Genome sequencing and analysis of pathogenic *Escherichia coli* strains

Vivek Kapur, Dept. of Veterinary and Biomedical Sciences, Penn State
Chitrita "Chobi" DebRoy, Dept. of Veterinary and Biomedical Sciences, Penn State
Subhashinie Kariyawasam, Dept. of Veterinary and Biomedical Sciences, Penn State
Edward G. Dudley, Dept. of Food Science, Penn State

Chai-Mei Tang, Creatv MicroTech Inc., Potomac, MD
Pete Amstutz, Creatv MicroTech Inc., Potomac, MD
Peixuan Zhu, Creatv MicroTech Inc., Potomac, MD
Daniel Adams, Creatv MicroTech Inc., Potomac, MD

Daniel R. Shelton, Environmental Microbial and Food Safety Laboratory, USDA/ARS, Beltsville, MD
Jeffrey S. Karns, Environmental Microbial and Food Safety Laboratory, USDA/ARS, Beltsville, MD

Pina Fratamico, USDA/ARS, ERRC, Wyndmoor, PA

Robert Mandrell, Produce Safety and Microbiology, USDA/ARS, WRRC, Albany, CA

Peter Feng, U.S. FDA, HFS-711, College Park, MD
Patrick McDermott, FDA, CVM, Laurel, MD

Jim Bono, ARS, USDA USMARC, Meat Safety and Quality Research Unit, Clay Center, NE

Cheryl Bopp, Enteric Diseases Laboratory Branch, CDC, Atlanta, GA
Nancy A. Strockbine, Chief, Coordinating Center for Infectious Diseases, CDC, Atlanta, GA

Allison O'Brien, Uniformed Services University of the Health Sciences, Bethesda, MD

Lee W. Riley, School of Public Health, University of California, Berkeley, CA

James R. Johnson, Infectious Disease Unit, VA Hospital, Minneapolis, MN

Timothy J. Johnson, Dept. Veterinary and Biomedical Sciences, University of Minnesota, MN

1. INTRODUCTION

The species *Escherichia coli* includes both non-pathogenic and pathogenic strains found in the intestinal tract of mammals and birds. While *E. coli* is the most thoroughly studied bacterial species in the microbial world, much remains to be discovered about the genetic potential of pathogenic and non-pathogenic strains of this organism. Over the past several decades, *E. coli* has served as a model organism for the study of many of the fundamental processes of prokaryotic molecular biology, biochemistry, and evolution. A number of genome sequences of *E. coli* were published and have been deposited in the public databases over the past decade, and reveal the incredible diversity of this species. The typical *E. coli* genome is between 4.5 and 5.5 million base pairs (Mbp) in length (Blattner et al. 1997; Perna et al. 2001; Welch et al. 2002; Rasko et al. 2008, Touchon et al. 2009), encoding approximately 4500 to 5500 genes. By extrapolating current data, it is predicted that only about 2200 genes are conserved in all members of the species (the core genome) (Rasko et al. 2008, Touchon et al. 2009). The pan genome, or the total number of unique genes found in a species, is currently estimated to be over 20,000. *E. coli* is also thought to have an "open genome", as approximately 300 new genes (ie. about 6% of the total genes in any given strain) are annotated with every subsequent genome sequenced. In comparison, it is suggested that the pan genomes of *Streptococcus pyogenes* and *Streptococcus agalactiae* only grow by 1.5% (about 30 genes) with each new genome sequence reported, and the pan genome of *Bacillus anthracis* is completely known after sequencing only four strains (Tettelin et al. 2005). Therefore, while the sequences of many *E. coli* genomes have been characterized thus far, there is still much to be learned about the genetic potential of this organism.

Strains of *E. coli* commonly associated with food poisoning and other serious human illnesses often produce shiga toxins (Stx), a family of related protein toxins encoded by lambdoid prophages with two major types of Stx1 and Stx2. Shiga toxin was originally described from *Shigella dysenteriae* by Japanese bacteriologist Kiyoshi Shiga. Further studies showed that Shiga-like toxins were also expressed in other bacterial species such as enterohemorrhagic *E. coli* strains. Over 100 serotypes of Shiga toxin-producing *E. coli* (STEC) have been associated with human infections, including the most common serotype *E. coli* O157:H7, a major food-borne pathogen that has been implicated in many food-poisoning outbreaks worldwide. It is estimated that *E. coli* O157:H7 causes greater than 73,000 cases of illness and 61 deaths in humans due to hemolytic uremic syndrome (HUS) and hemorrhagic colitis (HC) each year in the United States. A total of 70 serogroups of non-O157 STEC have been described in the literature, and non-O157 strains belonging to serogroups O26, O45, O91, O103, O111, O121, and O145 and others have become important public health problems in the United States, and cause an estimated 37,000 cases of illness and 30 deaths each year (Mead et al. 1999, Tozzi et al. 2003, Sonntag et al. 2004). Since 2000, non-

