THINK INDIA Journal

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

Comparative Analysis of Genomic Data To Determine Codon Usage and Amino Acid Usage in *Tropheryma Whipplei*

Amit Joshi[2] , Vikas Kaushik[1*] and Joginder Singh[1]

[1]Department of Bioinformatics, School of BioEngineering and BioSciences, Lovely Professional University, Punjab, India

[2]Department of Biological Sciences and Biotechnology, Surajmal Agarwal Private Kanya Mahavidyalaya, Kichha, Uttarakhand 263148

[*]Corresponding Author E-mail: vikas.14664@lpu.co.in

Abstract

The next-generation sequencing methods assisted the scientific community in creating an enormous amount of genomic data, which is found to be very useful in terms of comparing the structural properties and pathogenicity of various bacterial strains. *Tropheryma whipplei* causes constitutional diseases like Gastroenteritis, Endocarditis, where aetiology directly depends significantly on genetic sequences. In this research Twist strain and TW 08/27 of *T. whipplei* were selected for In-silico analysis. As the chosen bacterium is fatal in live form, so to design an effective preventive vaccination, the primary target is to analyze the available Genomic sets. In this research, it has been observed the higher similarity between these two strains in their genomic sets. Intra-strand biasness at codons develops the greater expression of genes mostly due to translational selection. But strand-specific mutational stress was not highly significant at amino acid usage levels. The basic research finding in this study was to determine the preference of GC rich codons in leading strands that were expressing rapidly in comparison to lagging strand in *T. whipplei*. Also, the characterization of two genomes was achieved by comparing

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

membranous genes, which is responsible for exhibiting significant diversity in strains of *T. whipplei*.

Keywords *Tropheryma whipplei*, Constitutional diseases, Etiology, Genomic diversity, In-silico analysis

Introduction

George Hoyt Whipple in 1907 identified with silver nitrate with hydroquinone based histological staining, bacterium in vacuoles associated with macrophages at connective tissue site of patients, which causes intestinal lipodystrophy (Whipple's disease) that result, disturbances in fat metabolic schemes [1]. Dr. George promulgated this communique two years after his graduating convocation from the Johns Hopkins University in 1905. Afterwards, he triumphed with Nobel Prize in physiology and medicine (for his "revelations pertaining to expansion of liver therapy for anaemia" in 1934). Whipple's disease is an extremely subtle malady. During 80 years past (1907-1987) there were 696 recited cases and the prevalence since 1980 has been approximately 30 individual exemplifications per year extensively. The affliction has a proclivity for Caucasian males (less than 3% in other indigenous faction), implying a concealed genetic propensity. Ordinarily, the age criteria for this ailment range from 40 to 60 year old individual with central tendency of 50 year. The proportional distribution of this disease gender wise is 8: 1 (male to female). Whipple's ailment occurs extra frequently among peasants (35%) and diverse edaphic or animal associated personage. The origination and development of Whipple's ailment remains esoteric. Intrusion or infiltration of this microbial entity is comprehensively distributed in the living torso, in conjunction with the Gastro-intestinal epithelia, Blood capillary system and lymphatic system, endothelium, hepatic organ, central nervous system, cardiovascular system, respiratory system, synovium sites of oesteological system, nephric tissues, and bone marrow (stem cell sites), and skin epithelia (dermal layers). Such sites exhibit a significant curtailment of inflammatory response to *T. whipplei*. Furthermore, this organism wields no striking cytotoxic repercussion upon host cells, thereby allowing hefty accretion of *T. whipplei* entities at zonal demarcation of malady. Host cellular proteins clout the course of infection without affecting bacterial genomics. The prevalence of Whipple's disease is very low in comparisons to its pervasive presence in the societal surroundings. Accordingly, it has been advocated that host

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

factors designated by immune defalcation is culpable for the reinforcement of whipple's ailment in humans. In sufferers with Whipple's ailment, tangential T cell procreation and propagation is diminished after invigoration with phytohemagglutinin and concanavalin, but subjects or convalescents have usual notch of immunoglobulin proteins, concluding a definitive glitch in cellular immunity process. In 2003 first cultured variety of Gram-positive *T. Whipplei* named as Twist Marseille strain get entrenched, and its complete set of DNA was partially sequenced with GenBank accession no. NC_004572 [2]. The total genomes of *T. whipplei* Twist and TW08/27 strains was completely sequenced and have GC content of 46%, which indicates low genome sizes and low GC content [3]. Diagnosis of whipple's ailment can be conducted by PCR based method by 16S rRNA-based detection method and *rpoB*-based molecular detection, *rpoB*-based molecular detection gives a 100% positive predictive value, although on a limited number of specimens [14].

