**International Journal of Medical and Health Research**
**ISSN: 2454-9142, Impact Factor: RJIF 5.54**
**www.medicalsjournals.com**
**Volume 2; Issue 3; March 2016; Page No. 03-10**

# *In silico* comparative genomic analysis of *Escherichia coli* E24377A and *Escherichia coli* HS

**Ammara Masood, Nadia Iqbal, Rubab Zahra Naqvi, Hira Mubeen**

National Institute for Biotechnology and Genetic Engineering, Faisalabad, Pakistan

**Abstract**
Comparative genome sequence analysis is a powerful technique for gaining insights into any genome of interest. E.coli E24377A is a human pathogen, which causes diarrhea. After the whole genome of two strains of *E.coli*, the virulent E24377A and non-pathogenic, commensal HS were sequenced; there is a hope that comparing the genomes will allow an identification of the genes responsible for its virulence and thus the development of treatment and control. Several virulent proteins are also being investigated, as virulence factors as potential vaccine or drug targets. Here we describe the comparison between the genomes of two strains of *E.coli* with few existing genomics databases and tools available in the public domain websites. By comparing nucleotide and protein sequences of the two strains, we investigate the existing differences and similarities.

**Keywords:** In silico Genomics, virulence factors; *E.coli* HS, *E.coli* E24377A.

## Introduction

Comparative genomics is a relatively new field of study that has arisen as a major tool to find meaning in newly sequenced genomes. In studying prokaryotic systems comparative genomics has proven especially useful, leading to a better understanding of systematics, bacterial lifestyle, virulence, and host pathogen interactions. One application of comparative genomics that shows a great deal of promise is the study of virulence. Each species has a different pattern of virulence, by comparing these species on the sequence level, they hope to identify genetic features that account for these differences (1).
Sequencing of whole microbial genomes is re-shaping the fields like microbiology, biotechnology, molecular biology, biochemistry, etc. *E.coli* was discovered by German pediatrician and bacteriologist Theodor Escherich in 1885 (2) and is now classified as part of the Enterobacteriaceae family of gamma-proteobacteria. *Escherichia coli* is a model laboratory bacterium, a species that is widely distributed in the environment, as well as a mutualist and pathogen in its human hosts. As such, *E. coli* represents an attractive organism to study how environment impacts microbial genome structure and function (3). It is commonly found in the lower intestine of warm-blooded animals. *Escherichia coli* represents a versatile and diverse enterobacterial species which can be subdivided into
(i) nonpathogenic, commensal
(ii) intestinal pathogenic
(iii) extra intestinal pathogenic strains.

Enterotoxigenic *Escherichia coli*, or ETEC, is the name given to a group of *E. coli* that produce special toxins which stimulate the lining of the intestines causing them to secrete excessive fluid, thus producing diarrhea. Enterotoxigenic *E. coli* (ETEC) is one of the most devastating pathogenic E. coli, which can cause watery traveler'a diarrhea.

**Strain E24377A***: E. coli* E24377A (pathogenic) has been shown to contain CFA pili types CS1 and CS3. Two additional toxins are thought to be responsible for the virulence, a heat stable and a heat labile enterotoxin. Other virulence factors include serotype O139:H28 and potential factors encoded on 6 uncharacterized plasmids. Traveler's diarrhea is caused by ETEC *E. coli* E24377A. The pathogenesis of diarrhea due to ETEC depends on 2 known virulence factors: fimbrial adhesins, which allow the organism to stick to intestinal epithelium and resist the clearing action of peristalsis, and elaboration of either or both of 2 toxins, a heat-stable toxin (ST) and a heat-labile toxin (LT). ETEC enterotoxins activate cyclic nucleotide synthesis in the gut epithelium that causes secretion by crypt cells of fluid and electrolytes into the small intestinal lumen, resulting in watery diarrhea. Advances in understanding the pathogenesis of ETEC diarrhea have led to the development of vaccine candidates against ETEC. (4)

**Strain HS:** *E. coli* strain *HS* is a human commensal. Strain *HS* colonizes the human gastrointestinal tract in challenge experiments, but no overt signs of disease occur. The genome sequence of strain HS represents the genomic baseline for colonization of the gastrointestinal tract. This isolate is serotype O9, motile, competent and amenable to genetic manipulation. (5)
Because of the multi-drug resistance nature of the Escherichial strains, we need the deeper understanding of the virulence factors. For that, the comparative analysis of genes and proteins may provide more insight on their resistance nature and virulence factors. Therefore, the availability of sequence data for the strains E24377A and HS provides a unique opportunity to compare their genes and proteins for the comparison of virulence nature between the two strains.