Table 1. Complete genome sequences of non-O157 STEC

Gene Bank Acc. No.	E. coli serotype	Strain #	Genome size	Publication
AP010960	O111:H-	11128	5371077 bp	Ogura et al. 2009
AP010958	O103:H2	12009	5449314 bp	Hattori et al. 2009
AP010953	O26:H11	11368	5097240 bp	Ogura et al. 2009
FM180568	O127:H6	E2348/69	4965553 bp	Iguchi et al. 2009

O157 STEC infections became notifiable to the National Notifiable Disease Surveillance System in the Unites States. Increases in the incidence of disease caused by non-O157 STEC may be due to a change in the pathogen and/or an increased awareness of their role in human illness. The importance of STEC strains to *E. coli* community is also evidenced by the large number of publications in the last 10 years (ca. 4500) as well as the holding of numerous international conferences on Shiga-toxin producing *E. coli* strains. Finally, although several strains of *E. coli* O157:H7 have been sequenced, the genomes of only 4 of the 70 non-O157 STEC strains (serogroups O26, O111, O103 and O127) have been characterized thus far (Iguchi et al. 2009, Ogura et al. 2009; Table 1). It is widely recognized that this lack of genome scale information considerably limits our understanding of the genetics, pathogenic potential, and evolutionary history of this important group of organisms.

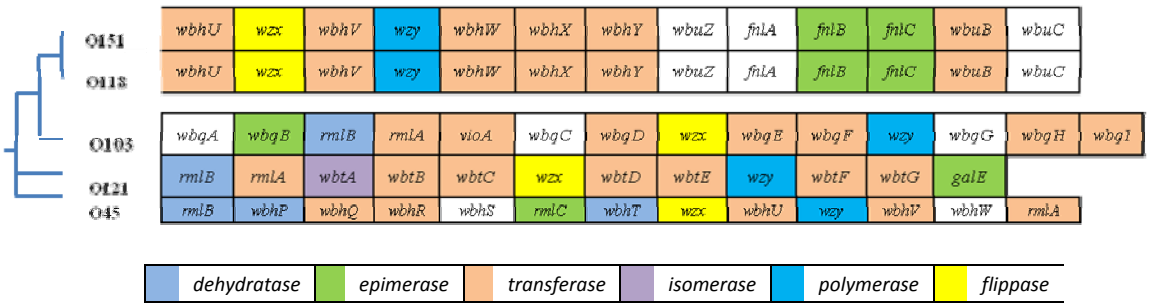
The proposed project seeks to help fill this knowledge gap through obtaining complete genome sequence information of five major human and animal pathogenic groups of *E. coli*. For two of these groups, the non-O157 STEC (n = 23 isolates) and non-STEC O157 (n = 4 isolates) groups, we propose sequencing of the genomes to completion to enable a rigorous comparative genomics analysis of this important group of human pathogens. For the other three groups, diarrheagenic *E. coli* (n = 21 isolates), Extraintestinal pathogenic *E. coli* (n = 8 isolates), and the reference isolates from international *E. coli* serotype collection (n = 112 isolates), we propose only draft sequencing so as to determine the genetic potential and mechanisms of pathogenesis of this important group of human and animal pathogens.

We note that the isolates were carefully selected by the community on the basis of their frequent isolation as human and animal pathogens, and the fact that they are well characterized in terms of virulence gene content and pathogenicity. We believe that the availability of the whole genome sequences of multiple O groups of STEC will enable researchers to: (a) Identify and determine the role and relevance of genes encoding virulence factors such as extracellular toxins, cell-surface antigens, and other molecules implicated as determinants of pathogenicity and disease specificity. (b) Study the molecular mechanisms involved in generating host specificity of clones recovered from human and animal infections. (c) Elucidate the molecular basis for the nonrandom association of certain bacterial clones with specific disease conditions in humans and other mammalian hosts. (d) Examine molecular mechanisms involved in the rise of new and unusually virulent bacterial clones and, (e) Identify specific genes and proteins suitable for use in the development of the next generation of diagnostic, therapeutic and immunoprophylactic agents.