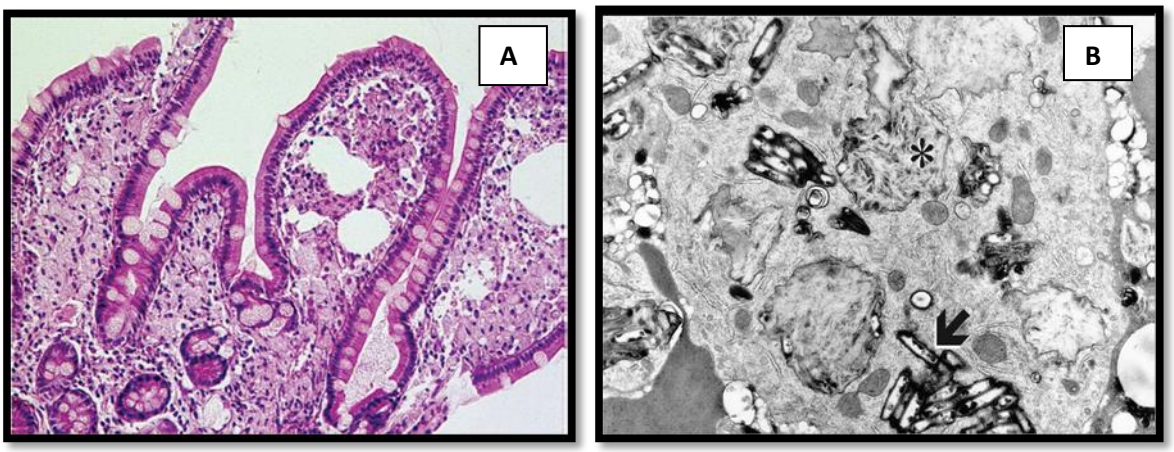


Figure 1 A. Whipple's ailment: aggrandizement of the villi repleted by frothy macrophages or histiocyte, by Hematoxylin-eosin stain method. B. clusters of bacterium species (arrow) and distinctive lysosomes (asterisk) observed in Gastrointestinal tissue macrophage in Whipple's ailment.

Next-Generation Sequencing (NGS) involves modern approaches which are cost-effective like Illumina Solexa, and Ion Torrent approaches that already replaced the Sanger, Maxam-Gillbirt methods [4]. The modern NGS techniques like Whole Genome Sequencing and Whole Exon Sequencing after the successful completion of Human Genome Project generates an extensive

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

volume of highly graded data which is required for computationally exhaustive bioinformatic or in-silico scrutiny, but data repository complications mostly resolved by the construction of Databases like NCBI GenBank, EMBL etc [5]. Most of the works on selected pathogenic bacteria *T. whipplei* is based on computational analysis of genomes that are accessible in the NCBI genomic directory (https://www.ncbi.nlm.nih.gov/genome/genomes/162#). Enlisted strains are available in this database are as follows:

Table 1. Available Genomic sequences for various *Tropheryma whipplei* strains at NCBI GenBank

Strain	Assembly	Size (Mb)	GC%	Scaffolds	Genes
Twist	GCA_000007485.1	0.927303	46.3	1	912
TW08/27	GCA_000196075.1	0.925938	46.3	1	893
Sali28	GCA_000359265.2	0.927465	46.4	1	872
DigADP25	GCA_000359325.1	0.883649	46.4	134	852
Neuro14	GCA_000376505.2	0.885853	46.4	154	858
Neuro20	GCA_000359345.1	0.883582	46.4	125	857
DigMusc17	GCA_000359385.1	0.884564	46.4	128	856
Neuro1	GCA_000346085.2	0.927567	46.3	1	866
Art1	GCA_000346105.2	0.927575	46.3	1	869
Dig7	GCA_000359205.2	0.927564	46.3	1	867
Dig10	GCA_000359225.2	0.927515	46.4	1	867
Endo27	GCA_000359245.2	0.927598	46.3	1	866
Bcu26	GCA_000359405.1	0.880271	46.4	138	856
Art29	GCA_000359285.2	0.927595	46.3	1	867
Dig15	GCA_000359145.2	0.927582	46.3	1	868
Pneumo30	GCA_000359305.2	0.927553	46.4	1	872
Endo32	GCA_000359165.2	0.927567	46.3	1	863
Dig9	GCA_000359365.1	0.880115	46.4	146	847
Slow2	GCA_000359185.2	0.927621	46.3	1	858