## Materials and methods

In order to retrieve the complete genome sequences, annotated gene and protein sequences list of E24377A and HS, we have used NCBI (National Center for Biotechnology Information). For the comparison of two genomes, we have used the genomic BLAST (Basic Local Alignment and Search Tool).

Following bioinformatics tools (software and databases) were used for comparative genomic analysis:
CLUSTAL W (http://www.ebi.ac.uk/Tools/clustalw/)
Justbio (www.justbio.com)
Microbesonline database (http://www.microbesonline.org/)
Microbial genome database (http://mbgd.genome.ad.jp/)
Integrated microbial genome (http://img.jgi.doe.gov)
TIGR(http://www.jcvi.org/cms/home/)
KEGG(http://www.genome.jp/kegg/)

**Results**

Comparative genomics and *in silico* studies have begun to reveal insights into gene and protein functions of many bacterial species and strains. In our present work, we would like to consider the comparison of genome features, the whole genome alignment and comparison of gene role category particularly virulence factors between two strains, E24377A and HS of *Escherichia coli*. Here we have made the comparative study by using the available public domain databases and tools and the results are discussed below.

**Comparison of the genome features:** Both E24377A and *HS* strain have similar cellular features table 1. The genome sequence of an organism provides information about the size of the genome, base composition, gene content, number of RNAs, number of direct and inverted repeats and other features. The genome features of two strains are obtained from "microbes online" and "integrated microbial genome" and the results are provided in table 1 for comparison. E24377A is the largest based on their genome size as compared to *HS*. No significant difference was observed in the GC content, number of rRNA genes, tRNA genes (Table1).

**Table 1:** Comparison of *Escherichia coli* E24377A and *Escherichia coli HS* genome features.

| Sr. # | Features | Escherichia *coli* E24377A | Escherichia *coli* HS |
|---|---|---|---|
| 1 | Strain | *E.coli*_E24377A, 331111 | *E.coli*_HS, 331112 |
| 2 | Genome size | 4980187 | 4643538 |
| 3 | DNA type | Linear | Linear |
| 4 | Genes | 5258 | 4628 |
| 5 | Ref sequence | NC_009801 | NC_009800 |
| 6 | RNA genes | 113 | 158 |
| 7 | Protein coding genes | 5145 | 4470 |
| 8 | GC content % | 50.62% | 50.82% |
| 9 | DNA coding no of bases | 4617796 | 4136299 |
| 10 | DNA G+C no of bases | 2654167 | 2359828 |
| 11 | Pseudo genes | 151 | 87 |
| 12 | rRNA genes | 21 | 22 |
| 13 | 5S rRNA | 7 | 8 |
| 14 | 16S rRNA | 7 | 7 |
| 15 | 18S rRNA | 0 | 0 |
| 16 | 23S rRNA | 7 | 7 |
| 17 | 28S rRNA | 0 | 0 |
| 18 | tRNA genes | 91 | 88 |
| 19 | Other RNA genes | 1 | 48 |
| 20 | Genes with function prediction | 3734 | 3481 |
| 21 | Without function prediction | 1411 | 989 |
| 22 | Genes with enzymes | 1216 | 1219 |
| 23 | Genes in orthologous clusters | 4896 | 4332 |
| 24 | No of chromosomal cassette | 1032 | 854 |
| 25 | Genes coding signal peptide | 864 | 788 |
| 26 | transmembrane coding protein | 1104 | 337 |
| 27 | Obsolete genes | 0 | 0 |
| 28 | revised genes | 3 | 4 |
| 29 | %age coding sequence | 85% | 85% |
| 30 | Cell Shape | Rod | Rod |
| 31 | endospore | - | - |
| 32 | Range | mesophillic | mesophillic |
| 33 | Motility | Yes | yes |
| 34 | Oxygen requirement | facultative | facultative |
| 35 | Habitat | Host associated | Host associated |
| 36 | Optimum temprature | 37°C | 37°C |
| 37 | Arrangement | Singles, pairs | Singles, pairs |

**Whole genome alignment**

The genome stability was analyzed between E24377A and *HS* using the genome alignment tool-Genomics BLASTn. They share 99% similarities (95073/95685 with 25/95885 gaps). Further genomic BLAST was again performed for proteins of these two strains using BLASTx. Complete gene distribution map of both genomes, is given in fig 1.

