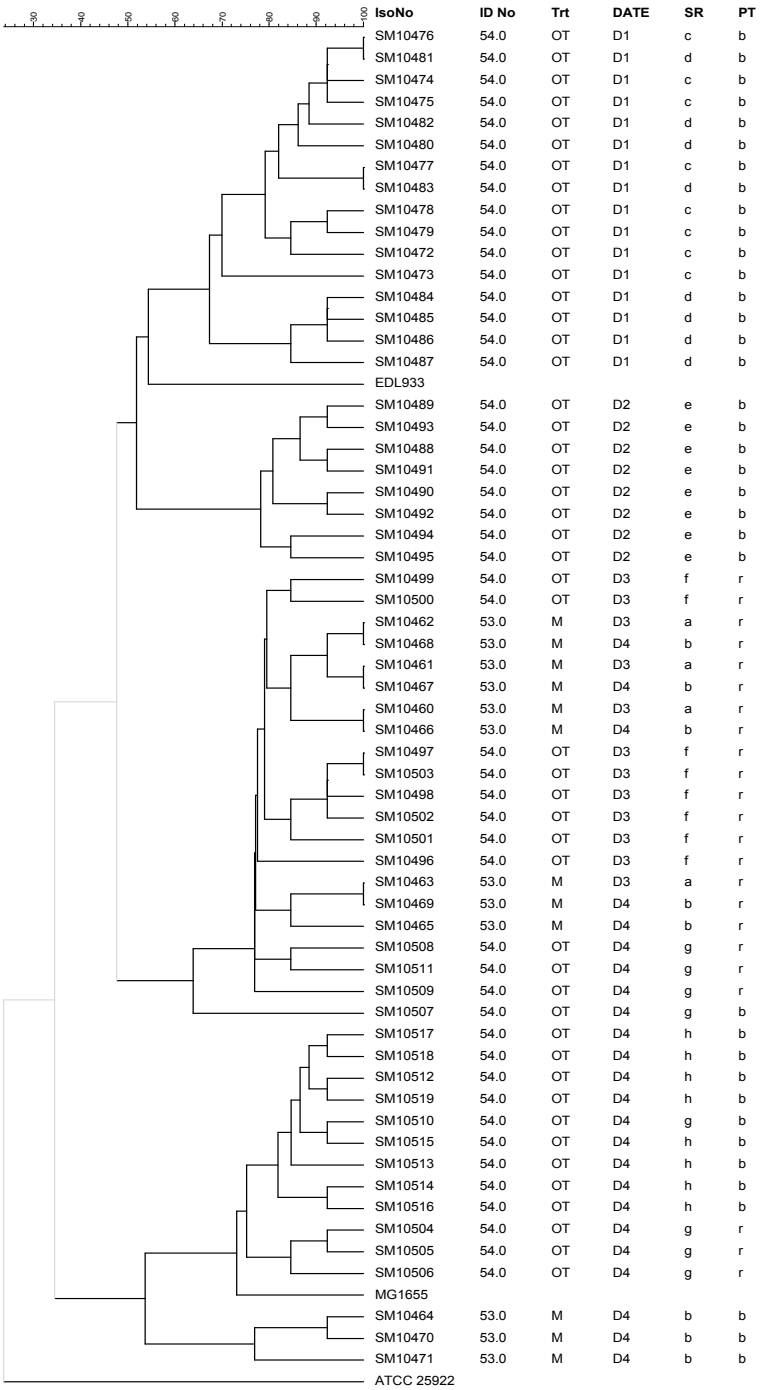

Fig 4 A UPGMA derived dendrogram from cluster analysis of a similarity matrix calculated using the multistate coefficient categorical. Dark solid lines indicate the location of significant clusters within the overall dendrogram. The similarity matrix is based on ALP analysis of *E. coli* isolates obtained from fecal swabs. The dendrogram is anchored by reference *E. coli* strains MG1655, ATCC 25922 (O6) and EDL933 (O157:H7). The column headings from left to right are: dendrogram, IsoNo (*E. coli* isolate), ID No (animal ID), Trt (antibiotic treatment), DATE (period of sampling), SR (sibling relationship), and PT (chromagar phenotype, r=red, b=blue).

8. Applications of ALP for the generation of other dataset types

The design of the primers used in the ALP-PCR method is based on known genomic DNA sequence information, thus facilitating the 'in silico' analysis of recently sequenced related microbial genomes. This creates information rich sites that are located on these genomes and can be exploited for the differentiation of these newly sequenced strains. In addition to the creation of an 'in silico' ALP dataset base on new genomic information; the specific amplicons may

provide data from melting assays (Tm prediction and high resolution melt curves), restriction fragment length assays (RFLP) and DNA sequence analysis (MLST). Thus, potentially four types of data may be derived from each amplicon to facilitate a high resolution definition of the relationship between various unknown isolates. Seven additional O157:H7 genome sequences were retrieved from GenBank and subjected to 'in silico' analysis (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html) along with the three *E. coli* reference strains. Data from only three ALP-PCR primer sets were analyzed using ALP polymorphism and amplicon Tm melting temperature prediction, combined with restriction endonuclease digestion using two different restriction enzyme analyses. This 'in silico' analysis resulted in the successful differentiation of these ten O157:H7 genomes into eight distinct groups (data not shown).