Methodology

All sequences of genomic and proteomic interests related to *T. whipplei* were retrieved from NCBI GenBank. For phylogenetic establishment between considered strains, BLAST-N tool is deployed. To reduce sampling related errors, annotated coding sequences up to 100 codons were excluded out. Analysed duplicate sequences, transposable elements with internal stop codons, non-translational fragments were also not considered during the analysis. Approximately 700 sequences were considered for this study.

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

Genes of the template and non-template strands were retrieved based on oriC in *T. whipplei* selected strains; for this analysis, Oriloc is used [6]. RSCU (Relative synonymous codon usage) and RAAU (Relative amino acid usage) among *T. whipplei* strains were depicted by deploying CODONW 1.4.2, that helps in applying COA (corresponding analyses).

GC content analysis for each codon positions was calculated for all selected coding sequences in the selected pair of strains for *T. whipplei*. Other useful specifications such as CAI (Codon adaptive index) [7], RSCU, RAAU, GRAVY Score (average hydrophobicity), the total number of repeatable feature of each and every selected codon and also average complexity quotient is analyzed for translated proteins to set up an systematic in-silico assay to determine factors influencing amino acid usage.

Nucleotide patterns, MEGA program used to estimate pairwise synonymous divergences (d_S) and moreover non-synomymous divergence divergences (d_N) [8].

Comparative genomic hybridization data is already available in GenBank accession numbers EF536098 to EF536142 [9], that describes major genomic changes or trends were related to transcriptome encoding for WiSP membrane protein family and also indicates about AT biasness. To get acquisition of the connotation of the correlation between the positions of concatenation along significant coordinates of correspondence study and microbial vacillating parameters use STATISTICA: Linear regression analysis. Prediction of the disarrayed regions within amino acid stretches determined by deploying GlobPlot. It is a web based assistance that grants the user to sketch the tendency within the query for protein systematic or globular structure.

Results and Discussion

The sequences of various strains of *T. whipplei* are retrieved from the NCBI website and then the BLAST-N tool is deployed to analyse the similarities between available genome sets for various bacterial strains. This genomic blast produces a dendrogram (figure 2). This depicts the phylogenetic relationship between all the considered strains of *T. whipplei*.