2. BACKGROUND AND RATIONALE

Strains of *E. coli* are classified on the basis of somatic "O" antigens that are present on the surface of the bacteria. The O antigen is the polysaccharide unit of the gram-negative lipopolysaccharide (LPS), which is exposed on the surface of the bacteria. O-antigens are important virulence factors that are targets of both the immune system and bacteriophages. O-antigen specificity is very important for adaptation (Reeves, 1992), and its variation plays an important role in evasion of the host defenses (Reeves, 1995). The *E. coli* serotypes, therefore, are of great importance in epidemiological studies, in tracing the source of the outbreaks of gastrointestinal or other illness, or for linking the source to the infection. The O antigen gene cluster is responsible for O antigen production. Fig 1 shows the composition of O antigen gene clusters for the biosynthesis of O antigens of O118 and O151 strains that only differ by two nucleotides out of 13283 (Liu et al. 2008), suggesting that these two O groups are likely clonal and identifying a genetic basis for the differences in antigenic structure and immunological

Fig. 1. Comparison of genes in the O-antigen clusters of different O serogroups (the aa changes are absent for O151)



reactivity of the O antigens between these two serotypes. Thus, **a better understanding of the DNA sequences within the O-antigen gene cluster as well as a comparative genome scale analysis of variation in content and sequence of other genes in the genome among STEC will shed new light on the genetic basis of the evolution and mechanisms of virulence of this important group of organisms.**

Recent studies from the DebRoy laboratory show that while most of the O antigen gene clusters are unique in terms of gene content, some clusters only differed by a few nucleotides (Liu et al. 2008, Wang et al. 2007), suggesting that O-serotyping of *E. coli* may lead to false conclusions about their genetic background. Hence, **the availability of complete genome sequences of the isolates of *E. coli* as proposed herein will enable the development of the next generation of molecular diagnostic tools for superior genetic and virulence typing of isolates of this major human and animal pathogen.**

It has been observed that the virulence genes that confer pathogenicity of *E. coli* strains are specific not only for the serotypes but also for the host species from which they are derived from. For example *E. coli* O2 and O78 carrying the virulence genes such as *iss* or *tsh* are commonly found in avian spp. whereas strains belonging to serogroup O147 carrying virulence

genes *fedA* are only recovered from pigs. While *E. coli* O157:H7 are highly pathogenic to humans and cause HUS, they are not pathogenic to cows, goats, giraffes and other animals and are abundant in these host species (DebRoy and Roberts, 2006). However, the molecular basis of virulence and host association for certain O serogroups of *E. coli* is not understood, and it is anticipated that **the availability of genome sequences of these isolates may help elucidate the molecular mechanisms involved in the rise of new and unusually virulent bacterial clones and their host association.**

Table 2. Frequency and serogroup of commonly recovered STEC isolates at the ECRC.

O type	Total # isolates	% of total collection	% stx+	*Path
2	2140	5.7	8.2	
5	588	1.58	36	
8	3044	8.2	1.5	
11	451	1.2	8.4	
26	544	1.47	24.68	
45	234	0.63	10	
76	173	0.46	20	
88	502	1.35	9.6	
91	596	1.61	14.6	
103	350	0.94	40	
111	413	1.11	43.6	
113	220	0.59	30	
121	124	0.32	37.5	
128	298	0.8	23.2	
145	178	0.48	6.5	
147	157	0.42	70.6	
153	318	0.85	26.7	
174	303	0.8	44.8	

52 other STEC O groups carried <4 stx+ isolates and were not included for sequencing

*Path yellow to red depicts moderately pathogenic → highly-pathogenic.

Strain Selection of isolates for complete genome sequencing. Three criteria were used in the selection of the isolates proposed for sequencing: (1) a high frequency of isolation of strains from diseased humans and animals; (2) common STEC recovered from major domestic animal species (e.g., cows, pigs, and birds) with major contribution to the infectious disease burden from STEC in humans; and (3) well-characterized and described strains used for the investigation of the evolution and pathogenicity of *E. coli* by a large community of scientists. In addition, contemporary non-O157 STEC strains isolated from food (USDA strain from ground beef), water and fresh produces are also proposed to be included.

It is important to note that the selected strains were identified to represent the extent of genetic diversity amongst STEC strains of *E. coli*, based on data from over 37,000 *E. coli* clinical isolates serotyped at Penn State's *E. coli* Reference Center over the past 50 years. This represents the most comprehensive surveys of its type. The Center collects data frequency with which each serotype is observed, the host species, and the presence of known and putative virulence genes (Table 2).

As noted above, much of the focus from an epidemiologic, microbiologic, genomic, and diagnostic standpoint has been on E. coli O157:H7 that are STEC. However, it is increasingly recognized that in addition to non-O157 STEC isolates being a major cause of disease, several studies show that strains of E. coli that carry the O157 antigen but that do not contain the shiga-toxin genes may be as frequently isolated from fecal specimens and or contaminated food and water as are shiga-toxin producting isolates of O157. This remarkable observation provides an excellent opportunity for a thorough comparative genomic analysis that will enable a better understanding of the genetic basis of evolution of virulence in this major group of pathogens.

Hence, we propose to carry out the complete genome sequencing of both non-O157 STEC (n = 23) and non-STEC O157 (n = 4) isolates that have been carefully selected based on

the criteria defined above and after soliciting broad community input. Table 3 below details the characteristics of the proposed isolates for sequencing.

