Comparative Genome Analysis of Three Pathogenic Strains of *E. coli, Salmonella* and *Shigella*

Chandra Shivani*[1], Kumari Abha[1], Grover Alka[1] and Nehra Sampat[2]

[1]Amity University Uttar Pradesh, Noida
[2]Birla Institute of Scientific Research , Jaipur
Email: schandra4@amity.edu, akumari@amity.edu, agrover@amity.edu, nehrasampat@gmail.com

ABSTRACT

Bacteria belonging to family *Enterobacteriaceae* are well-known for their association with pathogenicity in humans. Availability of complete genome sequences of multiple strains of several bacterial pathogens has contributed immensely to our understanding of the high level of similarities and differences among closely related organisms at sequence level. *Escherichia coli, Salmonella* and species of *Shigella* are among the best-studied genomes of diarrhea causing bacteria, yet there is much to be learned about the nature and evolution of interactions, bacterial diversity and pathogenesis. In this study a whole genome comparative genomics approach was used to analyze the genomes of *E.coliO157:H7 str sakai, Shigella dysenteriae* and *Salmonella typhimurium str.LT2*. The main objective of the study was to analyze the genomes of three diarrhea causing pathogens at sequence level. The study revealed that the three genomes have around 60% sequence similarity. Comparison of the chromosomes of different enteric bacteria identified a common set of so called “core genes” that are, in general, shared among enteric species. Also several highly conserved coding sequences that potentially have virulence function were identified between *E.coliO157:H7 str sakai & Shigella dysenteriae* and *E.coliO157:H7 str sakai and Salmonella typhimurium str.LT2*. This suggests that evolutionary conserved coding regions consist of some important virulence factor coding regions thus providing useful information for the identification of various diarrhoea causing elements. The results from the whole genome comparison of the three bacteria revealed significant virulence factor genetic heterogeneity between the *E.coliO157: H7 str sakai and Salmonella typhimurium* genomes, but their backbones are conserved.

KEYWORDS: Comparative Genomics, Conserved region, Diarrhea, Virulence factor.

***Corresponding Author:**

Chandra Shivani
[1]Amity University Uttar Pradesh, Noida
Email: schandra4@amity.edu, akumari@amity.edu, agrover@amity.edu, nehrasampat@gmail.com

INTRODUCTION

Comparative genome analysis of closely related species has substantial power to identify genes, define gene structure, highlight rapid and slow evolutionary change, recognize regulatory elements and reveal combinatorial control of gene regulation. The bacterial family Enterobacteriaceae contains some of the most devastating human and animal pathogens, including *Escherichia coli*, *Salmonella enterica* and species of *Yersinia* and *Shigella*[1]. Diarrhea is the second biggest killer of children globally, with more than 800,000 under-fives dying every year according to UNICEF. A quarter of these deaths occur in India.The most common organisms responsible for most cases of diarrhea obtained from pooled data worldwide include *Rotavirus, E.coli, Shigella, Vibrio cholerae, and non-typhoidal salmonella.* The family Enterobacteriaceae comprises facultative anaerobic gram-negative bacilli as *E.coli, Shigella, and Salmonella* which reside principally in the gastrointestinal tract of vertebrates and cause diarrhea[2].

So far, a complete genomic sequence of *E.coliO157:H7 str sakai, Shigella dysenteriae sd 197 and Salmonella enteric subsp enteric serovar typhimurium str.LT2* has been published in NCBI. For the gram-negative diarrhea causing *Enterobacteriaceae* family to maximize its ability to efficiently and effectively use this publicly available data, systematic research in the areas of functional and comparative genomics needs to be stimulated. These comparisons, will allow us to elucidate and analyze the coding or non-coding DNA sequences that are conserved [3].

In this report, we use comparative genomics approach to analyze the conserved coding sequence between narrow host range Enterobacteriaceae species: *E.coliO157:H7 str sakai & Shigella dysenteriae and E.coliO157:H7 str sakai & Salmonella typhimurium.* Analysis of conserved coding sequences among genomes from Enterobacteriaceae species was conducted to assess the general and Virulence properties of conserved coding sequences in these genomes and their correlation with diarrhea.