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

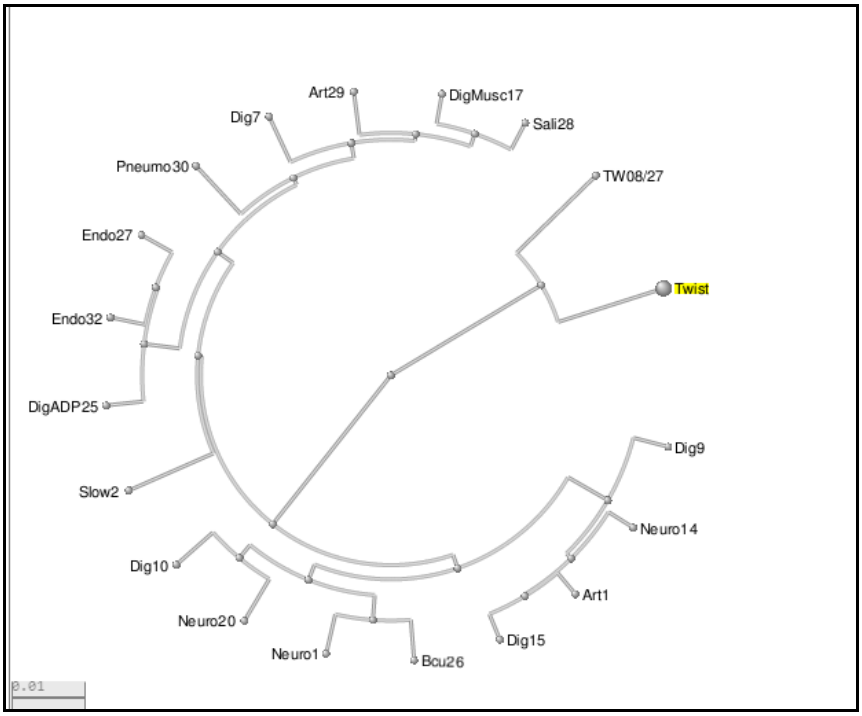


Figure 2. Dendrogram based on genomic blast

The above Dendrogram shows a relevant relationship between different strains of *T. whipplei* bacterium (https://www.ncbi.nlm.nih.gov/genome/?term=tropheryma+whipllei). It is based on genomic comparisons to analyze phylogenetic relationships between various strains of the selected bacterium.

In Figure 3 BLAST (Basic local search alignment tool) of NCBI is utilized for comparing nucleotide sequences of *T. whipplei* Twist and TW 08/27 strains as their complete genomes are available. It is useful in predicting that Twist strain and TW08/27; both strains have greater level of similarity from 408-555 K base pair regions in their genome sets. Score per cent identity in Figure 4, that indicates 99.66 per cent similarity in genomic analysis both considered organisms.