From a fundamental standpoint, the large dataset generated by this research will serve as an initial starting point for studies directed at understanding the ecology of *E. coli*. For example, it is well appreciated now that human pathogenic strains such as *E. coli* O157:H7 are not simply benign strains that have acquired a few toxin genes, but often are strains that share fewer than 70% of their genes with other well-characterized *E. coli* (Welch et al. 2002). The genome sequences can be used for providing insights into the complex phenomena such as host specificity and evolution of virulent strains. The rapid progression of *E. coli* O157:H7 from an unknown strain in 1982 to a widespread cause of foodborne illness worldwide a few years later highlights the critical need to understand the potential of this species to acquire and lose genes, and also to alter the regulation of existing genes. The data will also serve as the basis for correlation of genome sequence with pathogenicity. In addition, Shelton et al. of USDA have isolated and identified a number of *E. coli* O157 non-H7 serotype strains from environmental samples (Shelton et al., 2003; 2004; 2006). Serological and genetic characterization showed that some of these strains do not possess any virulence factors, raising a question about selection and genetic transfer and exchange of virulence genes between *E. coli* serotypes. Feng et al. of FDA have further investigated 19 strains of O157 non-H7 serotype and revealed that these strains were belong to a new clonal group with genetic distant to the known O157 STEC strains (unpublished data, 2009). In this project, we will select representative strain of *E. coli* O157 non-H7 serotype from USDA bacterial collections for whole genome sequencing to better understand the spread and evolution of the virulence genes in STEC.

Table 3. List of major serogroups of non-O157 STEC isolates and non-STEC O157 isolates proposed for complete genome sequencing.

	Strain No.	O type	H type	stx1	stx2	Host Species
1	1.2741	2	4	+	+	cow
2	97.0246	5		+	+	cow
3	5.0588	8		+	+	cow
4	97.0259	11		-	+	cow
5	96.1528	26	11	-	+	human
6	H30	26	11	+	-	human
7	95.0941	45	2	+	-	human
8	1.2264	76		+	+	goat
9	97.0264	88	25	+	+	cow
10	96.0497	91	21	+		human
11	99.0741	91		+	+	food
12	3.2608	103	2	+	-	horse
13	93.0624	103	6	+	-	human
14	4.0522	111		+	+	cow
15	JB1-95	111		+	+	human
16	96.154	113	12	-	+	human
17	5.0959	121	19	-	+	?
18	0.2732	121		-	+	pig
19	9.0111	128	2	+	+	human
20	4.0967	145	2	-	+	rabbit
21	2.3916	147		-	+	pig
22	3.3884	153		+	+	cow
23	97.0263	174		+	+	cow
24	USDA-2B	157	12	-	-	water
25	USA-3.2303	157	16	-	-	water
26	USDA.3003	157	45	-	-	water
27	ARG-HC7793	157	39	-	-	water

Strain Selection and draft genome sequencing of major pathogenic *E. coli* groups. As noted above, non-STEC E. coli are responsible for many different disease manifestations in humans. For instance, the diarrheagenic E. coli (DEC) strains are divided into different categories, enterohemorrhagic (EHEC), Enteropathogenic (EPEC), enteroaggregative (EAEC) and enterotoxigenic (ETEC). Some of these strains are very well characterized as such as the diarrheagenic *E. coli* (DEC) collection belonging to EHEC and EPEC categories (Read et al. 1999). EAEC strains are more common in infantile diarrhea in developing countries, other pathogenic *E. coli* that cause UTI in humans and necrotizing pneumonia in animals called Extraintestinal pathogenic *E. coli* (ExPEC), and ExPEC strains belong to 3 different groups, two of each representative groups have been included. Representative and well characterized isolates from each of these subgroups are proposed for draft genome sequencing as shown in Tables 4 and 5 (Appendix).

Finally, we propose draft sequence analysis of a total of 112 strains selected for sequencing represented in Table 6 (Appendix) are the O groups of which the O antigen biosynthetic genes are not currently sequenced. We chose the strains currently in the ECRC collection that were obtained from the Staten Institut (World Health Organization) and have been used as standard reference strains for O typing. By using these strains, we avoid the possibility of sequencing an *E. coli* strain that may be misclassified due to previously discussed issued with traditional antibody-based serotyping.

The sequence information generated here would also assist those who are developing rapid methods of detecting and tracking the spread of *E. coli*. Better and effective schemes for epidemiological studies can be developed to trace the source of the outbreaks of gastrointestinal or other illness, or for linking the source to the infection. This would help the community at large who work on pathogenic *E. coli* strains and those who are developing vaccines for certain O groups responsible for causing diseases in humans and animals.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nuc Acids Res* **25**, 3389-3402.

Blattner, F., Plunkett, G., Bloch, C., Perna, N., Burland, V., Riley, M., Collado-Vides, J., Glasner, J., Rode, C., Mayhew, G., Gregor, J., Davis, N., Kirkpatrick, H., Goeden, M., Rose, D., Mau, B. and Shao, Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453-1474.