MATERIALS AND METHODS

Genomic Sequences Information: Complete genome sequence of three Enterobacteriaceae species were obtained from NCBI (National Center for Biotechnology Information) (http://www.ncbi.nlm.nih.gov). The virulence factors of above bacterial pathogen were obtained through the Website of the Virulence Factor Data Base (VFDB) http://www.mgc.ac.cn/cgi-bin/VFs/compvfs.cgi?Genus=*Escherichia* .

1. Comparison of Genome structure at overall genome statistics: The overall nucleotide statistics features such as genome size, overall (G+C) content, total gene number, protein coding gene, structural RNA, gene density (gene/kbp), % coding, present as a global view of the similarities and differences of the genomes were obtained through the Website of the National Center of Biological Information (NCBI).

2. Identification of conserved region: Conserved region were identified by using Whole Genome VISualization Tool for Alignments (wg-VISTA)[4] . The core function of the VISTA suite of tools is to generate DNA from sequence two or more organisms with various types of annotation alignments and then visualize and analyze them. SLAGAN is an alignment program designed to work seamlessly with VISTA[5] . SLAGAN globally aligns DNA sequences of arbitrary length for the purpose of annotation and biological discovery using syntenic genomic sequences from two organisms). The pair wise *E.coliO157:H7 str sakai / Shigella dysenteriae sd 197 and E.coliO157:H7 str sakai / Salmonella typhimurium str.LT2* alignments were performed by SLAGAN alignment algorithm and were displayed with the wg-Vista graphical server, by applying default parameters. The wg-VISTA result presented graphically visualized showing the mapping of conserved regions in corresponding genomes as well as statistics result of conserved region [2,6,7,8]**.**

3. Comparative analysis of Virulence Factor coding regions: These Virulence Factor coding genes of individual bacterial strain were analyzed with other bacterial strain, using wg-VISTA alignment result of conserved region.

4. Phylogenetic analysis: A tree built based on Virulence Factor coding sequence data is called a gene tree since it is a representation of the evolutionary history of genes, as opposed to organisms.

RESULTS AND DISCUSSION

Comparison of Genome Structure at overall Genome Statistics:

A comparison of the genomes of these three pathogenic sequenced enteric bacteria immediately highlights some important common traits (Table 1). All the three genomes have a single chromosome, ranging in size from 4.3–5.4 Mb. Different strains may also harbor extrachromosomal DNA in the form of plasmids. The total number of genes is highest in *E.coli* with 5371 genes whereas gene density is highest in *Shigella* with only 4660 genes. As a classic feature of prokaryotes most of the genes in all three species are protein coding. Comparison of the chromosomes of different enteric bacteria identifies a common set of so called “core genes” that are, in general, shared among enteric species. These core genes can be regarded as genes that perform “household” functions associated with the common shared lifestyle of intestinal colonization and transmission (environmental survival). Such core genes may play a role in central metabolism or polysaccharide biosynthesis or encode common structural proteins [9] **.**

Table 1: Overall genome statistics.

General features	***E.coliO157:H7***	***Shigella dysenteriae***	***Salmonella typhimurium***
Length (base pairs)	5,498,450 nt	4,369,232 nt	4,857,432 nt
Total number of Gene	5371	4660	4620
Gene density (gene/kbp)	0.9768	1.0665	0.9511
Protein coding gene	5229	4270	4423
Structural RNAs	141	107	118
G+C content (%)	50.5	51	52.2
% Coding:	85%	76%	86%

Identification of conserved region

The wg-VISTA identify that conserved sequences in *E.coliO157:H7 str sakai, Shigella dysenteriae and Salmonella typhimurium are 65%, 67%* and 59% respectively (Table 2). The conserved sequences were identified that shared greater than 70% identity over at least 100 bp. E.*coli O157:H7 str sakai* have 61% coding region whereas Shigella *dysenteriae* and *Salmonella typhimurium* have 60% and 55% conserved coding region respectively. The percentage of conserved coding region was less in *Salmonella typhimurium* whereas *Shigella dysenteriae* contained most of the conserved non coding region. The wg-VISTA statistics result show that E.coliO157:H7 str sakai was more similar to *Shigella dysenteriae* than *Salmonella typhimurium* and genome rearrangement was more in *Salmonella typhimurium*.