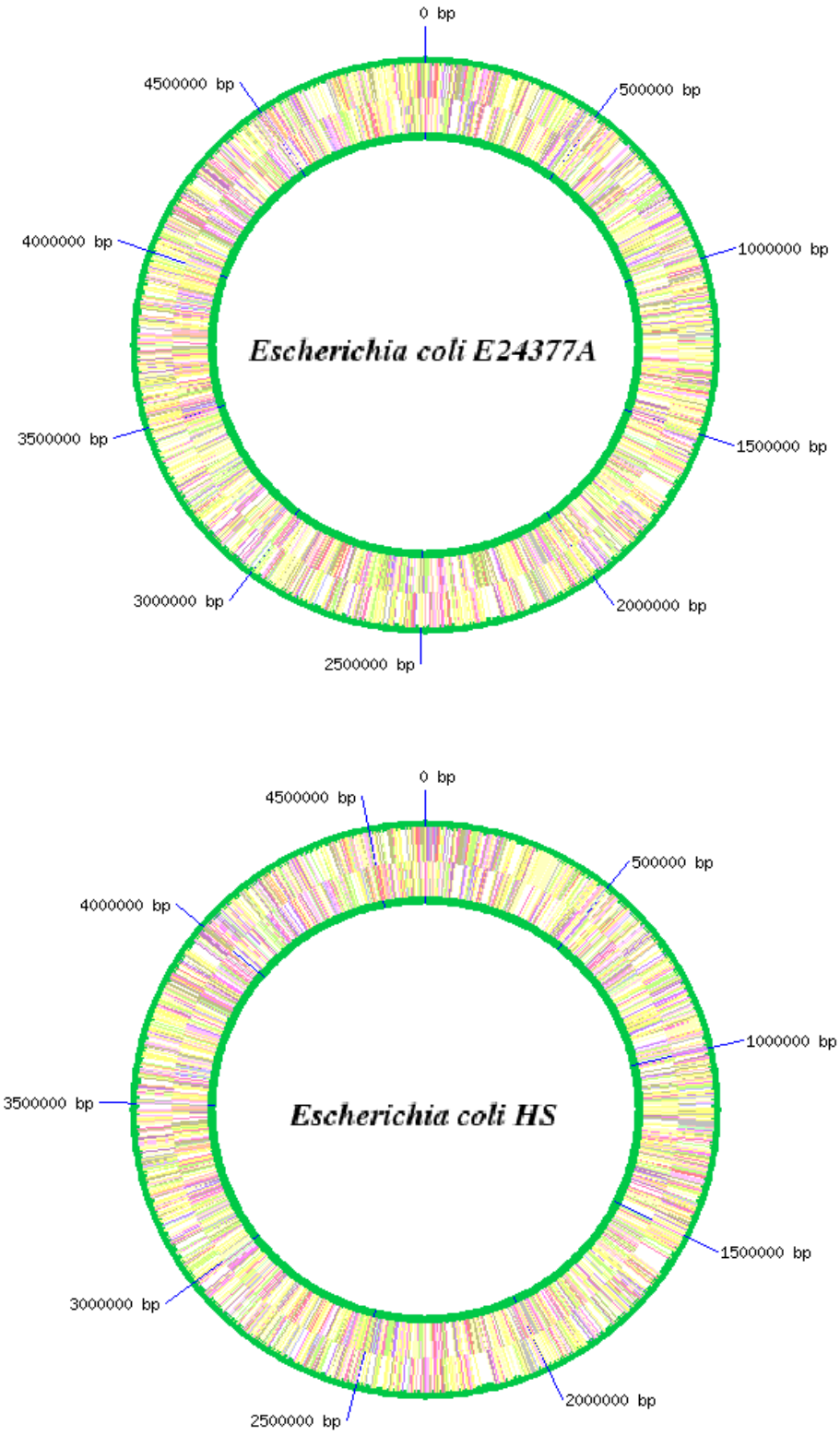


**Fig 1:** Complete Gene Distribution Map of two genomes E24377A and *HS*

**Gene Role Category Comparison**

In role category, the genes responsible for RNA processing and modification, Amino acid transport and metabolism, Lipid metabolism, Translation, ribosomal structure and biogenesis Posttranslational modification, protein turnover, chaperones and signal transduction mechanism of E24377A are nearly same as in *HS*, this suggests that the basic complement of proteins required for certain cellular processes in two strains are more or less same. Major cellular systems and features of E24377A that are notably different from the genome *HS* include the genes involved in Energy production and