Carver, T., Berriman, M., Tivey, A., Patel, C., Bohme, U., Barrell, B.G., Parkhill, J. and Rajandream, M.-A. (2008) Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24**, 2672-2676.

DebRoy, C., Fratamico, P.M., Roberts, E., Davis, M.A. and Liu, Y. (2005) Development of PCR assays targeting genes in O-antigen gene clusters for detection and identification of *Escherichia coli* O45 and O55 serogroups. *Appl Environ Microbiol* **71**, 4919-4924.

DebRoy, C. and Maddox, C. 2001. Assessing virulence of *Escherichia coli* isolates of veterinary significance. Animal Health Res. Rev. 1: 129-140.

DebRoy, C., Yealy, J., Wilson, R.A., Bright, B.D., Bhan, M.K. and Kumar, R.K. 1995. Antibodies raised against outer membrane protein interrupts adherence of enteroaggregative *Escherichia coli*. Infect. Immun. 63: 2873-2879.

DebRoy, C. and Roberts, E. 2006. Screening petting zoo animals for the presence of potentially pathogenic *Escherichia coli.* J. Vet. Diag. Invest. 18: 597-600

DebRoy, C., Roberts, E., Kundrat, J., Davis, M.A., Briggs, C.E. and Fratamico, P.M. (2004) Detection of *Escherichia coli* serogroups O26 and O113 by PCR amplification of the *wzx* and *wzy* genes. *Appl Environ Microbiol* **70**, 1830-1832.

DebRoy, C., Roberts, E., Jayarao, B. and Brooks, J. 2008. Extraintestinal Pathogenic *E. coli* (ExPEC) Induced Bronchopneumonia in a Horse. J.. Vet. Diagn. Invest. 20: 661-664

Iguchi A, Thomson NR, Ogura Y, Saunders D, Ooka T, Henderson IR, Harris D, Asadulghani M, Kurokawa K, Dean P, Kenny B, Quail MA, Thurston S, Dougan G, Hayashi T, Parkhill J, Frankel G. Complete genome sequence and comparative genome analysis of enteropathogenic *Escherichia coli* O127:H6 strain E2348/69. J Bacteriol.191:347-354.

Johnson, J.R., Johnston, B., Clabots, C.R., Kuskowski, M.A., Roberts, E. and **DebRoy, C.** 2008. Virulence genotypes and phylogenetic background of *Escherichia coli* serogroup O6 isolates from humans, dogs and cats. J. Clin. Microbiol. 46:417-422.

Johnson J.R., Miller, S., Johnston B., Clabots, C. DebRoy, C. 2009. Sharing of *Escherichia coli* sequence Type ST131 and other Multidrug-Resistant Urovirulent *E. coli* Strains among Dogs and Cats Within a Household. J. Clin. Microbiol. 47 (11): in press

Lan, R. and Reeves, P.R. 1996. Gene transfer is a major factor in bacterial evolution. Mol Biol Evol. 13:47-55

Liu, Y., Fratamico, P., DebRoy, C., Bumbaugh, A.C. and Allen, J.W. (2008) DNA sequencing and identification of serogroup-specific genes in the *Escherichia coli* O118 O antigen gene cluster and demonstration of antigenic diversity but only minor variation in DNA sequence of the O antigen clusters of *E. coli* O118 and O151. *Foodborne Pathogens and Disease* **5**, 449-457.

Mead, P.S., Slutsker, L., Dietz, V., McCaig, L.F., bresee, J.S., Shapiro, C., Griffin, P.M., Tauxe, R.V. 1999. Food-related illness and death in the United States. Emerg. Inf. Dis. 5:607-625.

Ogura,Y., Ooka,T., Iguchi,A., Toh,H., Asadulghani,M., Oshima,K., Kodama,T., Abe,H., Nakayama,K., Kurokawa,K., Tobe,T., Hattori,M. and Hayashi,T.2009. Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic *Escherichia coli*. Proc. Natl. Acad. Sci. U.S.A. (2009) In press

Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.-Y., Cohoon, M., de Crecy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E.D., Gerdes, S., Glass, E.M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A.C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G.D., Rodionov, D.A., Ruckert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O. and Vonstein, V. (2005) The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. *Nucl Acids Res* **33**, 5691-5702.

Perna, N.T., Plunkett, G., Burland, V., Mau, B., Glasner, J.D., 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.T., Potamousis, K.D., Apodaca, J., Anantharaman, T.S., Lin, J., Yen, G., Schwartz, D.C., Welch, R.A. and Blattner, F.R. (2001) G enome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409**, 529-533.

Ramachandani, M., Manges, A.R., DebRoy, C., Johnson, J.R. and Riley, L.W. 2005. Possible animal origin of human multidrug-resistant uropathogenic *Escherichia coli*. Clin. Infect. Dis. 40: 251-257.