Table 2: wg-VISTA statistics result.

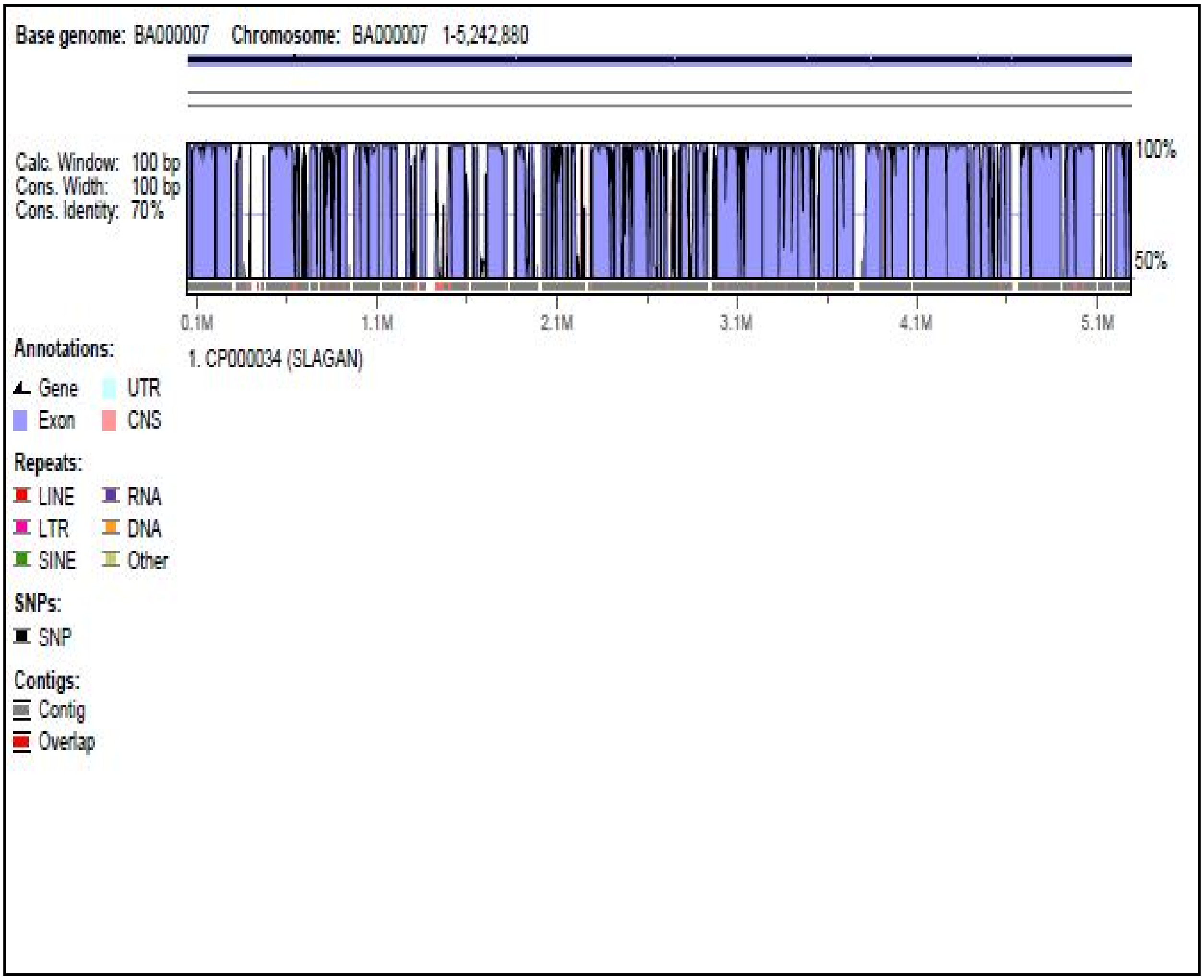
General features	***E.coliO157: H7***	***Shigella dysenteria***	***Salmonella tymphimurium***
1. Length of conserved regions	3158592	3509106	3089573
2. % of conserved region	65%	67%	59%
3.Length of conserved Coding region	2948588	3148000	2883933
4.% of conserved Coding region	61%	60%	55%
5.length of CNS	210004	361106	205640
6.% of CNS	4%	7%	4%