conversion, Cell envelope biogenesis, outer membrane, Cell motility and secretion. Locating the open reading frames in both genome is also very important as Open reading frame is continuous stretch of codons that do not have stop codon. The transcription termination site is located after the ORF, beyond the translation stop codon, because if transcription were to cease before the stop codon, an incomplete protein would be made during translation. Complete open reading frame map of both genomes are given in figure 2 and 3.

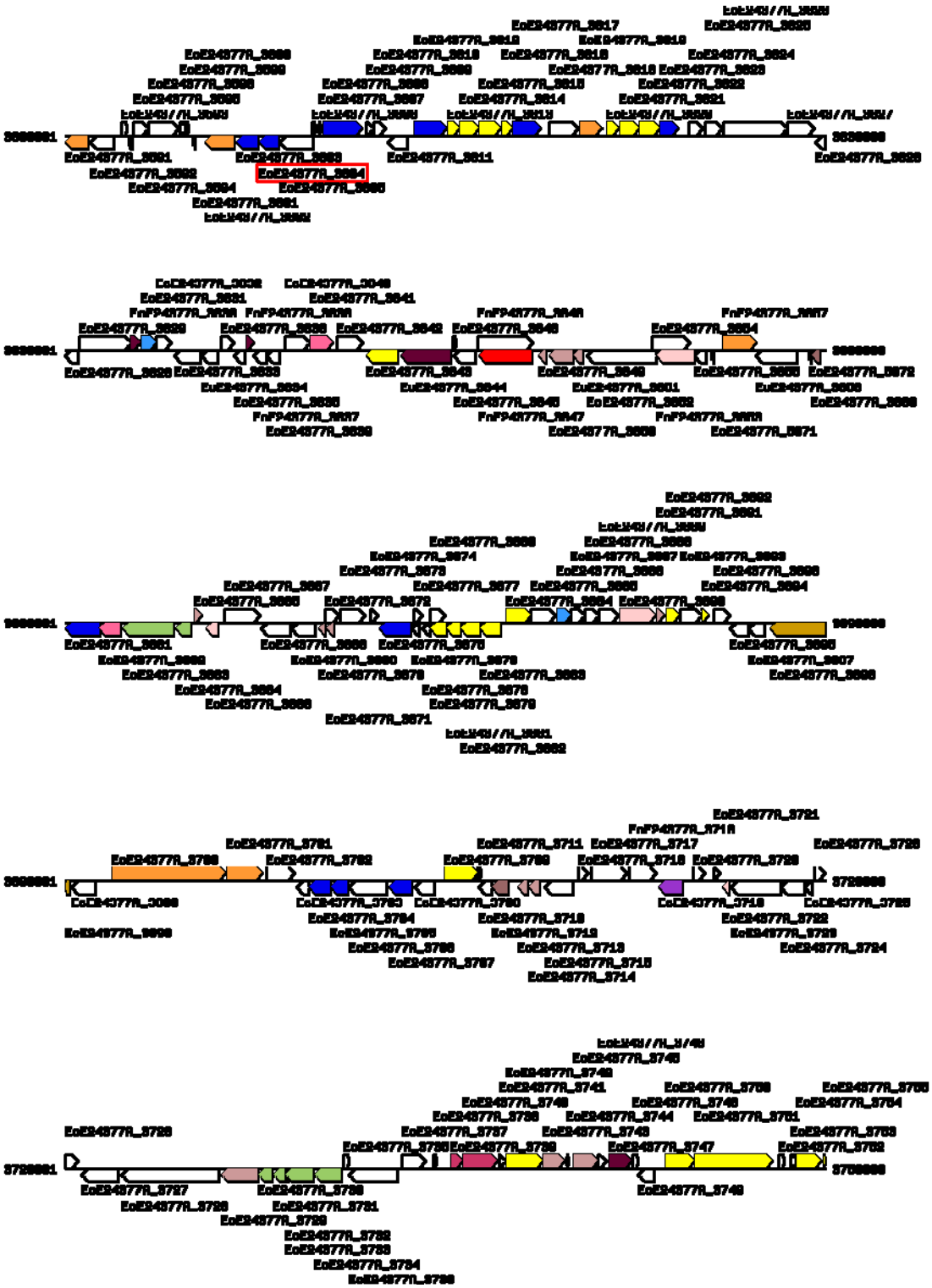

**Fig 2:** The complete ORF's map of the *Escherichia coli* E24377A. The thick colored arrows represent the position size and direction of transcription of the proposed ORF's of E24377A.