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

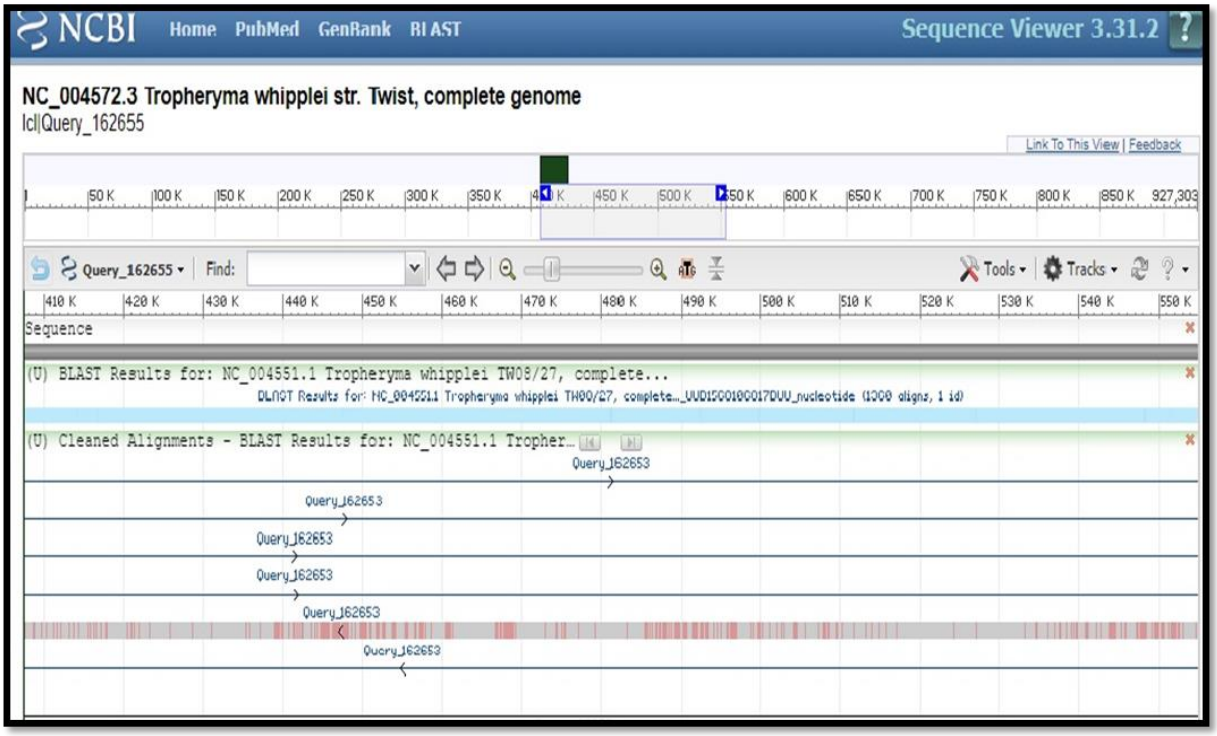


Figure 3. BLAST results for Twist and TW 08/27 Strains

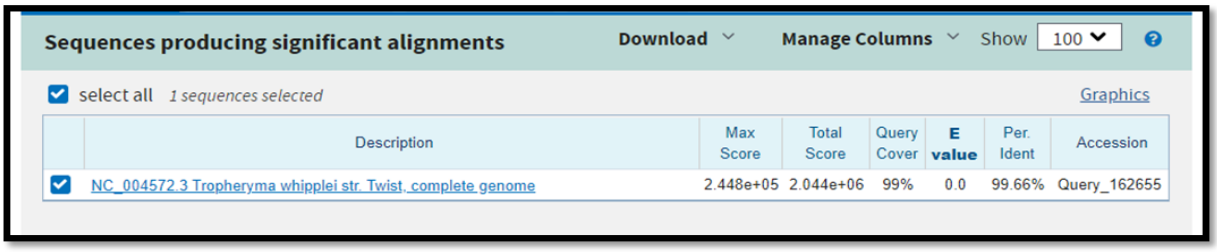
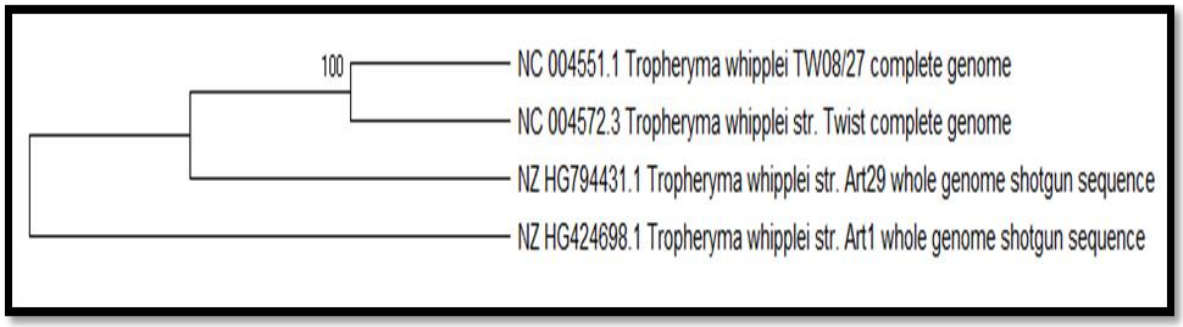


Figure 4. BLAST score per cent identity for Twist and TW08/27 (reference) strains

This also indicates similarity in proteomic sets that can be deployed in achieving the crucial task of vaccine designing by reverse vaccinology approach. In simple words, one can easily depict that the current study will circumscribe the welfare domain of human health.



ISSN: 0971-1260
Vol-22-Issue-16-August-2019

Figure 5: Results of Phylogenetic analysis among 4 strains of *Tropheryma whipplei*

In Figure 5 the evolutionary antiquity was analyzed by deploying the UPGMA approach. The optimal tree with the sum of branch length = 1.4633 is presented. The transformative gaps were determined by deploying the p-distance approach and it is in the number of amino acids unit variability per site. This investigation ramified 4 amino acid stretches. The respective stances with site indemnity of less than 95 percent were waived out, i.e., aligned or calibrated gaps of less than 5 percent, omitted data, and dubious bases were acquiesced at random position. Concluding datasets contain 925938 positions out of total locales under consideration. Evolution based transformative investigation were operated in MEGA X program. In figure 6 TW08/27 and Twist coding sequences on I^{st} and II^{nd} major axes were procreated by correspondence analysis on RSCU parametrical findings. The primary or first axis accounts for 9.21% and 9.18%, whereas secondary axis accounts for5.13% and 5.08% of total variable alterations in TW08/27 and Twist genomic patterns respectively. Above two plots shows open quadrilaterals that represent genes transcribed in the leading strand while closed or dark quadrilaterals indicate gene transcribed in lagging strand. While in the bottom two plots show dark or filled circles that represent transcribed genes in leading strand while open circles or unfilled ones indicate highly expressed genes in the leading strand. Such a scattered plot depicts the mutational biases which are strand-specific. Predicted sense strand sequences were found to be greater in leading strand (75.3% for TW 08/27 and 77% for TWIST strains genome) as compared to lagging strand. All such observations lead to conclude replication, as well as transcriptional selection together with asymmetric mutational bias, is the major cause for intragenetic variations in genomes of considered strains. This same figure also indicates most of the genes are clustered at the extreme end of the first axis in lower pictures, here strong correlation is predicted with leading strand sensed or coding regions. Also, low divergence at synonymous regions of highly expressed portions is inferred (d_S = 0.0057), this helps in concluding synonymous codon selection occurs at the translational level also [10].