Comparative Analysis of Virulence Factor Coding Regions

The comparative analysis of coding regions of virulence factor suggests that Shiga-like toxin Stx1A and Stx1B of *E.coliO157:H7 str* sakai were 99% similar with stxA and stxB of *Shigella dysenteriae* (Figure 1), which are not only important factors in disease pathogenesis of diarrhea but also responsible for the haemolytic uremic syndrome (HUS).Other virulence factors of *E.coliO157:H7 str sakai* such as auto transporter, iron uptake, hemin uptake show high homology with *Shigella dysenteriae* (Table 3).

Table 3: *E.coliO157:H7* Virulence factor match with *Shigella dysenteriae*.

Virulence factor	**Gene**	**E.*coliO157 sakai***		**Shigella dysenteriae sd**		**%Match**	**product**
		Start	**End**	**Start**	**end**		
Autotransporter	Aida/ECs1396	1444150	1446999	4297060	4298583	97%	AidA-I
Iron uptake	chuS/ECs4379	4388387	4389415	3289488	3290516	98.80%	hypothetical protein
Hemin uptake	ChuA/ECs4380	4389464	4391446	3287457	3289439	99.50%	Heme transport protein
	chuT/ECs4382	4392129	4393043	3285863	3286855	98.20%	putative Hemin binding protein
	chuX/ECs4384	4394413	4394907	3283378	3284493	97.40%	ShuX-like protein
	ChuY/ECs4385	4394907	4395530	3283378	3284493	97.40%	ShuY-like protein
	chuU/ECs4386	4395615	4396571	3281566	3283289	99.00%	putative Hemin permease
Shiga - like tocin	Stx1A/ECs2974	2924769	2925716	1284812	1283865	99.90%	Shiga toxin I subunit A precursor
	Stx1B/ECs2973	2924490	2924759	1285091	1284822	100.00%	Shiga toxin I subunit B precursor
	Stx2A/ECs1205	1266965	1267924	1284078	1284248	69.00%	Shiga toxin 2 subunit A



Base genome: BA000007 Chromosome: BA000007 1-5,242,880

Calc. Window: 100 bp
Cons. Width: 100 bp
Cons. Identity: 70%

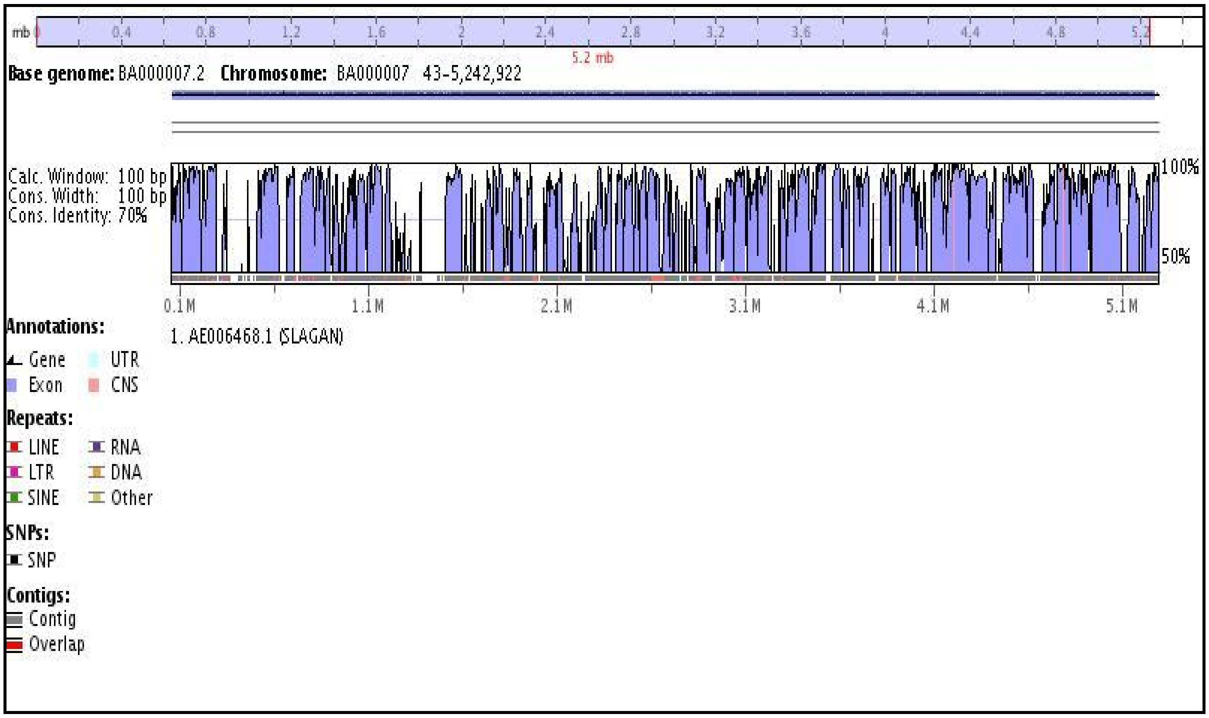
Figure 1: wg-VISTA result between *E.coliO157: H7 str sakai* (BA000007) & *Shigella dysenteriae sd* 197 (CP000034).