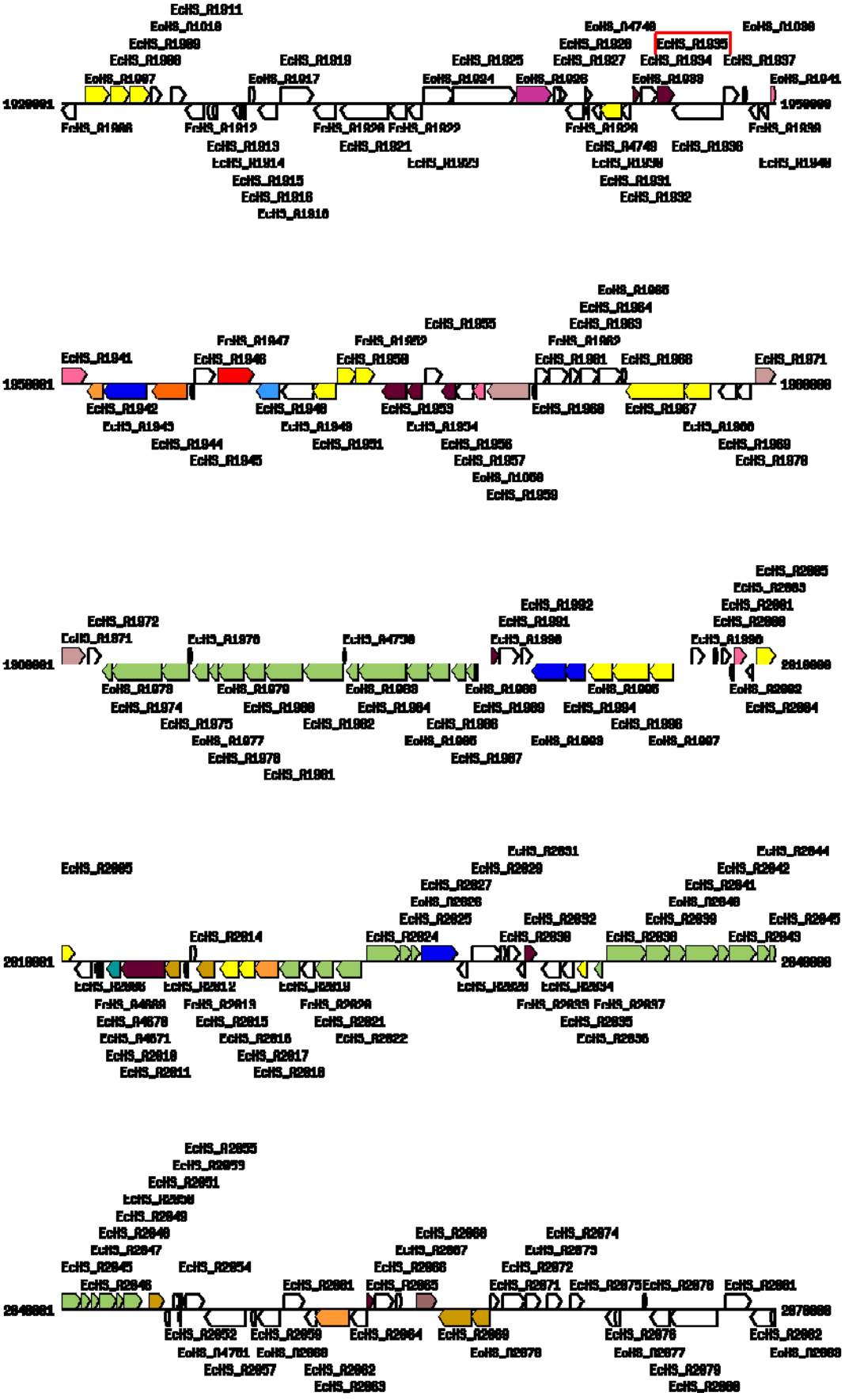

**Fig 3:** The complete ORF's map of the *Escherichia coli* HS. The thick colored arrows represent the position size and direction of transcription of the proposed ORF's of *E.coli* HS.