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

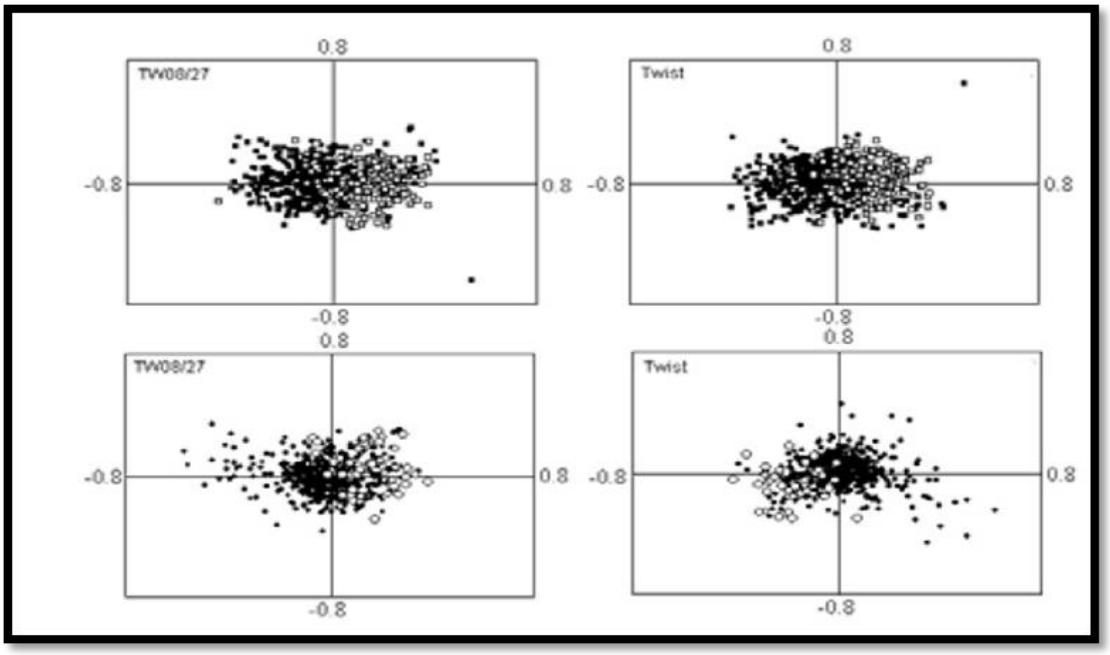


Figure 6. Correspondence study on synonymous transcriptome codon usage (RSCU values)