Shigella dysenteriae virulence factor as Enterobactin synthesis, Enterobactin transport, Heme transport, Shiga toxin were 99% similar with E.coliO157:H7 str sakai (Table 4). *Salmonella typhimurium* virulence factor as Fimbrial adherence determinants, Secretion system: TTSS, TTSS-2 translocated effector, Stress protein, Two-component system were more than 70% similar with E.coliO157:H7 str sakai (Figure 2). *Salmonella typhimurium* has no Shiga toxin virulence factor(Table 5)[10].

Table 4: Shigella dysenteriae Virulence factor match with E.coliO157:H7

Virulence Factor	**Gene**	**Salmonella *typhimurium LT2***		***E.coli sakai***		**%Match**	**Product**
		Start	**End**	**Start**	**end**		
Fimbrial adherence determinants	csgG	1228025	1228858	1459991	1460824	83.20%	putative curli operon transcriptional regulator
	csgF	1228885	1229301	1460851	1461267	81.10%	curli production assembly/ transport component
	csgE	1229328	1229723	1461292	1461681	79.30%	curli production assembly/ transport component
	csgD	1229728	1230378	1461686	1462336	81.60%	putative transcriptional regulator
	csgB	1231133	1231588	1463090	1463545	82.90%	minor curlin subunit precursor
	csgA	1231630	1232085	1463586	1464044	73.20%	major curlin subunit precursor
	csgC	1232147	1232473	1464103	1464429	72.20%	putative curli production protein precursor
	bcfC	25803	28424	653410	654427	67.80%	Fimbrial usher
	fimA	604130	604672	650741	651043	70.30%	fibrin
	fimC	605325	606017	651454	651850	71.60%	periplasmic chaperone
	fimH	608675	609682	655182	655772	71.60%	minor Fimbrial subunit
	fimF	609692	610210	656197	656303	71.00%	putative Fimbrial protein
	fimZ	610256	610888	656313	656939	71.90%	Fimbrial protein Z
	lpfA	3827449	3827985	4457511	4458047	70.20%	long polar Fimbrial protein A precursor
	stcD	2242832	2243839	2863364	2863465	72.50%	putative outer membrane lipoprotein
	stcC	2243855	2246344	2864164	2866634	71.50%	putative outer membrane protein
	stcB	2246358	2247041	22020	22417	70.60%	putative periplasmic chaperone protein
	stcA	2247097	2247627	2867848	2867954	71.80%	putative Fimbrial-like protein
Secretion system :TTSS	prgK	3014996	3015754	3719532	3719694	73.60%	needle complex inner membrane lipoprotein
	prgI	3016075	3016317	3720287	3720391	70.50%	needle complex major subunit