**Comparison of plasmids features**

*E. coli* HS has no plasmid DNA whereas *E coli* E24377A has 6 plasmids. As plasmids are important in pathogenesis so a comparison of these six plasmids were performed. (Table 2).

**Table 2:** Comparison of sequence attributes *Escherichia coli* E24377A plasmids**.**

| Component name | Ref Seq | Gen Bank | Protein count | Length (bp) | Av. CDS Length | GC content (%) | CDS coverage | Gene count |
|---|---|---|---|---|---|---|---|---|
| Chromosome | NC_009801 | CP000800 | 4755 | 4979619 | 907.097 | 50.6% | 87% | 4755 |
| Plasmid pETEC_5 | NC_009791 | CP000801 | 3 | 5033 | 628 | 49.6% | 37% | 3 |
| Plasmid pETEC_6 | NC_009789 | CP000798 | 5 | 6199 | 498 | 52.6% | 40% | 5 |
| Plasmid pETEC_35 | NC_009787 | CP000796 | 29 | 34367 | 736.655 | 51.6% | 62% | 29 |
| Plasmid pETEC_73 | NC_009788 | CP000797 | 68 | 70609 | 660.044 | 50.2% | 64% | 68 |
| Plasmid pETEC_74 | NC_009790 | CP000799 | 69 | 74224 | 717.261 | 49.8% | 67% | 69 |
| Plasmid pETEC_80 | NC_009786 | CP000795 | 68 | 79237 | 628.853 | 47.3% | 54% | 68 |

**Specific Target Identification for strain typing:**

Specific loci were investigated by BLASTn similarity searches for plasmid sequences against nr nucleotide database. 11 coding sequences were found very specific to *E. coli* E24377A plasmids and these are not present in any other living organism which is sequenced yet. These target sequences would be used for multilocus PCR diagnostics (Table 3).

**Table 3:** E24377A plasmids specific coding sequence for PCR diagnosis.

| Sr. # | Plasmid NCBI's Accession # | Location of CDS (bp) | NCBI's 0GeneID | Locus tag | Product |
|---|---|---|---|---|---|
| 1 | NC_0097876 | 3259-3753 | 5585705 | EcE24377A_F0003 | hypothetical protein |
| 2 | NC_0097876 | 8505-9449 | 5585687 | EcE24377A_F0009 | hypothetical protein |
| 3 | NC_0097876 | 28194-28400 | 5585686 | EcE24377A_F0034 | hypothetical protein |
| 4 | NC_0097876 | 28702-29328 | 5585669 | EcE24377A_F0035 | hypothetical protein |
| 5 | NC_0097876 | 60596-61735 | 5585670 | EcE24377A_F0070 | hypothetical protein |
| 6 | NC_0097876 | 61768-62934 | 5585696 | EcE24377A_F0071 | hypothetical protein |
| 7 | NC_0097876 | 62891-63664 | 5585675 | EcE24377A_F0072 | hypothetical protein |
| 8 | NC_0097876 | 69998-71284 | 5585634 | EcE24377A_F0081 | hypothetical protein |
| 9 | NC_009787 | 24197-25201 | 5585723 | EcE24377A_C0020 | hypothetical protein |
| 10 | NC_009787 | 27303-28130 | 5585714 | EcE24377A_C0024 | hypothetical protein |
| 11 | NC_009787 | 30904-31899 | 5585725 | EcE24377A_C0027 | hypothetical protein |
| 12 | NC_009791 | 1408-1503 | 5585895 | EcE24377A_A0001" | hypothetical protein |