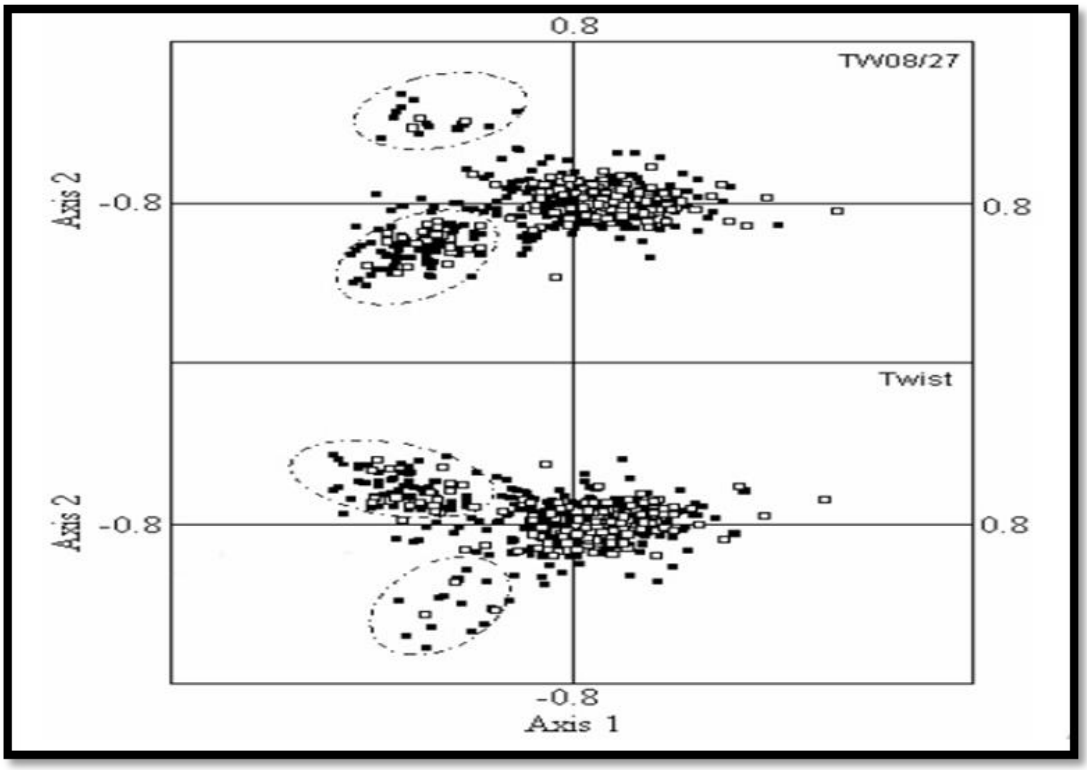


Figure 7. Correspondence study on Relative amino acid usage (RAAU)

In figure 7, plot show filled squares that exhibit genes transcribed in leading strand while empty squares show genes transcribed in lagging strand of replication sites of DNA. Correspondence

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

study on Amino acid Usage of encoded genomic outcomes or peptides depicts nullified segregation of proteins coded by leading and lagging strand nucleotide stretches. So separate analysis for each strand is deployed, and it was found Phe and Val are undoubtedly exhibited in G+U abundant codons of leading strand while Lys, Thr, and Asn are found to be abundant in A+T rich codons of lagging strand. In Table 2 we clearly state that there is a good correlation in terms of hydrophobicity and aromaticity [11] of expressed out proteins, as both are considered the crucial factors for amino acid variability in *T. whipplei* strains. This investigation leads to represent those bunch of genes depicting membrane-linked peptides, in conjunction with WiSP/WND group members and some debatable hypothetic proteinous structures. The cistronic groups of large clusters were found to hold the key for some crucial proteins for bacterial survival like cytochrome C subunits, integral membranous proteins and transporters of various nutrients. But one can conclude that most of them are associated with membrane proteins.

Table 2 Multiple Parameters for encoded proteins are described here

	TW08/27 strain			TWIST strain		
	Variability (%)	**Source of variation**	**Correlation Coefficient (r)**	**Variability (%)**	**Source of variation**	**Correlation Coefficient (r)**
1st AXIS	18.8	Gravy Score	-0.82	18.5	Gravy Score	-0.87
		Aromaticity	-0.67		Aromaticity	-0.68
2nd AXIS	14.6	Gravy Score	-0.58	14.3	Gravy Score	-0.47
		Aromaticity	-0.46		Aromaticity	-0.38
3rdAXIS	8.9	Size/Complexity	-0.79	8.9	Size/Complexity	-0.74
		GC_{12}	0.65		GC_{12}	0.66

Conclusion

In the present study of analyzing genomic and proteomic trends in *T. whipplei* based on NGS data already available in NCBI GenBank, it has been found that asymmetric mutational bias leads to cause variations in codon usage, and intra-strand biasness at codons develops greater expression of genes mostly due to translational selection. But strand-specific mutational stress was not highly significant at amino acid usage levels. The key conceptual finding was to determine the preference of GC rich codons in leading strands that were expressing rapidly in comparison to lagging strand in *T. whipplei*, although it was AT biased. The study also indicates