Rasko, D.A., Rosovitz, M.J., Myers, G.S.A., Mongodin, E.F., Fricke, W.F., Gajer, P., Crabtree, J., Sperandio, V. and Ravel, J. (2008) The pan-genome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J Bacteriol* **190**, 6881-6893.

Read, S.D., Betting, D.J. Whittam, T.S. 1999. Molecular Detection and Identification of Intimin Alleles in Pathogenic *Escherichia coli* by Multiplex PCR. J. Clin. Microbiol.37:2719-2722.

Reeves, P.R. 1992. Variation in O-antigens, niche-specific selection and bacterial populations. FEMS Microbiol Lett. 79:509-516

Shelton DR, Higgins JA, Van Kessel JA, Pachepsky YA, Belt K, Karns JS. 2004. Estimation of viable *Escherichia coli* O157 in surface waters using enrichment in conjunction with immunological detection. J Microbiol Methods. 2004 Aug;58(2):223-31.

Shelton DR, Van Kessel JA, Wachtel MR, Belt KT, Karns JS. 2003. Evaluation of parameters affecting quantitative detection of *Escherichia coli* O157 in enriched water samples using immunomagnetic electrochemiluminescence. J Microbiol Methods. 2003 Dec;55(3):717-25.

Shelton DR, Karns JS, Higgins JA, Van Kessel JA, Perdue ML, Belt KT, Russell-Anelli J, DebRoy C. 2006. Impact of microbial diversity on rapid detection of enterohemorrhagic *Escherichia coli* in surface waters. FEMS Microbiol Lett. 2006 Aug;261(1):95-101.

Sonntag AK, Prager R, Bielaszewska M, Zhang W, Fruth A, Tschäpe H, Karch H. 2004 Phenotypic and genotypic analyses of enterohemorrhagic *Escherichia coli* O145 strains from patients in Germany. J Clin Microbiol.42:954-962.

Sura, R., Van Kruiningen, H. J., **DebRoy, C**., Hinckley, L. S., Greenberg, K. J., Gordon Z., and French R. A. 2007. Extraintestinal pathogenic *E.coli* (ExPEC) induced acute necrotizing pneumonia in cats. Zoonoses Public Health 54:307-313.

Tatusov, R., Fedorova, N., Jackson, J., Jacobs, A., Kiryutin, B., Koonin, E., Krylov, D., Mazumder, R., Mekhedov, S., Nikolskaya, A., Rao, B.S., Smirnov, S., Sverdlov, A., Vasudevan, S., Wolf, Y., Yin, J. and Natale, D. (2003) The COG database: an updated version includes eukaryotes. *BMC* Bioinformatics **4**, 41.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., DeBoy, R.T., Davidsen, T.M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J.D., Hauser, C.R., Sundaram, J.P., Nelson, W.C., Madupu, R., Brinkac, L.M., Dodson, R.J., Rosovitz, M.J., Sullivan, S.A., Daugherty, S.C., Haft, D.H., Selengut, J., Gwinn, M.L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K.J.B., Smith, S., Utterback, T.R., White, O., Rubens, C.E., Grandi, G., Madoff, L.C., Kasper, D.L., Telford, J.L., Wessels, M.R., Rappuoli, R. and Fraser, C.M. (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial pan-genome. Proc Natl Acad Sci USA **102**, 13950-13955.

Touchon, Marie., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen6, E., Bonacorsi, S., Bouchier, C., Bouvet, O., Calteau, A., Chiapello, H., Clermont, O., Cruveiller, S., Danchin, A., Diard, M., Dossat, C., El Karoui, M., Frapy, E., Garry, L., Ghigo, J.M., Gilles, A.M., Johnson, J., Le Bougue´nec, C., Lescat, M., Mangenot, S., Martinez-Je´hanne, V. Matic, I., Nassif, X., Oztas, S., Petit, M. A., Pichon, C., Rouy, Z.,Ruf, C.S., Schneider, D., Tourret, J., Vacherie, B., Vallenet, D., Medigue, C.,. Rocha, E. P. C 2, Denamur E. 2009. Organised Genome Dynamics in the *Escherichia coli* Species Results in Highly Diverse Adaptive Paths. PLoS Genetics, 5(1)

Tozzi, A.E., Caprioli, A., Minelli, F., Gianviti, A, De Petris. L, Edefonti. A, Montini, G, Ferretti, A, De Palo, T, Gaido M, Rizzoni G. 2003. Shiga toxin-producing *Escherichia coli* infections associated with hemolytic uremic syndrome, Italy, 1988-2000. Emerg Infect Dis. 9:106-108

Tusnady, G.E. and Simon, I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics* **17**, 849-850.