**Table 4:** Comparison of plasmid sequences with genomes of *E. coli* E24377A and *E. coli HS* by Genomic BLAST.

| Query Plasmid | Matched plasmid | Score (BITS) | E value | Identities (%) | Gap (%) |
|---|---|---|---|---|---|
| NC_009791 PTEC_5 | PTEC_6 | 392 | 3e-108 | 78% | 9% |
| | *E.coli*E24377A complete genome | 204 | 2e-51 | 84% | 9% |
| NC_009787 PTEC_35 | PTEC_73 | 8309 | 0.0 | 94% | 0% |
| | PTEC_80 | 3033 | 0.0 | 94% | 0% |
| | PTEC_74 | 2957 | 0.0 | 96% | 1% |
| NC_009786 PTEC_80 | PTEC_74 | 9664 | 0.0 | 98% | 0% |
| | *E.coli*E24377A complete genome | 7169 | 0.0 | 98% | 0% |
| | PTEC_73 | 5299 | 0.0 | 99% | 0% |
| | PTEC_35 | 3033 | 0.0 | 94% | 0% |
| | *E. coli HS* complete genome | 2296 | 0.0 | 99% | 0% |
| NC_009788 PTEC_73 | PTEC_35 | 8323 | 0.0 | 94% | 0% |
| | *E. coli* E24377A complete genome | 6449 | 0.0 | 99% | 0% |
| | PTEC_74 | 6429 | 0.0 | 99% | 0% |
| | PTEC_80 | 5299 | 0.0 | 99% | 0% |
| | *E. coli HS* complete genome | 2061 | 0.0 | 97% | 0% |
| NC_009790 PTEC_74 | PTEC_80 | 9660 | 0.0 | 98% | % |
| | *E.coli* E24377A complete genome | 7249 | 0.0 | 99% | % |
| | PTEC_73 | 6429 | 0.0 | 99% | % |
| | PTEC_35 | 2957 | 0.0 | 96% | |
| | *E. coli HS* complete genome | 1236 | 0.0 | 95% | % |
| NC_009789 PTEC_6 | PTEC_6 | 392 | 4e-108 | 100% | 0% |

***In silico* Identification of Vaccine Candidates**

The primary candidates or bacterial proteins to be considered an antigen is its cellular localization. Proteins restricted to the cytosolic compartment are unlikely to be immunological targets, whereas surface associated and secreted structures are more easily accessible to antibodies, the primary immune effector against bacterial pathogen. Therefore, during present study all manually and computer annotated autotransporter protein sequences were retrieved from Uniprot KB database. Approximately 200 autotransporter protein sequences were collected from UniprotKB and used to analyze the different characteristics of these proteins to find unique proteins. After retrieval of protein sequences their homology were analyzed with human protein. The four sp (experimentally annotated genes) P01559, Q47185, A7ZMU8, A7ZQT4 show no homology with human proteins and selected for vaccine targets.

**Regular Expression (Pattern) of Drug Targets**

These unique drug targets were further studied by PROSITE databases to investigate their patterns (Table 5). Their patterns were analyzed against human proteomes by BLASTp for final confirmation of their uniqueness and to find that either this functional domain is present in this gene or not.

**Table 5:** list of regular expression pattern of drug targets

| Protein name | genes | Unique motif |
|---|---|---|
| A7ZMU8 | mgrB | STK----------LILSFSLCLMVLSC |
| A7ZQT4 | lgt | GRLGNFINGELWG |
| P01559 | Sta1 | CCELCCNPACAGC |
| Q47185 | Sta2 | CCELCCNPACTGC |

**Discussion**

In the present study, the genomes of two closely related strains of *Escherichia coli* E24377A (pathogenic) and *Escherichia coli HS* (non pathogenic, commensal) were compared with emphasis on genome organization and coding proteins. Genomics blast results showed that E24377A and *HS* genome sequences share more than 99% similarity. Most of the genome features of these two strains are more or less similar (Table 1). This is not surprising that because both strains occupy the same niche in the human gastrointestinal tract. Then the differences might have arisen after the divergence of these strains from other evolutionary lineages for adaptations in their host, these increase greatly in frequency in pathogens and appear to be associated with the ability to infect eukaryotes, perhaps reflecting a mechanism for evading host immune defenses.