	spaS	3031533	3032603	3723144	3723412	68.40%	type III secretion protein
	spaQ	3033385	3033645	3724771	3724931	69.60%	needle complex export protein
	invC	3036677	3037972	3728133	3728731	70.20%	type III secretion system ATPase
	invG	3041597	3043285	3732698	3733719	71.20%	outer membrane secretin precursor
TTSS-2 translocated effector	sseK1	4375350	4376360	3863923	3864600	73.00%	putative cytoplasmic protein
Stress protein	sodC	1130053	1130586	1202890	1202991	70.60%	superoxide dismutase precursor
Two-component system	phoQ	1317226	1318689	1610041	1612172	78.20%	sensor kinase protein
	phoP	1318689	1319363	1610041	1612172	78.20%	response regulator



Base genome: BA000007.2 Chromosome: BA000007 43-5,242,922

Calc. Window: 100 bp
Cons. Width: 100 bp
Cons. Identity: 70%

1. AE006468.1 (SLAGAN)

Figure 2: wg-VISTA result between *E.coliO157:H7 str sakai* (BA000007) & *Salmonella typhimurium* (AE006468).

Table 5: *Salmonella typhimurium str.LT2* Virulence factor match with *E.coliO157:H7 sakai.*

Virulence Factor	Gene	Salmonella *typhimurium LT2*		*E.coli sakai*		%Match	Product
		Start	End	Start	end		
Fimbrial adherence determinants	csgG	1228025	1228858	1459991	1460824	83.20%	putative curli operon transcriptional regulator
	csgF	1228885	1229301	1460851	1461267	81.10%	curli production assembly/ transport component
	csgE	1229328	1229723	1461292	1461681	79.30%	curli production assembly/ transport component
	csgD	1229728	1230378	1461686	1462336	81.60%	putative transcriptional regulator
	csgB	1231133	1231588	1463090	1463545	82.90%	minor curlin subunit precursor
	csgA	1231630	1232085	1463586	1464044	73.20%	major curlin subunit precursor
	csgC	1232147	1232473	1464103	1464429	72.20%	putative curli production protein precursor
	bcfC	25803	28424	653410	654427	67.80%	Fimbrial usher
	fimA	604130	604672	650741	651043	70.30%	fibrin
	fimC	605325	606017	651454	651850	71.60%	periplasmic chaperone
	fimH	608675	609682	655182	655772	71.60%	minor Fimbrial subunit
	fimF	609692	610210	656197	656303	71.00%	putative Fimbrial protein
	fimZ	610256	610888	656313	656939	71.90%	Fimbrial protein Z
	lpfA	3827449	3827985	4457511	4458047	70.20%	long polar Fimbrial protein A precursor
	stcD	2242832	2243839	2863364	2863465	72.50%	putative outer membrane lipoprotein
	stcC	2243855	2246344	2864164	2866634	71.50%	putative outer membrane protein
	stcB	2246358	2247041	22020	22417	70.60%	putative periplasmic chaperone protein
	stcA	2247097	2247627	2867848	2867954	71.80%	putative Fimbrial-like protein
Secretion system :TTSS	prgK	3014996	3015754	3719532	3719694	73.60%	needle complex inner membrane lipoprotein
	prgI	3016075	3016317	3720287	3720391	70.50%	needle complex major subunit