Wang, W., Perepelov, A.V., Feng, L., Shevelev, S.D., Wang, Q., Senchenkova, S.y.N., Han, W., Li, Y., Shashkov, A.S., Knirel, Y.A., Reeves, P.R. and Wang, L. (2007) A group of *Escherichia coli* and *Salmonella enterica* O antigens sharing a common backbone structure. *Microbiology* **153**, 2159-2167.

Welch, R.A., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., Buckles, E.L., Liou, S.R., Boutin, A., Hackett, J., Stroud, D., Mayhew, G.F., Rose, D.J., Zhou, S., Schwartz, D.C., Perna, N.T., Mobley, H.L.T., Donnenberg, M.S. and Blattner, F.R. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci USA* **99**, 17020-17024.

.

APPENDICES

Table 4. Diarrheagenic *E.coli* strains proposed for sequencing (n=21)

Strain No.	Original strain no.	Serotype	Date	Type	Country	Host	Reference
1	C54-58 (1b)	O55:H6	1958	EPEC	Suriname	Human	Read et al. 1999
2	5513-56 (2b)	O55:H2	1956	EPEC	USA	Human	Read et al. 1999
3	493/89 (3f)	O157:H2	1989	EHEC	Germany	Human	Read et al. 1999
4	5338-66 (6a)	O111:H21	1966	NK*	USA	Human	Read et al. 1999
5	C142-54 (6b)	O111:H12	1954	NK	Germany	Human	Read et al. 1999
6	750001 (7a)	O157:H43	1975	NK	USA	Pig	Read et al. 1999
7	902034 (7b)	O149:H2	1990	NK	USA	Pig	Read et al. 1999
8	C240-52 (9c)	O26:H2	1952	EHEC	Switzerland	Human	Read et al. 1999
9	900105 (10e)	O26:H11	1990	EHEC	USA	Calf	Read et al. 1999
10	RDEC-1 (10f)	O15:H2	1970s	EHEC	USA	Rabbit	Read et al. 1999
11	C309-64 (10g)	O128:H8	1964	EHEC	ND	Human	Read et al. 1999
12	C186-61 (10h)	O119:H11	1961	EHEC	ND	Human	Read et al. 1999
13	87-1713 (10i)	O145:H6	1987	EHEC	Canada	Human	Read et al. 1999
14	88817 (10j)	O70:H11	1988	EHEC	Canada	Human	Read et al. 1999
15	2254-75 (11a)	O128:H2	1975	EPEC	USA	Human	Read et al. 1999
16	A9619-c2 (11c)	O45:H2	1983	EPEC	USA	Human	Read et al. 1999
17	3350-73 (13a)	O128:H7	1973	NK	USA	Human	Read et al. 1999
18	C691-71 (14b)	O128:H21	1971	NK	India	Human	Read et al. 1999
19	F03	O4:H7	1994	EAEC	India	Human	DebRoy et al 1995
20	H16	O78:H-	1994	EAEC	India	Human	DebRoy et al 1995
21	B41	O101:HNM	1980	ETEC	USA	Pig	DebRoy Maddox, 2001

*NK: Not Known

Table 5: Extraintestinal pathogenic strains proposed for sequencing (n=8)

-

Strain	Original Strain No	Serotype	Date	Type	Country	Host	Reference
1	88.0368	O17:H18	1988	ExPEC	USA	Cow	Ramchandani et al. 2004
2	SEQ895	O17	NK	ExPEC	USA	Human	Ramchandani et al. 2004
3	Outbreak strain	O15:K52:H1	NK	ExPEC	UK	Human	Ramchandani et al. 2004
4	5.3169	O25:H4	2005	ExPEC	USA	Human	Johnson et al. (in preparation)
5	8.2256	O25:H4	2008	ExPEC	USA	Dog	Johnson et al. 2009
6	85.1284	O6:H31	1985	ExPEC	USA	Human	Johnson et al. 2008
7	85.0143	O6:H31	1985	ExPEC	USA	Dog	Johnson et al. 2008
8	6.1680	O4:H5	2006	ExPEC	USA	Cat	Sura et al. 2007

Table 6. Reference strains belonging to 112 different O groups proposed for shot-gun sequencing