9. Conclusions

In this study of *E. coli* isolates, the ALP method has proven to be very robust in discriminating amongst closely related *E. coli* strains and has proven its value in another study of *E. coli* (Rice, 2009) and one on *Salmonella* species (Rice, 2008). The ALP method is readily amenable to automation and a small set of *E. coli* or *Salmonella* strains can be analyzed in one day versus a four day time frame for a comparable set of unknown strains using the PFGE method thus facilitating bacterial source tracking endeavors. The principles underlying ALP-PCR method should be applicable to other microorganisms with known DNA sequence information. The full potential of ALP has yet to be explored due to the lack of automated experimental platforms yet to be developed combined with of vast amounts of prokaryotic DNA sequence data being generated.

References

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. *Journal of Molecular Biology*, 215, 3, 403 (1990)
[2] Blattner, F.R., Plunkett, G.,3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. and Shao, Y. *Science* 277, 1453 (1997)
[3] Centers for Disease Control, Atlanta, GA. USA http://www.cdc.gov/pulsenet/
[4] Chapman, P.A., Siddons, C.A., Cerdan Malo, A.T. and Harkin, M.A. *Epidemiology and Infection* 119, 245 (1997)
[5] Fratamico, P.M., Deng, M.Y., Strobaugh, T.P. and Palumbo, S.A. *Journal of Food Protection* 60,1167 (1997)
[6] Hancock, D.D., Besser, T.E., Kinsel, M.L., Tarr, P.I., Rice, D.H. and Paros, M.G. *Epidemiology and Infection* 113, 199 (1994)
[7] Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C.G., Ohtsubo, E., Nakayama, K., Murata, T., Tanaka, M., Tobe, T., Iida, T., Takami, H., Honda, T., Sasakawa, C., Ogasawara, N., Yasunaga, T., Kuhara, S., Shiba, T., Hattori, M. and Shinagawa, H. DNA Res 8, 11 (2001)
[8] Ivanoff, B. *Med J Indonesia* 7, 5 (1998)
[9] Mead, P.S. *Emerging Infectious Diseases* 5, 607 (1999)
[10] McClelland, M., Sanderson, K.E., Spieth, J., Clifton, S.W., Latreille, P., Courtney, L., Porwollik, S., Ali, J., Dante, M., Du, F., Hou, S., Layman, D., Leonard, S., Nguyen, C., Scott, K., Holmes, A., Grewal, N., Mulvaney, E., Ryan, E., Sun, H., Florea, L., Miller, W., Stoneking, T., Nhan, M., Waterston, R. and Wilson, R.K. *Nature* 413, 852 (2001)
[11] National Library of Medicine, National Institutes of Health, Washington, D.C. http://www.ncbi.nlm.nih.gov/genomes/microbes/microbial_taxtree.html
[12] Olive, D.M., and Bean, P. *Journal of Clinical Microbiology* 37, 1661 (1999)
[13] Pang, T., Levine, M.M., Ivanoff, B., Wain, J. and Finlay, B.B. Trends Microbiol 6, 131 (1998)
[14] Parkhill, J., Dougan, G., James, K.D., Thomson, N.R., Pickard, D., Wain, J., Churcher, C., Mungall, K.L., Bentley, S.D., Holden, M.T., Sebaihia, M., Baker, S., Basham, D., Brooks, K., Chillingworth, T., Connerton, P., Cronin, A., Davis, P., Davies, R.M., Dowd, L., White, N., Farrar, J., Feltwell, T., Hamlin, N., Haque, A., Hien, T.T., Holroyd, S., Jagels, K., Krogh, A., Larsen, T.S., Leather, S., Moule, S., O'Gaora, P., Parry, C., Quail, M., Rutherford, K., Simmonds, M., Skelton, J., Stevens, K., Whitehead, S. and Barrell, B.G. *Nature* 413, 848 (2001)
[15] Perna, N.T., Plunkett, G. III, Burland, V., Mau, B., Glasner, JD., Rose, D. J, Mayhew, G.F., Evans, P.S., Gregor, J, Kirkpatrick, H.A., Posfai, G. Hackett, J, Klink, S., Boutin, A., Shao, Y., Miller, L., Grotbeck, E.J, Davis, N. W., Lim, A., Dimalanta, E., Potamousis, K., Apodaca, J, Anantharaman, T.S., Lin, J, Yen, G. Schwartz, D.C., Welch, R.A. and Blattner, F.R. *Nature* 409 (6819), 529(2001)
[16] Reid, S.D., Herbelin, C.J Bumbaugh, A.C. Selander, R.K., and Whittam, T.S. *Nature* 406, 64 (2000)
[17] Rice, W.C. Letters in Applied Microbiology, 47, 158 (2008)
[18] Rice, W.C. Journal of Applied Microbiology, 106, 149 (2009)
[19] Riley, L.W., Remis R.S., Helgerson, S.D. et al. *New England Journal of Medicine* 308, 681 (1983)
[20] Sneath, P.H.A. and Sokol, R.R. (1973) *Numerical Taxonomy*. San Francisco: Freeman.
[21] Tenover, F.C., Arbeit, R.D., Goering, R.V., Mickelsen, P.A., Murray, B.E., Pershing, D.H., and Swaminathan, B. *Journal of Clinical Microbiology*, 33, 9, 2233 (1995)
[22] Todd, E. Lancet 336 (1990)
[23] Uzzau, S., Brown, D.J., Wallis, T., Rubino, S., Leori, G., Bernard, S., Casadesus, J., Platt, D.J. and Olsen, *J.E. Epidemiol Infect* 125, 229 (2000)
[24] Watanabe, H., Wada, A., Inagaki, Y., Itoh, K, and Tamura, K. *Lancet* 348, 831 (1996)