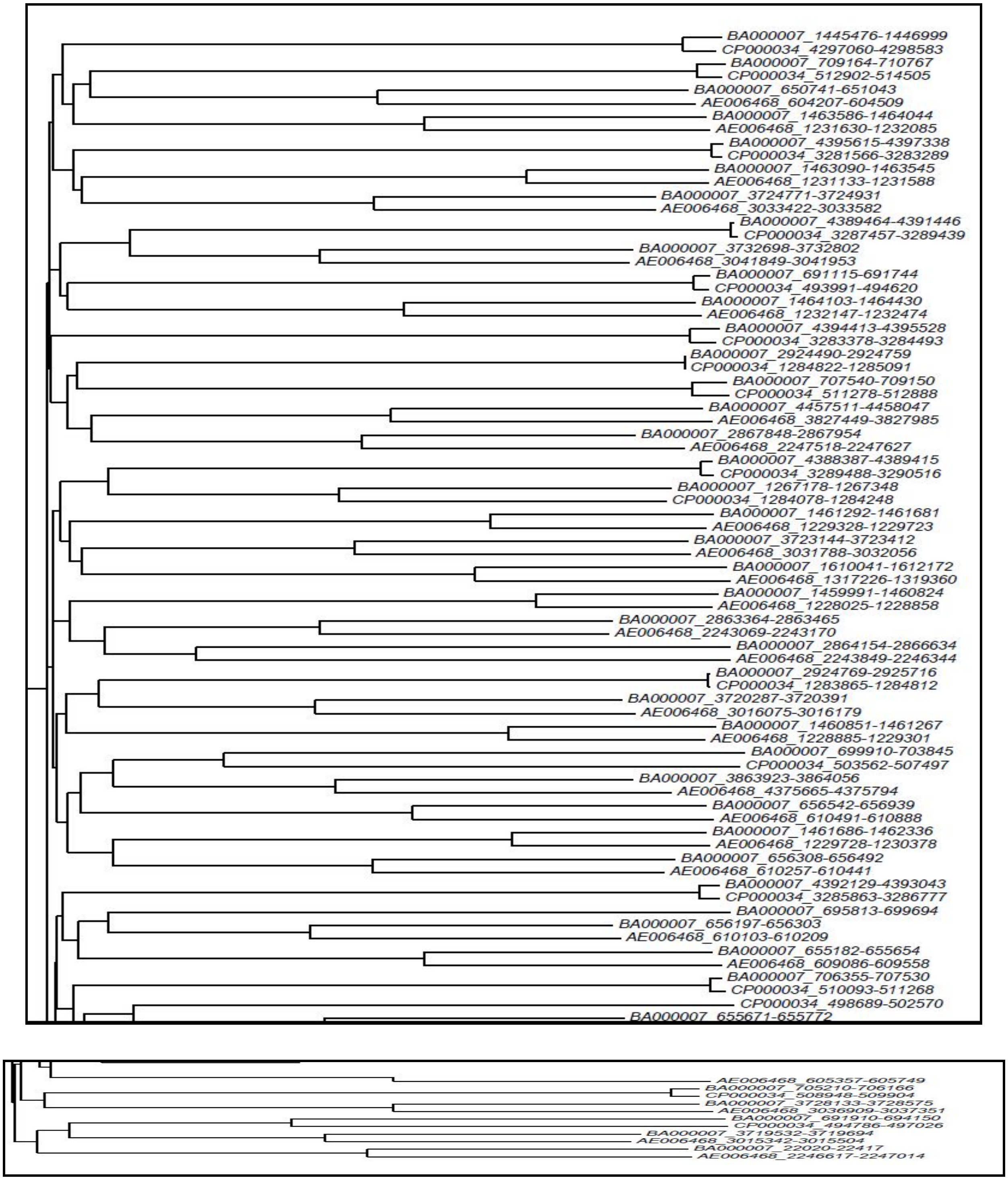
	spaS	3031533	3032603	3723144	3723412	68.40%	type III secretion protein
	spaQ	3033385	3033645	3724771	3724931	69.60%	needle complex export protein
	invC	3036677	3037972	3728133	3728731	70.20%	type III secretion system ATPase
	invG	3041597	3043285	3732698	3733719	71.20%	outer membrane secretin precursor
TTSS-2 translocated effector	sseK1	4375350	4376360	3863923	3864600	73.00%	putative cytoplasmic protein
Stress protein	sodC	1130053	1130586	1202890	1202991	70.60%	superoxide dismutase precursor
Two-component system	phoQ	1317226	1318689	1610041	1612172	78.20%	sensor kinase protein
	phoP	1318689	1319363	1610041	1612172	78.20%	response regulator

PHYLOGENETIC TREE

A phylogenetic analysis of virulence factor coding gene in conserved region show that *E.coliO157:H7* (BA) and *Shigella dysenteriae* (CP) were more similar than *E.coliO157:H7* (BA) and *Salmonella typhimurium str.LT2* (AE). The similarity between *E.coliO157:H7* (BA) and *Shigella dysenteriae* (CP) is not surprising as E. coli and Shigella strains are thought to have diverged from a common ancestor ~10 mya [11]. These results are in agreement with the findings that *E.coli* and *Shigella* has a common gene pool and can not be separated in to two groups [12].