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

that most of the genomic sets are dealing with membranous proteins that can be useful in determining pathogen-host interactions, dynamicity of pathogenicity islands [12], and furthermore in developing insight for quorum sensing [13] studies. Currently hydroxychloroquine (600mg/day) and doxycycline (200mg/day) jointly used for the treatment of whipple's ailment for tenure of 12 to 18 months, but requires long continuance in follow up is required [15], so it is time-consuming treatment process and only few handful trials were conducted [16]. Due to bacterium evolution and horizontal gene transfer that promotes gain of new antibiotic resistance genetic sequences; it becomes very crucial to implement new strategies to fight infections [17]. Despite the therapeutic interest, there are only a few considerations comprehending genomic scrutiny of *T. whipplei* and remarkably the literature based methods do not deploy the switch or reverse vaccinology approach for vaccine or drug discovery and designing, this will open novel prospects in the comparative genomic studies that will bring cannonade in pharmaceutical expansion for human health, welfare and social development.

References

1. Whipple GH (1907). A hitherto undescribed disease characterized anatomically by deposits of fat and fatty acids in the intestinal and mesenteric lymphatic tissues. *Bull. Johns Hopkins Hosp.* 18:382-393.
2. Raoult D, Ogata H, Audic S, Robert C, Suhre K, Drancourt M, Claverie JM (2003). *Tropheryma whipplei* twist: a human pathogenic actinobacteria with a reduced genome. Genome Res. 13:1800-1809.
3. Bentley SD, Maiwald M, Murphy LD, Pallen MJ, Yeats CA, Dover LG, Norbertczak HT, Besra GS, Quail MA, Harris DE, Von Herbay A, Gobble A, Rutter S, Squares R, Squares S, Barell BG, Parkhill J, Relman DA (2003). Sequencing and analysis of the genome of the Whipple's disease bacterium *Tropheryma whipplei*. *Lancet*.361: 637-644.
4. Blum HE, Oexle K (2014). Clinical Interpretation and Implications of Whole Genome Sequencing. *Next Generat Sequenc & Applic*. 1: 105.
5. Mardis ER (2019).The Impact of Next-Generation Sequencing on Cancer Genomics: From Discovery to Clinic. *Cold Spring Harb Perspect Med*. 3:9(1-9).
6. Frank, CA, Lobry, JR (2000). Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics* 16: 560-561.

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

7. Sharp PM, and Li WH (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*, 15:1281-1295.
8. Nei M, Gojobori T (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*. 3:418-426.
9. My-Van La, Nicolas C, Pascal B, Raoult D, Renesto P (2007). Comparative genomic analysis of *Tropheryma whipplei* strains reveals that diversity among clinical isolates is mainly related to the WiSP proteins. *BMC Genomics* 8:349
10. Das S, Paul S, Chatterjee S, Dutta C (2005). Codon and Amino Acid Usage in Two Major Human Pathogens of Genus *Bartonella* – Optimization Between Replication-Transcriptional Selection, Translational Control and Cost Minimization. *DNA Res.* 12:91–102.
11. Singer GA, Hickey DA (2003). Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene*., 317:39-47.
12. Gal-Mor O, Finlay BB (2006). Pathogenicity islands: a molecular toolbox for bacterial virulence. Cell Microbiol. 8(11):1707-19.
13. Barriuso J, Martínez JM .(2018). *In Silico* Analysis of the *Quorum Sensing* Metagenome in Environmental Biofilm Samples. *Front Microbiol* 9: 1243.
14. M Drancourt, A Carlioz, and D Raoult (2001). *rpoB* sequence analysis of cultured *Tropheryma whipplei*. Journal of clinical microbiology,7(39):2425-2430.
15. Lagier JC, Fenollar F, Raoult D. (2014). Whipple's disease and *Tropheryma whipplei* infections in internal medicine. When to think about it? How to treat? Rev Med Interne 35:801-807
16. Feurle GE, Moos V, Blaker H, Loddenkemper C, Morter A, Stroux A, Marth T, Schneider T. (2013). Intravenous seftriaxone, followed by 12 or three months of oral treatment with trimethoprim- sulfamethoxazole in Whipple's disease. J Infect 66:263-270.
17. Mondal SI, Ferdous S, Jewel NA, Akter A, Mahmud Z, Islam MM, Afrin T, Karim N. (2015). Identification of potential drug targets by subtractive genome analysis of *Escherichia coli* 0157: H7: an in silico approach. Adv. Appl. Bioinf. Chem. 8:49-63.