Sr #	O type	Strain	Serotype
1	O-2	U9-41	O2:K1:H4
2	O-5	U1-41	O5:K4:H4
3	O-6	Bi7458-41	O6:H2a:H1
4	O-8	G3404-41	O8:K8:H4
5	O-9	Bi316-42	O9:K9:H12
6	O-10	Bi8337-41	O10:K5:H4
7	O-11	Bi623-42	O11:K10:H10
8	O-12	Bi626-42	O12:K5:H-
9	O-16	F11119-41	O16:K1:H-
10	O-17	K12a	O17:K16:H18
11	O-18	F10018-41	O18ab;K-:H14
12	O-19	F8188-41	O19ab:K-:H7
13	O-20	P7a	O20:K17:H-
14	O-23	E39a	O23:K18ab:H15
15	O-25	E47a	O25:K19:H12
16	O-27	F9884-41	O27:K-:H-
17	O-29	Su4338-41	O29:K-:H10
18	O-30	P2a	O30:H-
19	O-33	E40	O33:K-:H-
20	O-34	H304	O34:K-:H10
21	O-35	E77a	O35:K-:H10
22	O-36	H502a	O36:K-:H9
23	O-37	H510c	O37:K-:H10
24	O-38	F11621-41	O38:K-:H26
25	O-39	H7	O39:K-:H-
26	O-41	H710c	O41:K-:H40
27	O-42	P11a	O42:K-:H37
28	O-43	Bi7455-41	O43:K-:H2
29	O-44	H702c	O44:H18
30	O-46	P1c	O46:K-:H16
31	O-48	U8-41	O48:K-:H-
32	O-49	U12-41	O49:K+:H12
33	O-50	U18-41	O50:K-:H4
34	O-51	U19-41	O51:K-:H24
35	O-53	Bi7327-41	O53:K-:H3
36	O-54	Su3972-41	O54:K-:H2
37	O-57	F8198-41	O57:K-:H-
38	O-60	F10167a-41	O60:K-:H33
39	O-61	F10167b-41	O61:K-:H19
40	O-62	F10524-41	O62:K-:H30
41	O-63	F10598-41	O63:K-:H-
42	O-64	K6b	O64:K-:H-
43	O-65	K11a	O65:K-:H-
44	O-66	P1a	O66:K-:H25
45	O-68	P7d	O68:K-:H4
46	O-69	P9b	O69:K-:H38
47	O-70	P9c	O70:K-:H42

48	O-71	P10a	O71:K-:H12
49	O-74	E3a	O74:K-:39
50	O-75	E3b	O75:K95:H5
51	O-76	E5d	O76:K-:H8
52	O-78	E38	O78:H-
53	O-79	E49	O79:K-:H40
54	O-80	E71	O80:K-:H26
55	O-81	H5	O81:K97:H-
56	O-82	H14	O82:K-:H-
57	O-83	H17a	O83:K-:H31
58	O-84	H19	O84:K-:H21
59	O-85	H23	O85:K-:H1
60	O-87	H40	O87:K-:H12
61	O-88	H53	O88:K-:H25
62	O-89	H68	O89:K-:H16
63	O-90	H77	O90:K-:H-
64	O-91	H307b	O91:K-:H-
65	O-92	H308a	092:K-:H33
66	O-95	H311a	O95:K+:H33
67	O-96	H319	O96:K-:H19
68	O-97	H320a	O97:K-:H-
69	O-99	H504c	O99:K-:H33
70	O-100	H509a	O100:K-:H2
71	O-101	H510a	O101:K-:H33
72	O-102	H511	O102:K-:H40
73	O-105	H520b	O105:K-:H8
74	O-108	H708b	O108:K-:H10
75	O-109	H709c	O109:K-:H19
76	O-110	H711c	O110:K-:H39
77	O-115	27w	O115:K-:H18
78	O-116	28w	O116:K+:H10
79	O-119	34w	O119:H27
80	O-120	35w	O120:K18a:H6
81	O-124	227	O124:H30
82	O-125ab	2745-53	O125ab:H19
83	O-125ac	2129-54	O125ac:H6
84	O-131	S239 (=H27w)	O131:K-:H26
85	O-132	N87 (=H30w)	O132:K+:H28
86	O-133	N282 (=H31w)	O133:K-:H29
87	O-134	4370-53	O134:K-:H35
88	O-136	1111-55	O136:H-
89	O-137	RVC1787	O137:H41
90	O-140	149-51	O140:K-:H43
91	O-142	C771	O142:H6
92	O-144	1624-56	O144:K-:H-
93	O-153	14097	O153:K-:H7
94	O-154	E1020-72	O154:K94:H4
95	O-156	E1585-68	O156:K-:H47
96	O-158	E1020-72	O158:K-:H23
97	O-160	E110-69	O160:K-:H34

98	O-161	E223-69	O161:K-:H54
99	O-162	10B1-1	O162:K-:H10
100	O-163	SN3B-1	O163:K-:H19
101	O-165	E78634	O165:K-:H-
102	O-166	3866-54	O166:K-:H4
103	O-169	1792-54	O169:K-:H8
104	O-170	745-54	O170:K-:H1
105	O-171	198	O171:K-:H2
106	O-173	L119B-10	O173:K-:H-
107	O-175	2533-54	O175:K-:H28
108	O-176	E29518-83	OX176:H-
109	O-177	E40874-85	OX177:H25
110	O-178	E54071-88	OX178:H7
111	O-180	86-381	OX180:H-
112	O-181	92-1250	OX181:H49