Finally, the results from this research suggest that Shiga toxin virulence factor is similar between *E.coliO157: H7 str sakai & Shigella dysenteriae* and Secretion system: Type III secretion system (T3SS /TTSS) virulence factor is similar between *E.coliO157: H7 str sakai & Salmonella typhimurium*. These in silico analyses revealed significant virulence factor genetic heterogeneity between the *E.coliO157: H7 str sakai & Salmonella typhimurium* genomes, but their backbones are conserved [13].

Figure No.3 Virulence Factor phylogenetic tree constructed for the Virulence Factor coding genes that present in conserved region of *E.coliO157:H7* (BA), *Shigella dysenteriae* sd 197 (CP)and *Salmonella typhimurium str.LT2* (AE) .



CONCLUSION

In conclusion, it can be said that the comparative genomic approach offers the potential to understand and differentiate the basic pathogenic mechanisms employed by different pathogenic bacteria.

FUTURE WORK

Present work can be extended on the structural as well as on the functional aspects especially of virulence factor proteins found within *E.coli sakai, Shigella dysenteriae and Salmonella typhimurium str.LT2* as three dimensional structures of proteins can be predicted and on the basis of these structures probable functions can be hypothesized. These organisms responsible for most cases of diarrhea and further causes hemolytic-uremic syndrome (HUS) which have no effective prophylactic or therapeutic approach for the prevention of HUS.

REFERENCES

1. Nataro JR and Kaper JB Diarrheagenic *Escherichia coli*. CLIN MICROBIOL REV . 1998; 11: 1–60
2. Ryan KJ and Falkow S. Enterobacteriaceae, An introduction to infectious diseases, Med Microbiol. 1994; 3: 328-332.
3. Liping W, Liu Y, Dubchak I. et al. Comparative genomics approaches to study organism similarities and differences, Methodological Review. J Biomed Informatics 2002 ;35: 142–150.
4. Brudno M, Malde S , Poliakov A et al. Glocal alignment: finding rearrangements during alignment. Bioinformatics 2003 ;19:1 i54-i62
5. Brudno M, Chuong B Do, Cooper GM et al. LAGAN and Multi-LAGAN: Efficient Tools for Large-Scale Multiple Alignment of Genomic DNA. Genome Res. 2003;13 (4): 721-731.
6. Frazer KA, Elnitski L,Church DM et al. Cross-Species Sequence Comparisons: A Review of Methods and Available Resources. Genome Res . 2003; 13:1-12.
7. Frazer KA, Pachter L, Poliakov A et al.. VISTA: computational tools for comparative genomics, Nucleic Acids Res. 2004; 32: W273-W279.
8. Mayor C, Brudno M , Schwartz JR et al VISTA: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics. 2000;16: 1046–1047.

9. Baker S and Dougan G. The Genome of *Salmonella enterica serovar Typhi*. Clin Infect Dis 2007; 45:S29–33

10. Hale T L. Genetic basis of virulence in *Shigella* species. Microbiol. Rev. 1991; 55:206-224.

11. Reid SD, Herbelin CJ, Bumbaugh AC et al, Selander RK, Whittam TS: Parallel evolution of virulence in pathogenic Escherichia coli. Nature. 2000; 406:64–67.

12. Gordienko EN, Kazanov MD, Gelfand MS.Evolution of pan-genomes of *Escherichia coli*, *Shigella spp*., and *Salmonella enterica*. J Bacteriol. 2013;195(12):2786-2792.

13. Ureta-Vidal A, Laurence Ettwiller L, Birney E. Comparative Genomics: Genome-Wide Analysis in Metazoan Eukaryotes. *Nat Rev Genet. 2003;* 4: 251-262.