The genomic comparison showed that E24377A is pathogenic and causes Traveler's diarrhea while *HS* is human commensal and non pathogenic. By genomics comparison, specific mutliloci were identified for PCR diagnosis. For this task, all the plasmids were analyzed and the unique coding sequences were gathered. These unique genes were only present in this E24377A strain and not observed in any other organism sequences yet. Therefore, they serve as targets for strain typing. Further Primers was designed for PCR amplifications. They would help in disease diagnosis. By the *In silico* study of E24377A strain drug targets were also identified. The specific unique protein sequences that are not present in human proteome can be used as drug targets. These drug targets were E24377A specific enterotoxin namely Heat-stable enterotoxin A2 (Q47185) and ST-IA/ST-P (P01559) encoded by Sta2 and Sta1 genes, respectively. (Table 5). Both Sta1 and Sta2 enterotoxin comprise 72 amino acid. Study of these genes from different isolates revealed variable numbers of contiguous copies of the motif CCELCCNPACAGC. These enterotoxins activate the particulate form of guanylate cyclase and increases cyclic GMP levels within the host intestinal epithelium. Bacterial lipoproteins have many important functions and represent a class of possible vaccine candidates. The prediction of lipoproteins from sequence is thus an important task for computational vaccinology. The A7ZMU8 and A7ZQT4 show no homology with human proteins and selected for vaccine targets. The *Escherichia coli* E24377A specific lipoproteins**,** diacylglyceryl transferase (A7ZQT4) and Protein mgrB **(**A7ZMU8) encoded by Igt and mgrB, respectively were identified unique as compared to human proteome. These lipoproteins were suggested as competent vaccine targets. mgrB lipoprotein comprises 47 amino acid and low magnesium induced hypothetical protein. It has a role as a transcriptional modulator of steroid biosynthesis. *mgr* locus profoundly affected extracellular protein production, suggesting that the locus may regulate many other genes as well (4). Prolipoprptein diacylglycerol transferase have 291 amino acid. This unique patch is conserved in this prolipoprotein family. This is also present in A7ZQT4 gene. By comparing with complete genome of 24377A this shows 73.1% identity. Following signal peptide-directed export of the prolipoprotein, processing occurs by the enzyme prolipoprotein diacylglycerol transferase (Lgt). Lgt uses phospholipid substrates and catalyses the addition of a diacylglycerol unit onto the thiol of a crucial conserved cysteine, which is located within the 'lipobox' motif at the cleavage region of the prolipoprotein signal peptide. This cysteine therefore forms the N-terminus of the mature lipoprotein (5).

**Conclusion**

Comparative genomics has advantages that it provides a deep insight into the genomes of two bacteria and further deep insight into two genomics. Aside from the indirect effects that might flow from vaccine and drug discovery, comparative genome sequencing has already had done impact in the clinical microbiology laboratory, most notably in the field of epidemiological typing, and to a lesser degree in diagnostic bacteriology and the study of antimicrobial resistance. Just as comparative genome sequencing can reveal differences useful for typing strains within a species, it can also be used to provide molecular diagnostic target that can distinguish closely related species, subspecies to find novel targets for immunodiagnosis. 4 *Escherichia coli* E24377A specific loci were identified during present study will be very helpful for PCR diagnostics. Discovery of novel drug targets would be incredibly valuable and effective for antibiotics and vaccine development. Further microbiologist and pharmacologist can work on the drug designing of these targets, so that a cure to traveler's diarrhea can be made.

**References**

1. Preuss P. Berkeley Lab Science Beat: Comparative Genomics at the Joint Genome Institute: an Interview, 2002.

2. Feng Weagant, Sand Grant M. Enumeration of Escherichia coli and the coliform bacteria. Bacteriological analytical manual. 2002; 29(9):24-29.
3. Chen SL, Hung CS, Gordon JI. Identification of gene subject to positive selection in uropathogenic strains of Escherichia coli, a comparative genomic approach. Epub 2006; 103(15):5977-82.
4. Jason W, David A. Analysis of the global transcriptional profiles of ETEC isolate E24377A. American society for microbiology, 2012.
5. Mary P, swati B, Steven M, Regino M. Precolonized human commensal E.coli strains serve as barrier to *E.coli* 0157:H7 growth in the streptomycin treated mouse intestine. Infection and Immunity., 2009, 2876-2886.
6. Thanh T Luong, Steven W Newell, Chia Y. Mgr, a novel global regulator in Staphylococcus aureus. Journal of Bacteriology. 2001; 185(13):3703-3710.
7. Salgado H, Santos ZA, Gama CS, Millan ZD, Collado VJ. Regulon DB: transcriptional regulation and operon organization in Escherichia Coli K-12. Nucleic Acids Res, 2001; 29:72-74.
8. David A Rasko. The pangenome structure of Escherichia coli: comparative genomic analysis of E.coli commensal and pathogenic isolates. Journal of bacteriology. 2008, 6881-6893.